Springer Series in Synergetics

Springer : COMPLEXITY

Andrei Ludu

Boundaries of a Complex World

Springer

Springer Complexity

Springer Complexity is an interdisciplinary program publishing the best research and academic-level teaching on both fundamental and applied aspects of complex systems – cutting across all traditional disciplines of the natural and life sciences, engineering, economics, medicine, neuroscience, social and computer science.

Complex Systems are systems that comprise many interacting parts with the ability to generate a new quality of macroscopic collective behavior the manifestations of which are the spontaneous formation of distinctive temporal, spatial or functional structures. Models of such systems can be successfully mapped onto quite diverse "real-life" situations like the climate, the coherent emission of light from lasers, chemical reaction-diffusion systems, biological cellular networks, the dynamics of stock markets and of the internet, earthquake statistics and prediction, freeway traffic, the human brain, or the formation of opinions in social systems, to name just some of the popular applications.

Although their scope and methodologies overlap somewhat, one can distinguish the following main concepts and tools: self-organization, nonlinear dynamics, synergetics, turbulence, dynamical systems, catastrophes, instabilities, stochastic processes, chaos, graphs and networks, cellular automata, adaptive systems, genetic algorithms and computational intelligence.

The three major book publication platforms of the Springer Complexity program are the monograph series "Understanding Complex Systems" focusing on the various applications of complexity, the "Springer Series in Synergetics", which is devoted to the quantitative theoretical and methodological foundations, and the "SpringerBriefs in Complexity" which are concise and topical working reports, case-studies, surveys, essays and lecture notes of relevance to the field. In addition to the books in these two core series, the program also incorporates individual titles ranging from textbooks to major reference works.

Springer Series in Synergetics

Founding Editor: H. Haken

The Springer Series in Synergetics was founded by Herman Haken in 1977. Since then, the series has evolved into a substantial reference library for the quantitative, theoretical and methodological foundations of the science of complex systems.

Through many enduring classic texts, such as Haken's *Synergetics and Information and Self-Organization*, Gardiner's *Handbook of Stochastic Methods*, Risken's *The Fokker Planck-Equation* or Haake's *Quantum Signatures of Chaos*, the series has made, and continues to make, important contributions to shaping the foundations of the field.

The series publishes monographs and graduate-level textbooks of broad and general interest, with a pronounced emphasis on the physico-mathematical approach.

More information about this series at http://www.springer.com/series/712

Andrei Ludu

Boundaries of a Complex World

Springer

Andrei Ludu
Embry-Riddle Aeronautical University
Department of Mathematics
Daytona Beach, FL
USA

ISSN 0172-7389 ISSN 2198-333X (electronic)
Springer Series in Synergetics
ISBN 978-3-662-49076-1 ISBN 978-3-662-49078-5 (eBook)
DOI 10.1007/978-3-662-49078-5

Library of Congress Control Number: 2016934064

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media (www.springer.com)

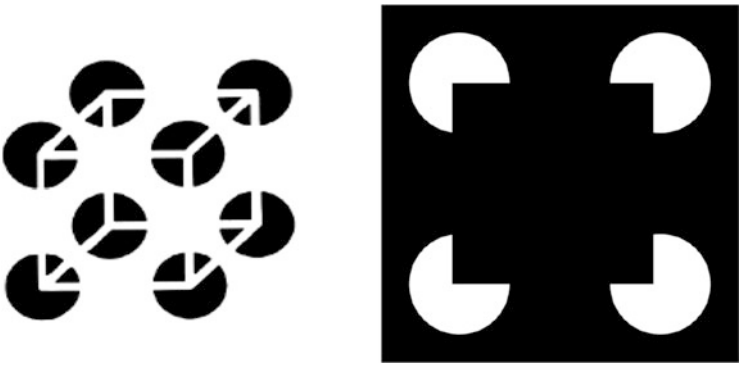
I dedicate this book to my wife, Maria, and to her perseverance and patience

Preface

Have patience with everything unresolved in your heart, and try to love the questions themselves as if they were locked rooms or books written in a very foreign language. Don't search for the answers which could not be given to you now because you would not be able to live them. And the point is, to live everything. Live the questions now. Perhaps then, someday far in the future, you will gradually, without even noticing it, live your way into the answer.

Rainer Maria Rilke

In most dictionaries the word 'boundary' is defined as something that shows where one thing ends and another begins or something that divides something else into two parts. In some dictionaries a boundary is also defined as the limit of a subject, activity, or experience. More mathematical dictionaries define a boundary as the closure of a given set, the points separating the set from its complement. Overall, the boundary seems to be defined by *divide et . . . separare* (see the figure below). In this book we try to discuss the many aspects of the boundary from a unifying point of view, an interdisciplinary angle. We shall examine how important the boundary is for the existence, dynamics, and stability of what it bounds. In other words, we shall discuss and exemplify the extent to which the boundary dynamics is essential for the dynamics of the interior. The book emphasizes the importance of the boundary as a glue, rather than as a separation between various interrelated topics.



Real and apparent boundaries and frames

The direction this book has taken has been strongly influenced, if not even determined, by Dr. Christian Caron, executive publishing editor at Springer and my editorial adviser, who encouraged me to approach the topic of boundaries in parallel with the increasing importance of interdisciplinary topics, complex systems science, and especially the booming new socioeconomic theories. He also drew my attention to several crucial papers for the contents of this book.

Another motivation derives from the mathematical claim that 'a boundary has no boundary'. I first heard this affirmation when at elementary school, from an uncle, accountant by profession and spare-time astronomer by vocation, and for many years I was intrigued by its strange duality. Of course, when I was taught a little geometry, I was able to prove it, but in this book I will try to express the essence of the assertion through several different forms of expression and not only purely mathematically.

Last but not least, another motivation for writing this book built up from many discussions I had with vision scientists, neuroscientists, painters, and art critics on the difference between painting on the circular (hence, infinite) surface of a vase (e.g., ancient Greek vases) and painting on a regular flat surface delimited by the rigid boundary of a frame. The question is whether the existence of a frame has consequences for the power of creativity and for the freedom in choosing a subject. I do not think we found an answer, but at least we raised a question.

The importance of boundary is discussed in this book from several points of view: artistic, sensorial, neuroscientific, social, physical scientific, epistemological, and mathematical. Let us give a simple example of the type of argumentation we develop throughout the book. In a series of psychological studies on delayed gratification, performed in the 1960s by Walter Mischel et al. [1], studies known as the *Standford marshmallow experiment*, the subjects (children) were offered a choice between one small reward (marshmallow, cookie) provided immediately and two such small rewards if they waited for a short period of time. The children could eat one marshmallow right away, but if they waited for fifteen minutes without giving in to the temptation, they would be rewarded with a second marshmallow. The researcher interviewed a girl from the group of children who had waited patiently for the late (hence, double) reward: How did she do it? The girl explained how she "put a frame around the marshmallow" so that it became more of an abstract photograph than a real treat. This comment illustrates a situation where the importance of the frame, even an abstract or imaginary one, grows beyond its supportive mission.

This book is not aimed only at expert readers in some field or another and is thus written in a self-contained manner. The mathematical approaches are introduced gradually where needed, and they are exemplified as intuitively as possible. However, to avoid trading rigor for simplicity, we do introduce some of the sophisticated and counterintuitive concepts in a self-explanatory form: words and images, rather than formulas. Formulas and equations, main theorems, and proofs are always provided as supplements, along with intuitive examples.

The book is divided into three parts. In the first part the human, artistic, and social components are dominant. The second part is dedicated more to mathematical language, including both continuous and discrete mathematics, and the interface

between them. In the third part, we present several applications of the theoretical aspects introduced in the first two parts. Specifically, we cover elements of network theory, big data, and examples from the physical sciences. For the second and third parts, the reader will need a grasp of linear algebra and elementary calculus. This book is a graduate level text: I try to maintain an approachable and fairly constant level throughout, in spite of the breadth and diversity of topics covered, ranging from the visual arts to differential geometry. Globally speaking, this book is mainly addressed to readers interested in complex systems with boundaries, especially those occurring at the interface between different domains like art and neuroscience, fluid dynamics and computer science, networks and geometry, etc.

The book tries to present the concept of boundary from very different angles, yet in a uniform and integrated structure, by focusing on the same major theme, viz., the extent to which the structure of the boundary of a system controls the evolution of the system as a whole. The goal of this book is to present models of phenomena that occur mainly on closed, compact surfaces with boundary, especially where nonlinear and complex solutions are involved. We are acutely aware that the book is still far from being a comprehensive study of the importance of the boundary for systems that have one.

After a motivational introduction, Chap. 2 begins by describing the influence and importance of the frame for the visual arts. The reader is taken on a tour of different opinions and different artistic media relating to boundaries, frames, and contours. Beginning with Sect. 2.5, we move smoothly from the visual arts to vision itself and then to the neuroscience of vision and other perceptions. Section 2.7 describes the possibility of comparative studies involving many senses, different types of perception, and mathematical language.

In Chap. 3, we study the importance of boundaries for social systems. The feedback effect of boundaries on social relations is also described in Sect. 3.3. Sections 3.4–3.6 introduce the elements of social metrics and their relations to social boundaries and socio-mathematics. In the last two sections of Chap. 3, namely, Sects. 3.7 and 3.8, we develop the theory of topological social boundaries, providing a few examples of topological patterns and discussing modern trends.

The second part of the monograph contains an exposition of the basic topics in mathematics and the physical sciences which relate to boundaries in a nontrivial way. We introduce the mathematical language of boundaries using topological and geometrical tools. In Chap. 4, we discuss topology and differential manifolds and also differential forms and fiber bundles. Section 4.7 considers the effects of perturbations of the boundaries on the evolution of dynamical systems, and Sect. 4.8 introduces the elements of cobordism. In Chap. 5, we describe the basic features of discrete mathematics in order to balance the mathematical content of the book. We elaborate on graph theory, limiting ourselves mainly to topological, geometrical, and boundary affirmations, formulas, examples, and theorems. We introduce the elements of algebraic topology and homology in Sects. 5.5–5.7, and we conclude with a unifying overview of discrete and continuous methods.

The third part of this book describes applications. In Chap. 6, we briefly discuss the concept of boundary and boundarylessness in the philosophy of science. In

Chap. 7, we continue the presentation of examples by introducing networks and in particular by describing the Internet as a system with boundary. In Chap. 8, we give examples from the modern topic of big data. We examine three domains of research concerning very large data sets, viz., the dimensionality of data sets, topology of data sets, and holes in data sets.

Chapter 9 is devoted to liquid boundaries, and examples of 3D liquid drops, 2D drops, bubbles, shells, double bubbles, antibubbles, etc., are presented and discussed. Sections 9.7–9.9 present an interesting parallel between Leidenfrost drops, hurricanes, and rotating volumes of liquid confined in containers. The chapter ends with three appendices containing a few mathematical elements needed to better understand the above applications. In the conclusion in Chap. 10, we summarize the most important features of boundaries across the various fields of human knowledge.

Acknowledgments

The success and finalization of this book have been supported by interesting discussions, quasi-infinite patience, and constant encouragement from my family, who guided me through all the aspects and challenges encountered in creating a book with such a complex theme, including the difficulties in choosing graphics and working with images and long files. I am further indebted to Dr. Christian Caron for his skillful advice and assistance at all levels in the completion of the manuscript. I would also like to thank Stephen Lyle for his apposite questions and impeccable solutions, essential for clarifying the final manuscript in all its dimensions, and Mrs. Gabriele Hakuba whose professional help and unbounded patience gave me momentum and support to finalize the book. I am very thankful to the Hunt Library of Embry-Riddle Aeronautical University and especially indebted to the Interlibrary Loan Department which offered me the opportunity to obtain all the necessary references in the fastest, most professional, and most efficient ways. I am especially thankful to my daughter Delia Krimmel and to the Springer editorial staff, who devoted long days to proofreading my manuscript, also providing me with high-level discussions and valuable feedback. Working through her proofreads was pretty much like rewriting the book. During the work on this manuscript, I benefited from discussions with friends and colleagues, over the years, and from all the unexpected, highly unpredictable, and nonlinear answers that my students give to my questions.

Daytona Beach, FL, USA Andrei Ludu
January 2016

Contents

Part III Applications

Part I
Arts and Nonlinear Systems: 'Nonlineart'

In their inspiring interdisciplinary anthology of essays on *Framing Borders in Literature and Other Media*, Wolf and Bernhart (editors) point out that, in the last few decades, any signifying human act, meaningful perception, cognition, or communication involves *frames* [2]. Of course, they use the word in its most generic sense as referring to a structure, skeleton, plan, system of reference, opening, closure, or boundary. In arts that develop and fulfil themes over a certain period of time, such as literature, music, and the performing arts, or what Wolf calls the 'temporal media', the beginning as an introduction, overture, or preamble represents the essential meaning of the frame's beginning. The mission of this initial frame is to label and structure the cognitive development of the given work of art. There is a similar sequence of events in dynamics and other disciplines involving mathematics and applied mathematics, the physical sciences, astronomy, etc., where time evolution is the core of the theory. The frame is the law, the reference, the coding, and in that sense there is no escape from the frame into another non-realistic world.

There is a second use of the boundary metaconcept in its spatial context as a natural physical border, or enframing. From the art perspective, this boundary represents painting, photography, film, and all visual arts. Yet, there are connections between the two types of boundary concepts, and there is interference between the uses. For example, the frame of a blank canvas tells the painter to begin painting. The viewer, shopper, collector, and art critic attempts something completely different. In the first instance, the frame plays the role of structure, and it is an overture for the artist, in fact, a limitation which can sometimes be stressful and painful for the process of creation. However, upon completion of the work, the frame becomes its boundary, and in a simplistic way, its glue.

In the first part of this book we pursue this pioneering research on the frame-theoretical approach in the visual arts, and develop it into a more mathematically oriented language. We maintain the connection with visual perception and physiology, and orient it towards findings in neuroscience, and even a global network framing of knowledge. This first part investigates the consequences of the existence/nonexistence of frames for the psychology of the visual arts. In particular,

we elaborate on how the perception, cognition, and characteristics of visual arts and their productions are influenced by the interaction with the frame and boundary of the artwork. After an introductory chapter in which we seek to decode the message of the frame for the spectator, the book continues with chapters dedicated to the arts, networks, biology, and neuroscience. The first three sections of Chap. 2 discuss the sensory influence of the boundary of a given work of art on our perception and judgment. We discuss the way our brain processes the visual information coming from a framed image, and how this information is enriched, transformed, or distorted by the existence of its boundary.

In Sect. 2.4, we make connections between the way our brain processes visual information and the most novel computational models for image storage, recognition, and reconstruction. In Sect. 2.5, we debate the way different schools and artists use the boundary dilemma.

Section 2.6 focuses on the field of perception and we discuss how the physiology of vision is influenced by the information coming from a bounded image with well determined and definite boundaries.

Finally, Sect. 2.7 compares two people who have done much to further our understanding of the framing and bounding paradigms: René Magritte and Bernhardt Riemann, the first through his surrealist paintings, and the second by founding what we know today as the theory of functions of complex variable, branch lines, and Riemann surfaces. Other art forms, including theater and film, are briefly described in this section.

Chapter 1
Introduction

The origins of the words ‘boundary’, ‘bound’, and ‘frontier’ can be traced to two different etymologies. The oldest comes from expressing the *repetition* of a loud noise, reverberated by surrounding mountains or walls, possibly an echo. A more recent etymology can be traced to an expression meaning ‘under obligation’.

The first occurrence of the root word is noted in Medieval Latin as *budina*, still of uncertain origin. Only around the thirteenth century AD do we find the word *bombitire* in Vulgar Latin, meaning to echo back, to buzz, or to mumble. From here the word is seen again around 1170–1230 in Middle English and Anglo-French in the form *bounde*. Also mentioned in Old French as *bone*, *bonde*, or *bodne*, and in Gaulish as *banna*. Some authors attribute the roots of the word to the Vulgar Latin *tinnitire*, or *tentir* in Old French. A third possible root comes from the Vulgar Latin word *bombus*, meaning a deep, hollow noise, a buzzing or booming sound. Finally, some authors attribute its origin to the Greek *bombos*, also meaning deep and hollow, an echoic sound once again.

In Anglo-Latin, around 1200, we find *bunda*, or *born* in Old French, having already established the meaning of limit, or boundary stone. Around 1400, the word *frontiere* is used in Old French to indicate the prow of a ship, the front rank of an army, facing, or neighboring, believed to arise from the word *brow*. In the mid-fourteenth century, we find *boundary*, *border*, and *bound* used in the sense of fastened, or in the figurative sense of compelled, from *bounden* (the past participle of *bind*).

The first attested use of the word in North America is in the early fifteenth century, where it has the specific meaning of the front line of an army, or again a borderland.

A deviation from the geometrical meaning was recorded around 1550, with the sense of ‘under obligation’, or even ‘made fast by tying’. We find the word back in Europe in 1580, as *bombe* in French and *bomba* in Italian. The word is recorded around 1850 in Old French as *bondir*, meaning to rebound, to resound, or to echo.

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_1

At this time it is also being used to express leaping, rebounding, making a noise, or beating (a drum).

Back in the North American continent, we find the word used to describe mortar shells, or specifically an 'explosive device placed by hand or dropped from an airplane' in 1909. In an American 1920 census report, we find the word 'frontier' used to refer to the margin of 'that settlement which has a density of two or more to the square mile'.

From the point of view of physics, the boundary of a system can exist in two situations: when it separates a system from the rest of the environment (if the system is part of, or embedded in some environment), or when the existence of collective coordinates provides a faithful description of the system, as happens when a system reaches a point where its phase space can be reduced in dimension (Fig. 1.1).

The theory of boundary in science has a living history, from navigation around the Earth to navigation in the universe. Boundary problems are encountered at any physical scale, from quark–gluon droplets and heavy nuclei, to liquid droplets, swimming cells, labs-on-chips, and all the way to neutron stars and black holes. Historically, as almost every new theory begins to develop, another tends to ignore the existence of any natural boundaries. With the development and the occurrence of contradictions and experimental improvements, theoreticians are eventually forced to recognize the various fundamental limits within the theory.

On Earth, we live on the boundary of a compact spherical domain, so our geographical world is the boundary of the sphere, and consequently our world has no boundaries. It is also a 2D world, and the other 2D geographical/geometrical entity that is unbounded is the Euclidean plane, which happens to be flat. Therefore,

Fig. 1.1 An artistic procedure to underline the dynamical duality, inside–outside, of a frame. Pere Borrell del Caso, *Escaping Criticism*, 1874 oil on canvas, Banco de España, Madrid. Public Domain: https://commons.wikimedia.org/wiki/File:Escaping_criticism-by_pere_borrel_del_caso.png

our world must be flat! This was one of the first documented association fallacies in the perennial need to understand nature by models.

Another example is provided by the concept of the speed of an object, a wave, or information. Measured speed is a real number. Numbers are unbounded, hence velocities must be unbounded. This is once again an association fallacy created by theoretical knowledge. This fallacy contributed to the delay in understanding the geometry of our world as Minkowskian, as opposed to Euclidean, and relativistic as opposed to classical. The unboundedness of speed was shown to be false, and a boundary for velocities was established in the relativistic and causal world.

Another example is provided by the cross-multiplication rule which works very well for linear systems. For example, if yesterday I wrote one page, and today I wrote two, I figure that following this trend I will eventually be able to write a whole book in one day. This is the *fallacy of linearity*, which ruled the scientific world for thousands of years, until people figured out that this world is predominantly nonlinear.

In a more general sense, the boundaries of a signal largely determine its properties and qualities. As a first example, a signal carrying one pure note, one frequency, played for an unlimited duration is fully described by that frequency, and its spectrum consists of just one spectral line, one Fourier component. On the other hand, if this pure frequency is generated only for a short interval of time, one needs more parameters to describe the signal. It is necessary to know the time duration, which actually means knowing one more frequency, given by the reciprocal of this time interval, and one also needs to know how the signal started, i.e., the signal phase. The shorter the signal duration, the more information one needs to know to reproduce the signal. In the limit as the duration approaches zero, this pure frequency signal becomes a noise, and hence contains almost the whole spectrum of acoustic frequencies. In other words, there is a trade-off between duration and complexity. In engineering, this is known as the bandwidth–gain trade-off, in physics and mathematics as the principle of uncertainty, and in economics as the opportunity–cost trade-off. No matter what the field of application, the philosophy of this principle is that, by bounding one parameter in a feedback system, the bounds of some other correlated parameter must be extended.

A similar trade-off of bounds can be obtained by playing two sounds with very close frequencies. The closer in frequency the two signals, the longer the time needed to detect this difference. In other words, when we listen to these signals, we hear the phenomenon of beats, that is, a modulation of the resulting signal with a very low frequency equal to the difference between the two original frequencies. In order to detect beats, one needs to wait long enough to be able to count them, so here we have a trade-off between measurement time and proximity of the two frequencies.

Another interesting situation where the boundedness of a parameter generates significant effects is in differential equations. It is well known that a linear differential system will have a unique solution when we are given enough initial conditions. The unique solution is in general smooth to some extent and can be extended without limit in time. Linearity involves a high degree of predictability. By negation, a

set of non-unique solutions should occur in principle from a nonlinear differential system. Non-uniqueness versus time means that the solution is invariant under time translations, which implies that its time extension must be finite, or bounded. A time-limited signal is zero everywhere outside some bounds, so does not affect the time axis (at infinity) if it is translated in time. It follows that temporally bounded solutions must emerge from nonlinear systems. Moreover, since all identifiable patterns must have a certain degree of time localization, and consequently a degree of boundedness, it follows that pattern generation is intimately connected with nonlinearity, via the property of being limited or bounded.

More examples of the importance of boundedness in science arise from string theory, quantum gravity, and cosmological models. For example, it has been conjectured that the maximum entropy that can be enclosed by a spatial boundary is given by a fraction of its surface area [3]. In physics, the Bekenstein bound

$$S \leq \frac{2\pi k_{\mathrm{B}}}{\hbar c} RE$$

provides an upper limit to the entropy S, or information, that can be contained within a given finite region of space of radius R which contains a finite amount of energy E. In this equation the quantities k_{B}, $\hbar = h/2\pi$, and c are the Boltzmann constant, the rationalized Planck constant, and the speed of light in vacuum. In computer science, this implies that there is a maximum information processing rate (the so-called Bremermann limit) for a physical system that has a finite size and energy.

In computer science, Moore's law [4], which states that the density of independent electronic units increases as a power law versus time inside a given independent functional system, may become another example of an association fallacy. By applying the laws of physics to the process of computation and by applying Moore's law, an extrapolation of current exponential improvements, two more decades would result in computers processing information at the scale of individual atoms [5].

Another example where the boundary is important relates to thermodynamics, and the definition of the thermodynamic temperature. A thermodynamic system must have a number of constituents of the order of the Avogadro number ($N_{\mathrm{A}} \sim 6 \times 10^{23}$). This means that the space of mechanical degrees of freedom and constituent configurations has a huge number of dimensions of the order of the Avogadro number. Let us keep the system in thermal equilibrium, by holding it in contact with a thermostat. This constant temperature constraint is described mathematically by one equation, so the particles in the system must occupy on average a hypersurface of their phase space. Given the huge number of dimensions of this space, the topological difference between the sphere and its boundary is null, and practically all points inside the hypersurface are actually localized on the hypersurface, that is, at the boundary. It follows that, in a large enough system, the thermodynamic laws require the dynamics to happen mainly at the boundary of the space.

Another question arising from these discussions is whether complexity involves a boundary, and if the answer is affirmative, what its structure, topology, and geometry might be. Traditional complex systems are large networks (like neuronal networks,

the Internet, the World Wide Web), physicochemical systems (turbulence, patterns, spin glasses), life science systems (biological morphogenesis, genetic algorithms, evolutionary dynamics, the immune system, socially interacting species), and more recently, intelligent robotic systems (ant-like robotic systems), as well as social, economic, and political systems (e.g., the evolution of cooperation, evolutionary economics) [6].

Since the existence of a boundary involves measurement of localization (topological by neighborhood or metric), one can start by specifying how to measure complexity. One might say that, in the first approximation, complexity could be evaluated through measurement of self-organization [7]. Traditionally, measuring self-organization by entropy decrease does not work, since there are low temperature systems (like Ising systems or Fermi liquids) with low entropy, but which have no organization, and on the contrary, there are biological systems that are thermodynamically driven by increasing in entropy [6, 7].

The option for a system to be complex or to have a boundary is a direct consequence of the observer's interaction with it. When Marcel Proust once commented on Rodin's sculptures, he mentioned that there is some inherent impersonality in a sculpture because the spectator can understand everything about it by moving around and observing. In contrast, in a painting, the viewer is guided by the artist. What creates this difference is either the 2D or 3D aspect, or the interaction of light with the artistic material, or the change of reference of the spectator, or all of these variables interacting in our visual brain. We do not know. For both art works are bounded, one by its own surface, and the other by a frame. What we do comprehend is that, in these effects, we find the signature of complexity.

Chapter 2
Boundaries in Visual Perception and the Arts

> *Like all walls it was ambiguous, two-faced. What was inside it and what was outside it depended upon which side of it you were on.*
>
> Ursula K. Le Guin, *The Dispossessed: An Ambiguous Utopia*, 1974

In this chapter we study several examples and theories from the visual arts in order to analyze the importance of the boundary and frame for these types of artistic creations and perceptions. The main goal of this study is to investigate whether the 'feeling and perception' of a visual boundary is more of an artistic emotional effect, or a psychological effect, or whether it simply has a neurological and physiological explanation through the mechanism of visual perception.

Visual perception is the interpretation of the environment by processing the visible light information assimilated through the visual system. Light travels from direct and indirect sources into the eyes and is projected onto the retina. The multiple paths of the visual information from the retina are pieced together to form a perceptual representation of an object. Then a meaning is attached to the perceptual representation, the object is identified, and human decisions are made. In this process, the information is generated by a 3D distribution of sources, mapped onto a 2D surface (retina), and then mapped again into a 3D (or even a more complex Riemann surface[1]) structure in the central nervous system.

This chain of singular morphisms provides sensory information (especially in the retinal information segment) with some ambiguity. Visual perception is inherently ambiguous since infinitely many real everyday scenes can generate the same retinal information. Some scholars predicate that perceptual inferences are analyzed in the brain through Bayesian inference. However, the inference predicament is easily solved, thanks to additional knowledge referring to prior constraints: "everyday perception uses prior knowledge" [9].

At an associative level, perceptions of the surrounding environment are linked to human decisions. In real life, making a wrong decision is generally costly, thus

[1] For example, each eye projects onto three of the six layers of the primate lateral geniculate body in a very particular way: each half-retina is mapped three times onto one geniculate body: twice onto the parvicellular layers, and once onto the magnocellular layers [8].

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_2

the process of visual perception is associated with a certain risk in the orbitofrontal cortex. The connection between daily perceptions, decisions, and risks brings more affective motivation for the accuracy and faithfulness of the visual observation and perception process.

But what about the perception of art? How does one's mind reduce the perceptual ambiguity in visual art? It is well known that visual ambiguity can create conflicting situations and bistable (or unstable) perception, as in the case of the Necker cube. Without a critical attitude, the inherent arbitrariness of the artistic approach will not stimulate much artistic emotion. Mamassian's solution [9] is that "without specifying a task, the question of how good one is at looking at a painting becomes irrelevant, and the notion of risk associated with an alleged wrong perception becomes meaningless". Consequently, the low level of risk will involve a weak affective motivation, and hence a minimal interest in perception. A plausible path in the perception of visual arts is to return to the challenges of everyday perception. However, in addition to everyday perception which mainly uses prior knowledge, Mamassian considers that visual arts should use conventions for the elimination of ambiguities. The conventions can be inspired by prior visual knowledge, like placing dominant characters at the vanishing point, or can be arbitrary, like placing one eye of a person in a portrait on the vertical median of the frame, or having the face lit from the above-left.

Several researchers advocate this convention-driven way of resolving ambiguity. In her observations, Sylvia Pont et al. [10] concluded that visual awareness is frequently non-veridical. Human observers do not use natural image cues, i.e., perspective in a formal sense, but instead apply a template. Subjects maintain notions of how things should appear in a 'canonical view'. This also seems to be true in particular from the evolutionary perspective, since the canonical view is what viewers form mainly from their visual experiences, and they strengthen prior beliefs about this canonical template in a Bayesian sense. Haber considers that having two simultaneous realities (on the one hand, the flat 2D reality of the picture surface, and on the other, the 3D reality of the natural scenes) makes pictures more complex than natural scenes, while simultaneously the flat reality of the picture makes it possible for the viewer to perceive the layout of space in pictures more easily and accurately [11].

With philosophers and thinkers like Aristotle, Roger Bacon, Thomas Aquinas, Francis Bacon, and John Locke, we find scholars of various disciplines who are favorable to the peripatetic principle,[2] attributing some mathematical finality to the sensorial perceptions [12]. These scholars consider the type of mathematics that we humans handle in our everyday reasoning and thinking processes as having a partially sensorial ascendant. Therefore, some basic mathematical concepts must somehow be related to special human sensorial experiences and abilities. We communicate and think in terms of language: written language (chains of symbols

[2] *Nihil est in intellectu quod non prius in sensu*: nothing is in the intellect that was not first in the senses.

read one at a time) or spoken language (chains of sounds and musical notes, heard consecutively), which are basically chains of algebraic symbols. Conversely, interesting algebraic structures have been detected in language and music; see, for example, Noam Chomsky's minimalist program and syntactic structures, or Blanchard et al.'s game of structure and chance [13]. However, we can also express ourselves and think in terms of images, which are instantly perceived, and connected to principles and theorems of geometry in the most sub-cognitive forms. The fast developments of neuroscience and brain mapping [14] add more support to this peripatetic approach. The majority of the most highly appreciated mathematical tools for signal processing, data mining, and image reconstruction, to enumerate just a few of applications, have been shown to work identically to certain functions of our brain. For example, in image reconstruction, the Gabor filter [15], a commonly used image compression and recognition tool, works similarly to the *simple cells* [16] of the primary visual cortex. Such recent developments show that humans see the world primarily like a barcode scanner.

2.1 Is Our Visual Perception Two Dimensional or Three Dimensional?

The most common definition of the words 'frame' and 'boundary' relates to the act of being surrounded by an edge and exhibiting a 2D aspect, rather than 3D. Bounding or framing simply represent the action of eliminating some degree of freedom, or restricting dimensions: the boundary has one dimension less than its inside. Therefore, recognizing the frame's tendency to be 2D, we may wonder how the brain analyzes and represents space as the subject (human, animal) moves within 2 or 3 dimensions.

Questions about perception of a 3D world that apparently surrounds us have occupied humankind for centuries, and have been debated in the arts, philosophy, mathematics, physics, and lately also neuroscience [17–25]. We live in a world perceived as a 3D space, but when we travel we navigate by referring to a 2D map that accounts for distances on a surface.

If ancestral drawing was not born just as a means for representation and communication, but sprang up as a cognition product and a natural language, the neural representation of a 3D space must have occurred just as a more efficient mapping of the surroundings.

Recent experiments on rats [25] have shown that the hippocampal place cells and the grid cells [21, 22] exhibit vertically-elongated firing fields, indicating that the rat brain may encode the third dimension (elevation) in their motion with less accuracy than the horizontal dimensions. In a similar study, Savelli and Knierim [20] determined that the vertical dimension is encoded in the brain with less precision than the horizontal plane. Hayman et al. [19] studied the encoding of 3D space by place cells and grid cells, and found an anisotropy disfavoring the existence of

the third dimension in rat brains. This group of researchers affirms that we do not yet know whether other areas of the brain encode the third dimension, or whether mammals simply do not need that information to survive: "An animal has a mosaic of maps, each fragment of which is flat but which can be oriented in the way that is most appropriate. Or maybe in our heads, the world is simply flat" (Kathryn Jeffrey in an interview [19]).

Given this asymmetry, it is then legitimate to ask which came first: the 'invention' and use of the third dimension as a survival skill, or its preexistence in the brain as a visually-based neural mechanism needed for 3D navigation? Did we mentally escape from our 2 dimensions at some point in our evolution, or was the ability to use a third dimension already encoded in our brains? Researchers are not certain, as this relationship essentially winds itself into a cycle [21]. Yet, some scientists believe [23] that natural selection drove the development of systems in ancestral mammals to allow their brains to make rapid calculations in order to be able to grab moving insects or other prey quickly and accurately.

Even more technically, some recent studies [21] accept that it is still unknown whether rodent place-coding has a homologue in humans or whether human navigation is driven by a different, visually-based neural mechanism. Maybe the constant human need to navigate throughout space and beyond has enabled human brains to transfer easily from few to many dimensional spaces. Such a hypothesis is supported by studies performed with fMRI (functional magnetic resonance imaging) in 2013 by Mrucksez et al. [17]. It is plausible that, at some point in the evolution of anatomically modern humans, their brain became able to practise abstract thinking and symbolic behavior, and consequently to learn to project real-life events in higher than three-dimensional abstract spaces.

In the seventeenth century, we learned from Descartes and Fermat that space has 3 dimensions. We have since continued to stick to this 'belief bias', supported by thousands and thousands of experiments, data, and verifications. Of course, with the development of modern physics, we learned that space and time involve extra dimensions, and that a grand unification of all interactions, especially gravity and quantum gauge fields, is possible only in even higher dimensional spaces, where particles are strings and trajectories are branes. Even if our physical space appears to have only three large dimensions and, with duration, a fourth, nothing prevents a complete theory from including more than four dimensions. In the case of string theory, consistency requires spacetime to have 10 dimensions. Recently, however, a possibility has surfaced that refutes the belief that the universe has 3 or more than 3 dimensions.

When are we ever going to find out if the world is flat or not, and whichever it is, identify what it is that we really perceive? Well, given the very latest scientific findings, it seems that the archaic *flat Earth* belief may rule again. A glimpse of a new physics that could supplant our current understanding may shift the foundations of what we know about space and time. In a very unexpected way, we may return to the belief in a flat universe, after our trip into the third, fourth, tenth, and even the twenty-sixth dimension.

At the Planck scale (shorter than 5.4×10^{-44} s for time, and less than 1.6×10^{-35} meters for space), it is speculated that the concepts of time and distance break down. At such a tiny scale, all fundamental forces become 'equally strong', and the quantum uncertainty principle becomes the absolute rule [24]. In the world on the laboratory scale, we know that atoms, photons, and other quantum objects fluctuate and obey quantum uncertainty relations between their position and momentum, or between their energy and duration. At the Planck scale, however, space and time cease to be deterministic, smooth, and continuous since they begin to fluctuate themselves.

Apparently, recent calculations predict that, on the Planck scale, space is 2D, and the third dimension is inextricably linked with time [18]. If this is the case, then our 3D universe is nothing more than a hologram of a two-dimensional universe. At Fermilab in Batavia, Illinois, an experiment called the holographic interferometer is in preparation, supposed to measure the quantum noise of space itself. The measurement of this so-called holographic noise may allow us to take a major step forward in our understanding of how spacetime, relativity, and quantum mechanics coexist at the Planck scale, and consequently how spacetime emerges as a structure.

The comparison with holograms is used because, as in the case of a regular laser hologram (a 3D image coded on a 2D surface), the universe may be built in a similar fashion, i.e., its higher-dimensional information may be coded on a flatter and lower-dimensional component. If this hypothesis is true, classical 4D spacetime becomes just the approximate behavior of 2D quantum matter over long durations, viz., an illusion resulting from entanglement of the Planck world with geometrical degrees of freedom.

There is a long history of human obsession with the notion of 2 dimensions. The connection between the mathematical cognition processes in our brain and the phylogenetic and ontogenetic processes (through which this mathematics was constructed in our brain by prior perceptions) reveals a connection between the most important 2D mathematics and 2D neurological mapping processes. From drawings, paintings, photographs, and movies to the 2D theory of complex functions as the richest subject of continuum mathematics, there is historic evidence supporting the idea that our brain favors a 2D map of the world. This is certainly true for this statement: "A picture is worth a thousand words," or the famous Casorati–Weierstrass theorem on the ability of complex functions to reach any value 'they want' close to a singularity, or the recent tendency to understand the 'big data' phenomenon as a 2D structure, or finally the successful signal and system theories based on 2D input–output diagrams. Our human obsession with 2 dimensions seems to be the product of millions of years of evolution, and we seem to give in to Franz Kafka's counsel: "Follow your most intense obsessions mercilessly."

2.2 Message of the Frame

There is one special way of viewing the surrounding environment: through frames. We view a lot of things through boundaries: pictures, photographs, movies, TV, pages, reading glass frames, windows, mirrors, screens, wind shields, etc., and we make a decision based on the evaluations and processing of these framed 2D images. The art critic must decide on the value of the artistic massage from one particular painting, and the pilot must decide on the landing procedure based on the image received from the framed windshield in a visual regulation approach. Traditional visual artwork is generally bounded in space because we need to view it all at once: it must have bounds. The bounds can be formed by the very frame of the painting, or by the walls around a stage in a theater. The same goes for a structureless clutter of objects placed on a table. Place the same objects in a shelf, frame them, and suddenly the clutter becomes organized and structured, and acquires depth.

The viewer's eyes follow a path around the frame and such a cycling motion may bring a sense of infinity. Indeed, because the frames are spatial boundaries around 2D domains, they have no boundaries themselves, so a sort of flavor of infinity is brought by the frame to its captured image. Consequently, the bounded/unbounded properties of the frame may induce or enhance the perception of depth.

Why do taller windows look nicer than small ones? The windows on a building are preferably constructed as tall rectangles because they can then let in as much light as possible from the sky (and also let us know the weather outside). This is contrary to the case of a car windshield which is always made wider and less tall, because the images seen from a car require a specific field of vision, more horizontal, and more 2D.

In this section, we shall argue the pros and cons of a frame and its importance for, and effects on, image perception, from physiological reactions to artistic knowledge and interpretation. Taken together, current literature states that frames, margins, and boundaries induce manifest effects on perception, at different levels in the processing of image information. Because a boundary not only bounds, but isolates, divides, and induces cyclical perception in a steep and nonlinear way, we shall gather these features under a common term which we like to call *nonlineart*.

In the following, we enumerate the main effects of the existence of a frame or boundary as they are presented in the literature:

1. The frame offers historical and chronological recognition, and sometimes even underlines the ontology.
2. The frame stands as a message of discontinuity in the domain of the visual field: 'the artistic message is from this point to this point'. The frame brings compactness and support to the image within. Consequently, it brings extra

wavelengths in addition to the original visible spectrum of the original image.[3] These extra wavelengths induce specific illusions and perception alterations.

3. As opposed to the unframed image, which identifies with the whole ambient situation and is endless, taking over the whole visual field and consequently immovable (there is no other space available to move it to), the framed image is locally defined so it is movable: it is local, but not localized.
4. A framed image is finite, yet connected to infinity: the frame is a boundary and is itself boundless. The boundary is a closed curve, and cyclic, so the frame invites the gaze to engage in an infinite number of loops.

> Any 2D map of the world is a contradiction in terms because, on the one hand, it is represented by a flat picture with boundaries, and on the other, it represents a boundless shape. It is interesting to see how the shape of medieval world maps changed over time as a consequence of increasing knowledge about the Earth's shape: from Eratosthenes to Magellan and Mercator. The oldest 'T-O' type of map, to use the name given by John Gillies [26], is framed by the map of the church perimeter, and the building itself is displayed through sacred geography. An example is the Psalter world map c. 1265. The Ptolemaic world map has a bent, yet flat shape (see Fig. 2.1). As early as 1530, we find a 'cordiform' world map by Peter Apian, which suggests an inward bend around a spherical object. The Chinese world map from 1418, drawn up by Mo Yi Tong, is the first in which periodic boundary conditions tend to substitute a 3D shape.

5. The frame plays the inside–outside game of separation of worlds. The frame of an image is a consequence of the nonlinear interaction among the image elements, an interaction (known as image unity) that holds them together against the natural dispersive forces of the emptiness of the plane, and of their own individualities. See for example Fig. 2.2.
6. There are images insensitive to the existence of a frame, like obsessive repetition of patterns (grids, lattices, quilts), or self-similar structures (fractals, chaos). Such images contain their boundless extension to infinity inside their structure. In contrast, framed images, being compactly supported and not necessarily self-similar or partitioned, are inherently nonlinear. Only nonlinearity can warrant the non-uniqueness of localization necessary for their local definition.
7. A frame with a particular shape can enhance some artistic effects overall. For example, if we study Degas' famous painting of the ballerinas in their rehearsal room, the paintings often depict the ballerinas' bodies cut by the frame, and the frame is rather shallow.

[3] When we rapidly turn off a smooth musical note generated by an electronic device, we hear a sort of snapping, crackling noise. This short signal contains many more wavelengths than the original note. This is just the mathematical effect of abruptly cutting a smooth harmonic signal. The shorter the train of notes, the more extra wavelengths it includes. If we try to listen to one pure musical note for a very short interval of time, we will actually hear only a crackling, which consists in a pulse of white noise containing almost all wavelengths on top of that note. In Sect. 2.4, we elaborate more on this effect, known as the Fourier uncertainty principle.

Fig. 2.1 Ptolemaic world map, second century AD, reconstructed c. 1400. The Earth is displayed as a bent shape, despite the then accepted understanding that the world was flat. This is one of the first instinctual attempts to represent the world in 3 dimensions. Public Domain: https://commons.wikimedia.org/wiki/File:Claudius_Ptolemy-_The_World.jpg

Following Martin Heidegger's ideology, the way technology functions for humans is fundamentally by *enframing* (das Gestell), without a necessary means to an end, but rather as a mode of human existence. Heidegger says that what is revealed in the world, what has shown itself as itself, requires first an *enframing*. He gives Gestell an active role, not reducing it to a simple display apparatus of some sort. In his *Posed Spaces: Framing in the Age of World Picture*, John Gillies [26] considers that *enframing* is the engine for moving forward by 'gathering together', for the purpose of revealing, presenting, and understanding.

Commenting on *enframing* philosophy in his study *Museums and the Framing of Modernity*, Donald Preziosi asks: "What can it mean, then, to be a meaningful image in a world in which there exist frames for everything, and where virtually anything can serve as a frame? What kind of entity, then, is a frame?" Preziosi claims that the elements inside a frame are 'museologised' and somehow re-fabricated in opposition to what is not visible inside. A frame, once created, exists mainly within its circumference and possesses an outside.

Fig. 2.2 Artistic representation of the mechanism of separation and competition between the imaginary 'inside world' and the real outside world, using the frame technique. The 'image' does not fulfil the symmetry laws of reflection, so it must be a true repetition of reality, while the frame has the role of a separation, rather than a final double bar line. René Magritte, *La Reproduction Interdite* (1937, oil on canvas). Permission for reproduction from ©2015 C. Herscovici, Artists Rights Society (ARS), New York

2.3 Importance of the Frame to the Image Inside

Without the ability to localize objects in the environment, it would be nearly impossible to perform important functions in daily life, including obstacle avoidance, navigation, or the development of spatial representations to guide behavior (see Fig. 2.3). Similarly, localization and realization from the image delivered in a 2D photograph require a minimal reference frame (see Fig. 2.4).

What makes the frames important is definitely the improvement in the perception of complex scenes by introducing additional depth cues. It is argued by many authors (Ebenholtz and Benzschawel, Sigman, Goodenough and Flannagan, Lee and Aronson, see for example references in Ebenholtz and Glaser [27]) that eccentrically located retinal patterns (like frames and boundaries) serve the functions of artificial horizon or orientation guidance system by providing a reference signal for egocentric orientation perception. It is likely that such peripheral patterns act automatically without the need for prior perceptual or cognitive processing for size, shape, and depth.

Of course, depth and perspective also require a focal system processing for the recognition and identification of specific patterns; some visual illusions result from processing within the focal system. The way objects in a painting are arranged relative to the frame (of the composition) can be amended by prior expectations based

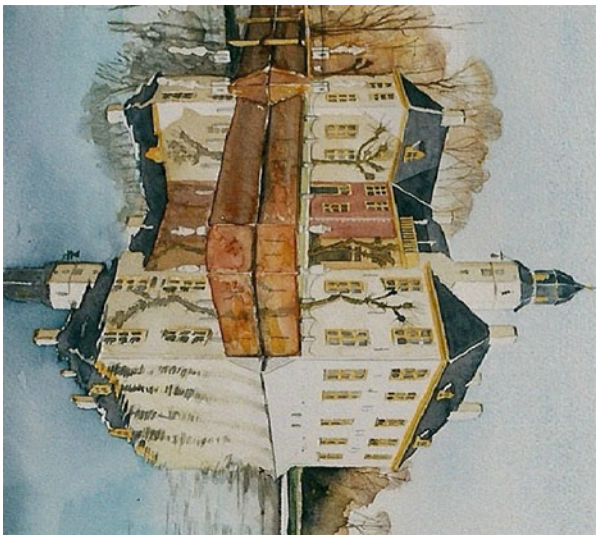


Fig. 2.3 Detail of a painting, artificially rotated through 90° here. Without an explanation, we are almost unsure which one is the real house and which is the reflection

Fig. 2.4 In the complete painting, the boundary of the water allows identification of the real world. *Freylemaborg in autumn reflected in the moat*, ©Anna Poelstra Traga (2012, watercolors) http://www.annapoelstratraga.com

on perceptual organization able to disentangle the patterns from the background. This idea is supported by the Gestalt principles relating to evolutionary criteria inspired by statistical regularities in the natural environment [9]. At the same time composition conventions in the visual arts do not always obey the above-mentioned prior expectations. For example, many objects may be arranged according to some principles of harmony that are similar to the Gestalt principle of organization, but

the placement of individual objects of importance in a painting appears to follow conventions that are quite different from everyday visual expectations.

The perception of the orientation of the objects represented in a framed image is influenced by the orientation of the surrounding frame, and precedes the multiple stages of size and form brain processes; this situation is known from experiments relating to the 'rod-and-frame' effect (RFE) [27]. This effect represents the influence of a large surrounding frame upon the apparent orientation of a rod enclosed within it, the latter appearing to tilt in a direction opposite to that of the frame. Among other conclusions, Ebenholtz and Glaser demonstrated the failure of large frames to exhibit the depth separation effect. Their measurements of the RFE show that depth processing is essentially uninfluenced by processing stages associated with global perceptual properties such as size and shape. In other words, our visual brain decides on the perspective and depth properties independently of and previously to perceptions and judgments based on size and shape. A framed image will talk to the brain in a very particular way, creating the perception of an open window towards an extra depth field. This happens independently of the retinal angle (of course, within certain biological limits) or the shape of the frame, while the same image projected on a blank background will induce the sensation of emerging from the background, and jumping out towards the viewer. Ebnholtz and Glaser show that "large and small frame effects are quite different phenomena", because they relate to functional differences in the focal visual systems.

Paul Duro [26] affirms in his study *Containment and Transgression in French Seventeenth-Century Ceiling Painting* that the frame and the system of perspective are mutually supporting of each other, i.e., symbiotic systems. Depth perception in 2D images is a monocular ability, i.e., it does not rely on stereoscopic cues. Linear perspective (convergence of illusionary parallel lines) is a monocular depth cue to perspective projection. As shown above in the RFE effect, presentation of certain patterns of flat trapezoidal shapes gives the illusion of slated 3D rectangles.

In 2006, Saunders and Backus [28] presented experiments on the quantitative psychophysical measurement of depth perception from perspective convergence. Human subjects understand 3D relationships from monocular images, with very little variability with respect to the projected sizes or slant conditions. In addition, textured images provide better depth recognition (see Figs. 2.5 and 2.6). Through his process of unweaving and reconfiguration, Esparza displaces the textile's potency as a clean-cut symbol for the socio-economic and political issues that arise from living in a border town, El Paso, Texas, a melting pot of varied traditions. The boundary between these traditions, and Mexican–American cultures, is seen in art as a graphic that relies heavily on the border and frame ideas. Even if the threads lie flat on the two walls, running in only two dimensions, the effect of using many different angles of intersection between the bundles provides an overall perception of three-dimensionality, of filling the whole exposition room in different directions.

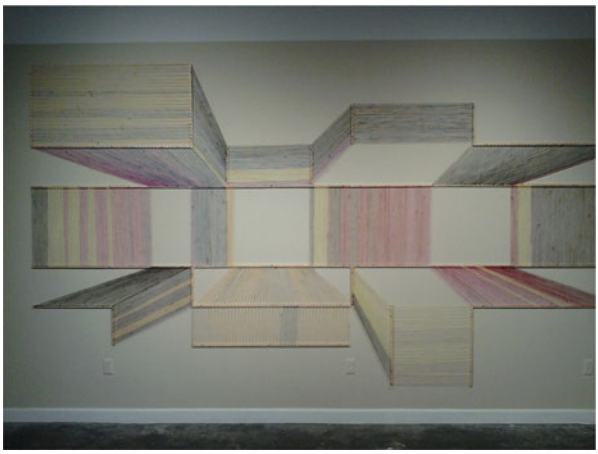


Fig. 2.5 Example of a flat structure delivering a strong feeling of depth. Adrian Esparza, *Spectra Work I*, 2014. Deconstructed sarape textile, wood, nails, bamboo furniture, and dimensions variable. Courtesy of Taubert Contemporary and Houston Center for Contemporary Craft, Houston, TX. Photography by Logan Beck

Fig. 2.6 Another example of a flat structure delivering a powerful perspective effect. Adrian Esparza, *Spectra Work II*, 2014. Deconstructed sarape textile, wood, nails, bamboo furniture, and dimensions variable. Courtesy of Taubert Contemporary and Houston Center for Contemporary Craft, Houston, TX. Photography by Logan Beck

Besides the largest frame surrounding an entire image, if we also consider local frames, like contours and simple shapes, there is evidence of a conditional relation between depth cues and contours. First discovered by Schuman in 1904, and later developed by Kanizsa in 1955 [29], *subject contours* are the special patterns in

Fig. 2.7 Examples of a white triangle and shading cues from Kanizs's subjective contours. Even in the absence of an enclosed outline (no white *triangle*, and no really drawn letter), our visual brain tells us that the negative spaces are a *triangle* and the letter F. Courtesy of Houston Center for Contemporary Craft

which the observer has the illusory perception of a contour even in homogeneous visual field areas (see, for example, Fig. 2.7). Subjective contours occur in the absence of an abrupt change in brightness, and sometimes they appear as bounding a white opaque figure which is in front of other opaque elements. The illusion of an opaque image in front of the others is so strong that small identical elements of patterns drawn inside the subjective contour and outside it appear to have different sizes through an effect of difference in apparent depth. Coren [29] explains the appearance of subjective contours simply as the edges of a subjective plane with a surface that 'ought to be present' in our mind on the basis of the available depth cues of the patterns.

Moreover, in an effort to explain a neurological effect of de-blurring and optimization of the clarity of the image by a mental process of compensation, Georgeson and Sullivan introduced the concept of contrast constancy [30]. This principle states that different patterns can appear to have the same contrast when their physical contrasts are equal, despite gross differences in the contrast thresholds for the patterns. This apparent contrast must be independent of the contrast sensitivity function, and even when the visual information is blurred by optical or neural processes, it can be restored by an active process of compensation.

In a rather surprising study, researchers Giesler et al. [31], by analyzing human eye movements while searching for and detecting targets in a complex naturalistic background, discovered that the most needed quality is a visual memory that can suppress the inhibition of return. The details are not so crucial to the study, but their location is essential. In order to be able to integrate across, one needs to remember *where* the details are. Not only do these findings explain why humans perform so well in extended visual search tasks, given that they have relatively poor memory for image details, but they reiterate the importance of returning to fixed points (like milestones, frames, etc.) within the image.

Another important effect generated by the existence of a frame around a 2D image is the genuine enhancement of the image, by providing an extra reference frame for the tilt aftereffect (TAE) [32]. While inspecting the image, our eyes generate sequences of retinal maps that are processed by our visual brain at a certain rate. A periodic incidence of a standard fixed-angle frame in our visual field should enhance the stability of a such a retinal frame. While analyzing the environment, we have the impression of a stable world, despite the continuously changing positions of our eyes, head, and body. Recent neurophysiology studies provide indications of possible mechanisms. Visual perception is based on several reference frames among which the retinotopic coordinates, the craniotopic coordinates, and body coordinates. Extensive experimental studies performed by Knapen et al. show that the critical factor in the TAE is the "correspondence between the adaptation and test locations in a retinotopic frame of reference, whereas world-centric and head-centric frames of reference do not play a significant role".

Typical studies show that, in the absence of a system of reference, subjects manifest either a tendency to underestimate (foveal biases) or to overestimate (peripheral biases) target eccentricity. A study performed on the influence of visual boundaries on peripheral localization [33] proved that peripheral biases and non-linear scaling metrics are evident when localization occurs in spaces without clearly visible boundaries. The researchers found that boundaries of the visual space modulate biases in judging target location, and hence influence depth and perspective perception. Their results provide evidence that visual boundaries influence both the reference frame in which localization forms, and the metric imposed by this reference system on the image.

The presence of external boundaries allows a regularization in the perception of the eccentricity, and also acts like a switch from a peripheral to a foveal bias, depending on the degree to which boundaries enclose a region that is separate from the observer. It is known that 3D surfaces serve better than 2D cues as informative cues for defining a functional space. For example, a 4 year-old reorients himself/herself within a certain layout defined by short walls, and fails to do so if the layout is marked by colored tape.

The *dual reality of pictures* (see, for example, [11]) is the human ability to perceive depth from either a flat 2D picture or a 3D environment (and according to some authors this ability works not only for human subjects). A study of the influence of tilting surrounding frames on the detection of bilateral symmetry in the fronto-parallel plane was performed by Herbert et al. [34]. By using dot patterns, Herbert noticed that the orientation of rectangular boundaries influences the speed of detection of associated symmetric patterns. The obliquely or vertically oriented rectangular frames slow down the detection of symmetries along vertical or oblique axes. Patterns were detected faster if they were parallel to the frame axes.

Fig. 2.8 Emphasis of a historical narrative using frames. Giotto, *Legend of Saint Francis*, Scenes 4–6, 1297–99. Upper Church of S. Francesco, Assisi, Italy. Public Domain: https://commons.wikimedia.org/wiki/File:Giotto_-_Legend_of_St_Francis_-_-04-_-_Miracle_of_the_Crucifix.jpg#/media/File:Giotto_di_Bondone_-_Legend_of_St_Francis_-_Scenes_Nos._4-6_-_WGA09116.jpg

The type of frame is relevant to the functional space generated in the observer's mental projection of the space inside the frame. Studies were performed on fMRI response in subjects exposed to views of artificially-created indoor scenes bounded by three different types of frames: high 3D (wall), low 3D (curb), and 2D (stripe). The measured multivoxel pattern activity across the retrosplenial cortex (RSC) scene-selective region showed a significant classification for the size of space when defined by the wall and the curb, but not when defined by the stripe. Such studies show that 3D surfaces serve better than 2D surfaces as informative cues to define a functional space [35].

Frames can also induce temporal dependence. For example, comic strips create a narrative effect. The same temporal dependence effect is used in the door of Santa Sabina in Rome, or in Michelangelo's *Gates of Paradise* in the Florence Baptistery. The panels accommodated by the whole frame provide an effective structural tool which emphasizes the historical narrative. Wolfgang Kemp, or Rico Franses [26], for example, address the impressive continuity in the flow of the biblical narrative, tunneling from frame to frame in Gothic stained-glass windows of churches and cathedrals in northern France (in Paris, Chartres, or Bourges). The frame rapidly grew in importance within the next 150 years, as we can see in the *Life of St. Francis*, painted at the end of the eleventh century (see Fig. 2.8).

It was not only during the Gothic and Renaissance that such panelled partition of the frame was used to narrate stories. More recently, Ryman's paintings develop out of units that are reminiscent of pickets, bricks, and tiles. With a fixation for modularity, and in a reaction to architecture, objects, and materials, Ryman recasts in his *Scrapwalls* elements and features suggesting the *Gates of Paradise* and

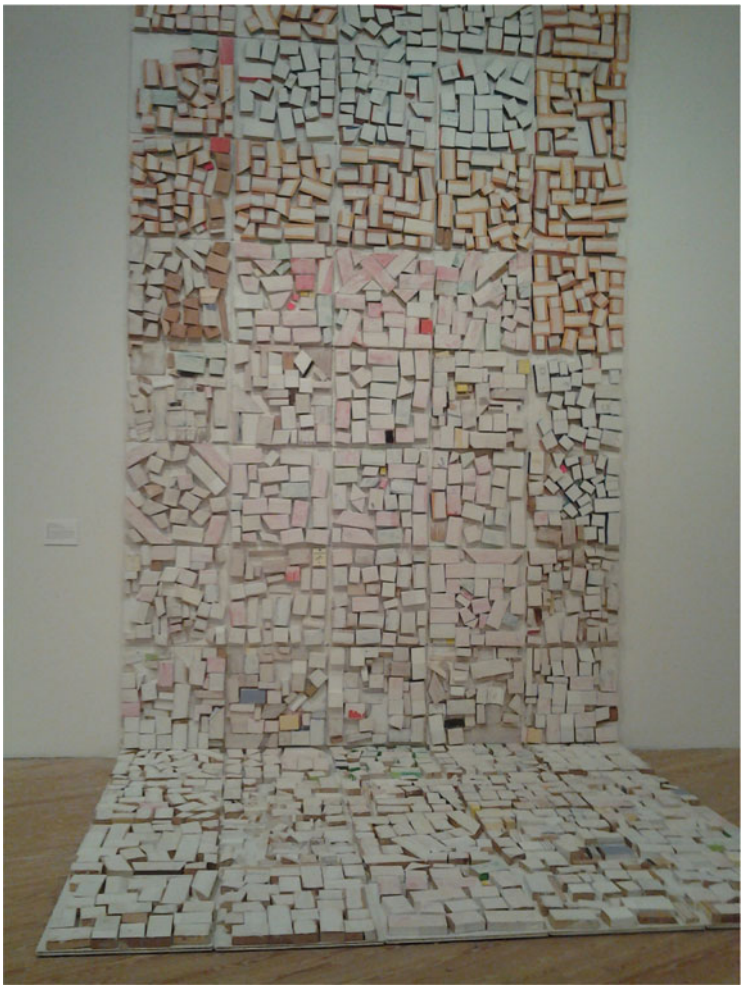


Fig. 2.9 Use of modularity to create a sense of narration and time arrow. Cordy Ryman, *Rafterweb Scrapwall V2* (2012–2013). Acrylic, shellac, and enamel on wood, 30 × 10 feet. Courtesy of the DODGEgallery, NY, Cordy Ryman, and the ZURCHER Gallery, Paris, NY

Rodin's *La porte de l'enfer* (see Fig. 2.9). Similar tendencies using the gate motif as an artistic instrument, and the power of the frame (gates, doors, windows) in focusing attention can be found in many modern artists from René Magritte (see Sect. 2.7), to Don Dahlke (repetitive usage as in *Seductive Silence*), Konstantin Somov, Ben Aaronson, and Neil Simone, or in Samuel Yellin's huge brass doors for the Bok Tower in Central Florida.

Another situation when an art object involves a narrated story can be found in the case of amphorae, vases, and urns (see more details in Sect. 2.5.4). Such symmetric art objects manifest rotational unboundedness at the price of not being able to see the whole message at once. A painted vase requests the motion of the body, like sculptures. A painted vase has a hidden bulk, has secretes because involves cosmic circularity and mystery. There is always an unseen part of it, a next thing to do.

Of course, frames also play a role in cinematographic art. Kemp's *Narrativity of a Frame* [26] gives a very touching example of how important the effect of the frame is in cinema, for both artists and business people. The first ever wide-screen movie (Cinemascope) was *How to Marry a Millionaire* by the Romanian director Jean Negulescu in 1953. Nevertheless, Twentieth Century Fox held this film just to present it together with *The Robe*, directed by Henry Koster and made after the *Passion of Christ* in the same year. As Kemp comments, film producers opted for "the era of big films to be ushered in by the biggest theme of all".

2.4 What Type of Mathematics Does Our Visual Brain Possess?

A successful approach in the mathematics of functions, transformations, transfers, and signals is provided by the theory of orthogonal expansions and multiresolution analysis. This theory gives a complete answer to the question of how we can understand a new and complicated signal using a well studied database of simpler signals. Of course, the procedure is an approximation, and this leaves open the question of how well we can approximate everything.

The starting point in the theory of approximation is the celebrated Weierstrass theorem (see, for example, Theorem 1 in Sect. 4.1). This asserts that any finite-duration continuous signal can be approximated globally and as accurately as we like with power functions (i.e., every continuous function defined on a closed interval has a sequence of polynomials converging towards it in norm). Powers and their linear combinations, the polynomials, are the simplest functions to use, and they work well around singular points. The latter occur when signals break, bifurcate, or approach infinite values (see Fig. 2.10). However, there are many situations where smooth functions can barely be approximated with powers, polynomials, or even orthogonal polynomials (see Fig. 2.11). This happens mainly because the conditions for the Weierstrass theorem involve local constraints, and in order to approximate well with powers, one needs continuity of the signal at every point.

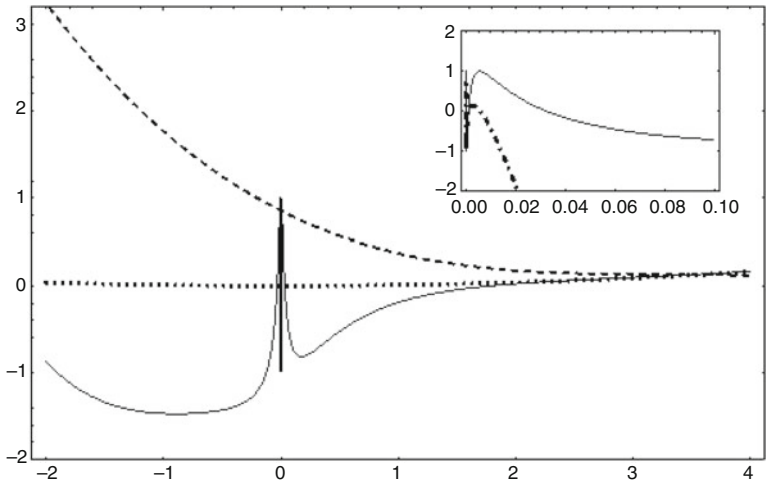


Fig. 2.10 The power functions (*dotted and dashed curves*) are very good approximations in the asymptotic regions of a function (*solid curve*) towards $+\infty$ (*main frame*) and around the singularity at $x = 0$. *Inset*: approximation with negative exponent argument

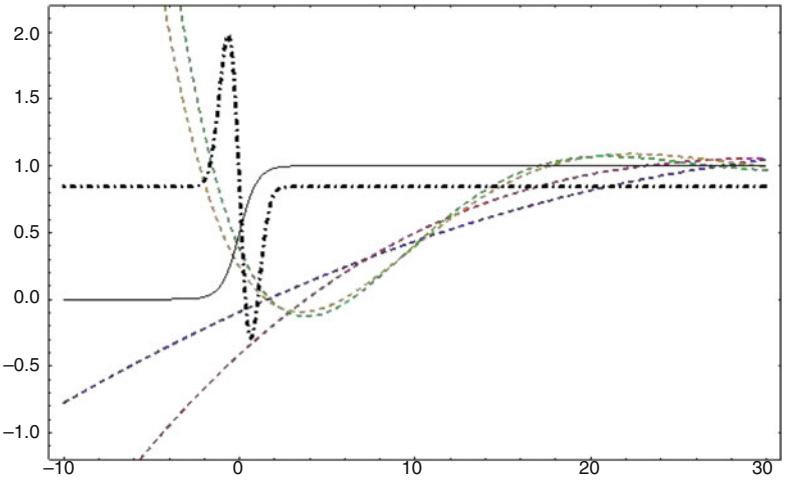


Fig. 2.11 Complete failure to approximate a hyperbolic tangent with polynomials (*dashed curves* represent polynomials, and orthogonal polynomials of order up to 10). However, one single Gabor filter function (*dot-dashed curve*) catches the essential behavior of the function

The approximation is improved if we change the database from powers to trigonometric functions (complex exponentials), or from the field of orthogonal polynomials to the field of Fourier transforms. Not only do the results improve, but the convergence criteria relax compared with the previous approach. For a Fourier series to approximate globally to any accuracy, or to converge in the norm to the function, the only requirement for a finite duration signal is that the area under the square of the signal shape should be finite, i.e., the function must belong to the L^2 class of functions.

The Fourier analysis of a signal provides a very good insight into signal dynamics, namely, the *general principle of uncertainty* [36]. In its simplest formulation, the signal theory principle states that a signal cannot be localized in both time and frequency. For example, a pure musical note of frequency f played for a time interval t is represented mathematically by the sine function $\sin(2\pi ft)$, which extends by definition and by construction from $t \to -\infty$ to $t \to +\infty$. The signal is infinitely extended in time, and its Fourier spectrum is infinitely narrow in frequency space: just one spectral line f (see the right-hand frame in Fig. 2.12). On the contrary, a point localized in space (a signal whose width in time approaches zero) contains all possible harmonics and its Fourier spectrum is infinitely extended (see the left-hand frame in Fig. 2.12). In other words, a signal $f(t)$ and its Fourier transform

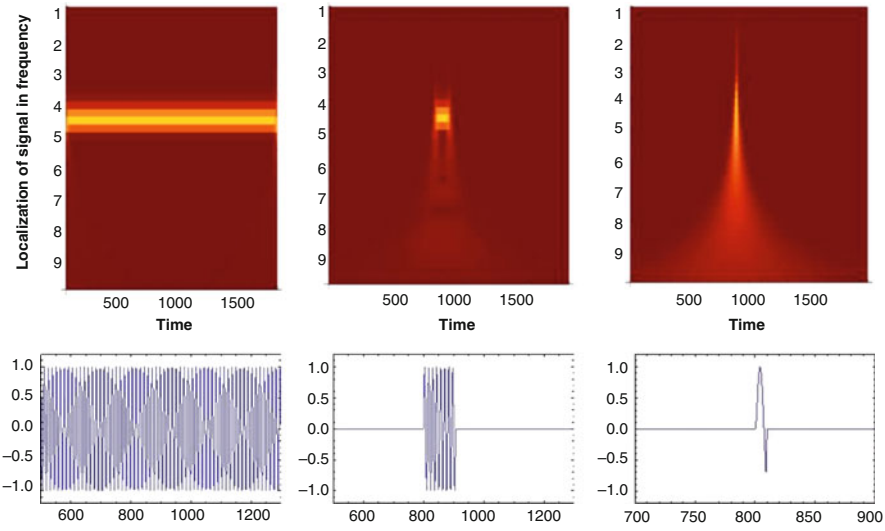


Fig. 2.12 *Upper frames* are density plots of the spectral distribution versus time (*horizontal axis*) and frequency (*vertical axis*, labeled in number of octaves) of a sine signal, while the *lower frames* represent the corresponding signal plotted against time. The *left-hand frame* represents a portion of an infinitely long sine function and its spectral distribution shows as a narrow stripe of localization in frequency. The *right-hand frame* represents a highly localized signal in time, and we can see how its frequencies diffuse through the whole spectrum, especially downwards, towards the lowest frequencies. The *middle frame* shows an intermediate situation, a sine signal with the same frequency but finite duration

$\hat{F}(f)$ cannot be simultaneously well localized. A good quantitative characterization of this uncertainty principle is provided by Hardy's theorem, which states that, if a real function of real variable $f(t)$ and its Fourier transform $\hat{F}(f)$ are simultaneously smaller in magnitude than a sufficiently narrow Gaussian profile, then the function must be zero. That is,

$$\left(\begin{array}{c} |f(t)| \leq \mathrm{e}^{-(t/\lambda_t)^2}\,, \\ \text{and} \quad |\hat{F}(f)| \leq \mathrm{e}^{-(f/\lambda_f)^2}\,, \\ \text{and} \quad \lambda_t \lambda_f < 2 \end{array}\right) \quad \Longrightarrow \quad s(t) \equiv 0\,. \tag{2.1}$$

Considering the Gauss bell function $\exp(-t^2/\lambda^2)$ as a prototype of a localized signal with width λ, (2.1) tells us that, if a signal $f(t)$ has its envelope bounded by a Gaussian shape (so that it is more localized than a Gaussian), and if we also expect its spectrum to be more localized than a Gaussian in the Fourier space of frequencies, then the signal is simply zero. The Hardy theorem is generalized by the Paley–Wiener theorem in many dimensions, including the well known case in the plane [37]. Such uncertainty theorems became famous through Heisenberg's principle of uncertainty, or Robertson's relation of uncertainty. Technically, the Fourier–Hardy and Paley–Wiener theorems bring a more general mathematical understanding than the limited quantum mechanics point of view [36].

The uncertainty principle can easily be verified in everyday life. John Wolfe's website provides many examples showing that the closer two notes are in tune, the longer one needs to listen in order to perceive the difference (see [38] for more details). When musicians tune, they listen to a note for a long time by removing beats, which are regular pulsations of loudness produced by notes that are nearly in tune. If the frequency difference is very small, then we hear an interference beat at very long time intervals. Therefore, small differences in tone require a longer 'measurement' time.

Regarding measurement approximations, we have to admit that the Fourier approach works well for 'smooth enough' signals of various types. However, when signals contain strong discontinuities, or singularities, it becomes harder to tackle these local disturbances while keeping the approximating functions under the same global constraints. The solution is to use pieces of functions, and adapt the right piece to the right local behavior. Since changing the 'degree of discontinuity' should be one of the properties of the set of functions in the database, in order to accommodate various discontinuities and patterns on different scales, pieces of functions are required to conserve their orthogonality even after a change in scale. A good solution for this problem is the *wavelet approximation* method (Haar 1909 [39]), represented by finitely supported (Daubechies 1988 [40]) pieces of trigonometric functions (Morlet 1983 [41]) that can be moved around by translations, or squeezed and expanded by compression and dilation.

Windowed Fourier analysis can be generalized from trigonometric functions to more general 'scaling functions'. Such wavelets adapt locally to the degree of smoothness or discontinuity by changing their scales. This approach is called multi-

resolution analysis (MRA), or multi-scale analysis (Mallat 1989 [42]). Furthermore, the convergence condition for wavelet approximation to a signal involves class L^p functions (the area under the pth power of the signal should be finite). If this condition is fulfilled, pointwise convergence of the multi-resolution series is guaranteed almost everywhere [43]. The series formed by dilations and translations of the same scaling function approaches the required signal with the same rapidity at any point. This property offers a better quality of convergence than the global one (convergence in norm) for either polynomials or Fourier series.

Since their development, wavelets have been successfully applied to signal and image processing, visual reconstruction, numerical analysis, turbulence, fractals, economics, forensic sciences, and quantum mechanics, among many other applications. By virtue of the Paley–Wiener theorem, wavelets and their multi-resolution analysis work under the same principle of uncertainty, in the same way as Fourier analysis does.

Among different types of wavelets for 2D MRA, there is one which has recently raised much interest because of its similarity to the way our visual brain works [16]. The Gabor filter is a Gaussian shape modulated by a plane sine wave [15] (Fig. 2.13). In this way it is localized, yet has patterned structure and scale (the sine wavelength), and it also has an orientation in the plane through the angle between the plane axes and the plane wave [44]. In Fig. 2.14, we present an abstract diagram with different patterns of spatial frequencies and orientations. In Fig. 2.15, we analyze this image with a Gabor filter with different scalings and orientation.

Not only scientists, but some painters have given great importance to a variable scale approach among their techniques. In Kinkead's painting *White Horse* (see Fig. 2.16), for example, large rectangular impasto patterns are used for the uniform background, and smaller scale rectangles for the horse face details.

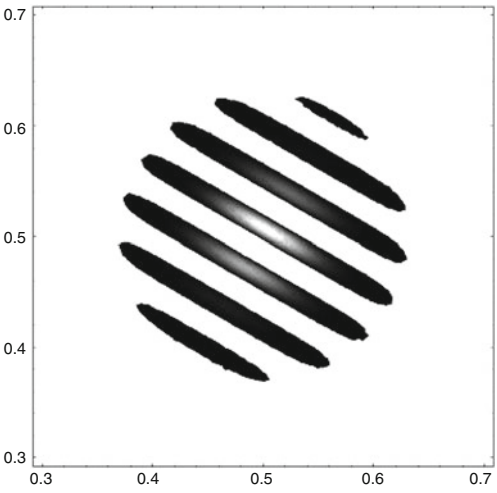


Fig. 2.13 Density plot of a Gabor filter function of wavelength $\lambda = 0.6$ m, spatial frequency $f = 0.1\ \mathrm{m}^{-1}$, and orientation $\theta = 60°$

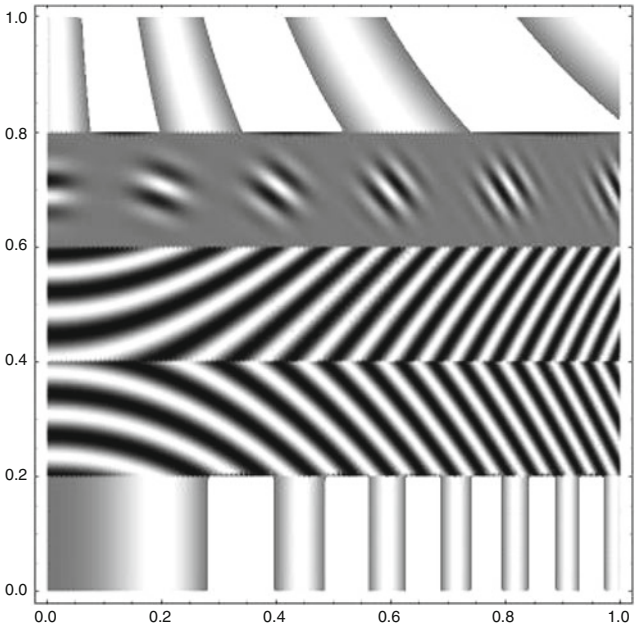


Fig. 2.14 Example of combined patterns using different scalings and orientations of a Gabor filter function

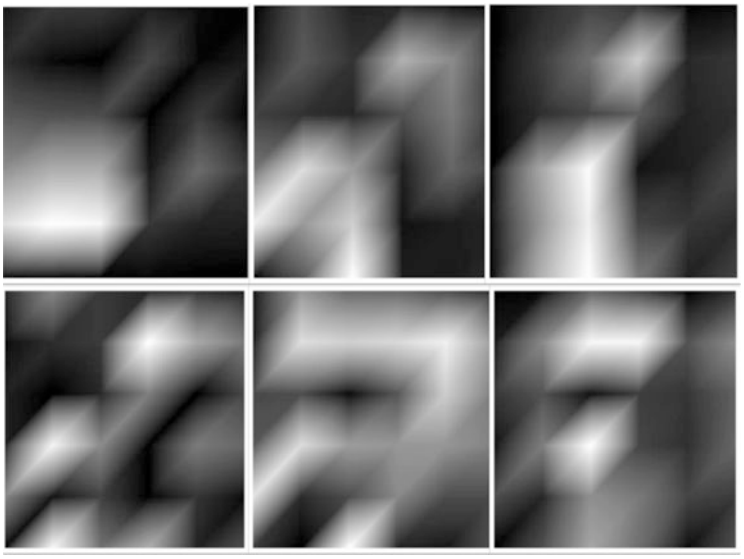


Fig. 2.15 MRA of the patterns in Fig. 2.14 using the Gabor filter of width $\lambda = 0.1$ m from Fig. 2.13. From the *upper left corner, clockwise*: $\theta = 0°, f = 0.2\ \text{m}^{-1}$, $\theta = -45°, f = 0.1\ \text{m}^{-1}$, $\theta = 90°, f = 0.1\ \text{m}^{-1}$, $\theta = -60°, f = 0.2\ \text{m}^{-1}$, $\theta =$ all angles, $f = 0.15\ \text{m}^{-1}$, $\theta = -45°$, $f = 0.2\ \text{m}^{-1}$

Fig. 2.16 Rebecca Kinkead, *White Horse (on white)* oil on wax and linen, 64 × 58 in, Gallery Mar, Park City, UT. Original painting ©Rebecca Kinkhead LLC, http://www.rebeccakinkead.com/, by courtesy of the Artist

2.5 Framed Versus Non-Framed

Imagine a peaceful view of the ocean on a calm and cloudless day at Daytona Beach, Florida. The view can be visually simplified as two parallel horizontal lines, one separating the beach from the ocean, and the other separating the ocean from the sky. While the solid beach and gaseous sky are motionless, the water is in a continuous, yet predictable state of motion. The sand and the sky have the same color and known pattern, and even the ocean's tidal elevation is a periodic and bounded function of time. The wave's wavelength and amplitude and their angle with respect to the shoreline are the same. Now visualize yourself inside a fine arts museum, while in front of you is a large painting with the same ocean landscape we have just imagined. Does this painting generate the same perceptions in our brain? Does it create the same feelings, thoughts, or state of mind?

The answer is: not quite, and we will try to analyze why. Obviously, there is a significant component of sensorial information which is missing in the case of the painting of the ocean, like the sound of waves, the breeze, the smell of salt water and seaweed, and the heat of the sun. All these are important, but the essential difference is that the painting is bounded, while the real scenery is not.

Fig. 2.17 One of Magritte's favorite themes, namely the 'window painting' and the 'painting within a painting', presents a cycle in which the viewer must choose one as real and the other as representation. The painter also uses bounds for the real scenery in order to eliminate all the differences. *The Human Condition*, 1945. René Magritte (Belgian, 1898–1967). Watercolor, crayon over graphite, ink, and gouache; 42.2 × 32.2 cm. Courtesy of The Cleveland Museum of Art. Bequest of Lockwood Thompson 1992.274

A frame sends a supplementary visual message to the brain, but no matter how smooth the frame is, and how well it fits the painting, it still introduces a significant discontinuity in the distribution of scales in the painting. In Fig. 2.17, we reproduce one of René Magritte's favorite themes, the cyclicity of real versus fabricated, as a possible artistic response to this question. The artist's answer is that a framed reality (the sea seen through the arcade) becomes indiscernible from any of its possible representations, one being for example, a frameless painting. The black

ball can easily roll in either direction of this indifferent balance between reality and representation.

2.5.1 The Necessity of a Frame

Matisse felt that "the four sides of a frame are among the most important parts of a picture", similar to attending a concert, where listeners can experience different effects according to the shape and dimensions of the hall (from *The Contingent Nature of any Act of Framing*, Henri Matisse 1943). While describing Clement Greenberg's criticism, Welchman claims the frame to be finally understood as the fitting condition of the shape of the canvas. He calls the frame a 'morphological impress' (John C. Welchman, *In and Around the Second Frame*, in [26]).

The natural world is believed to present all possible scales, patterns, and spatial frequencies with a $1/f$ type of distribution (see Sect. 2.6). The frame breaks this natural distribution because it cuts the continuous infinite extension of the landscape into a finite window. Any finite piece of an infinitely extended mathematical wave, or any compactly supported signal generates a totally different Fourier spectrum than a similar but infinitely extended signal. An example is shown in Fig. 2.46, which will be analyzed in the sections below.

The power spectrum of the compactly supported signal, i.e., the distribution of intensities of the frequencies contained in the signal, is wider since it contains a richness of lower frequencies and a long decaying tail of higher frequencies. It also holds weak resonances at some discrete frequencies with a potential relation between the scales in the signal and its size. The same phenomenon happens when we compare the real landscape, unbounded and infinitely extended in the retinal space, with a 2D framed image. What is bounded always has a richer spectrum. This phenomenon is a direct consequence of the relation, or a general principle, of uncertainty between the size of an original signal and the size of its integral transformation or representation.

In Sect. 2.6, we describe an interesting example of pattern selection (performed by the visual brain) from the painted image which will support our point of view regarding the importance of the frame for a painting. It is known from foveated vision that the angular distribution of the spatial resolution of the human visual system, i.e., the acuity to detect periodic patterns, decreases rapidly with the angle measured from the center of the crystalline-retina optical axis [8, 30]. In order to accommodate this behavior, the visual cortex resolves rather longer wavelength patterns towards the peripheral angles, as opposed to a richer spectrum of patterns towards the center of the visual axis. In other words, out of a multi-scale wavelength image, finer details are lost over longer wavelengths, by the visual brain, towards the boundaries of the perceived image. The information concentrated in finer details

(higher spatial wavelengths) is lost in the peripheral vision regions, while the information encoded in longer wavelengths is more equally detected over the whole visual field. A typical example of this effect is illustrated by the way one perceives the famous and hidden smile from Leonardo da Vinci's painting *Mona Lisa*, dating to 1503–1506 [8].

Coming back to the relation between image perception, pink noise, and Bayesian explanations, we note that the perception of the spatial frequency description of a painting is ambiguous, because it depends on the distance between the viewer and the painting, and also on the viewer's eye position (variable spatial resolution of retina with respect to the foveal point). In nature the spatial frequency distribution is similar to a $1/f$ noise distribution of frequencies, so a natural image is dominated by low frequencies. In art, some painters prefer to maintain a certain ambiguity in the visual interpretation. We refer back to da Vinci's *Mona Lisa*, where the ambiguity of the facial emotion results from the superposition of two conflicting sources of information in two different spatial frequency bands: the smile is visible in low spatial frequencies, while a neutral emotion appears in higher spatial frequencies [8]. Similarly, in *Slave Market with the Disappearing Bust of Voltaire*, by Dali (oil on canvas 1940) a disproportionate bust of the philosopher Voltaire can be noticed only if the viewer focuses on a larger and coarser scale, while alternatively, two small nun figures are seen when the viewer focuses visually on a finer scale [45]. Mamassian [9] points to another explanation for this kind of ambiguity, namely from the interpretation (or observation) of the shadows or dark surface material. Paintings have very different frequency distributions from natural images, and the convention on spatial frequency is a characteristic of their style. Another example is provided by the experiments of Intriligator and Cavanagh [46] showing that spatial frequency perception is equally a function of the focus of the attention point, the resolution limit of attentional selection, and the grain of attention, which are themselves inhomogeneous across the visual field.

In his 1970 article *No Thought Exists Without a Sustaining Support*, Mel Bochner (see, for example, [47]) emphasized that boundaries, or enclosures, are described 'conditions of positions' and hence reflexively reliant upon language. In his art installation *Theory of Boundaries*, presented in 1969–1970 at the National Gallery of Art, in the East Building Concourse Gallery 29F (Fig. 2.18), Bochner demonstrates the interest of visual arts in the physical and mechanical processes that involve their creation. Through his constructions, Bochner invites viewers to reflect on the most basic cognitive process involved in seeing the structural relations between objects. These painted frames (see Fig. 2.18) connect the space with the linguistic prepositions describing space. In our opinion, parts of his work accomplished in 1968–1970 raise purely mathematical questions, e.g., under what circumstances does the image of a boundary remain a boundary? In his essay based on an interview with Bochner, Kranjec [48] asks rhetorically whether some species-specific human abilities, like analogy and metaphor, are possible because of the deployment of those specific relational language prepositions, deeply grounded in relational cognitive domains, like space and time.

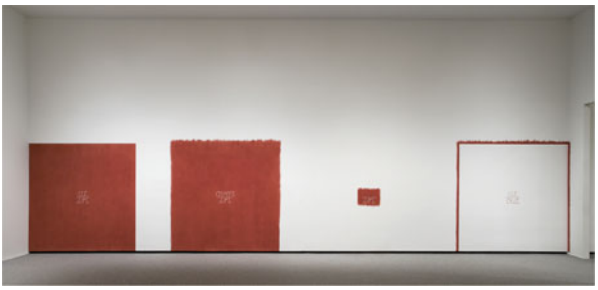


Fig. 2.18 Examples of art installations which point toward spatial, numerical, and linguistic themes, and relate (at least in an abstract way) the objects one to the other rather than to the objects themselves. The frames painted on the walls represent more than boundaries. They are co-boundaries between the space itself and the semantic of prepositions describing space, or even more abstract conceptual categories like emotions or morality. Mel Bochner, *Theory of Boundaries*, chalk on pigment on wall, 1970, size determined by installation. National Gallery of Art, Washington, D.C. ©Mel Bochner, http://www.melbochner.net/ by courtesy of the Artist

Artists show how they invest in the process and the structures of their art, and how viewers can emphasize the logic determining the relationships between the colored surface, the border, and the quality of enclosure in each of the four squares in the *Theory of Boundaries*, what Bochner calls a 'language fraction'. He explains the language fraction as the ratio between the existence of the (tangential) relationship of the image to the border, and the image position in regard to the sense of enclosure (enclosure considered as condition of position). Minimalists Robert Ryman, or Fred Sandback, for example, take this language to an extreme. In Ryman's *Midland* series (1976), he spins threads around four nails in the shape of a rectangle. Sandback hangs long parallel lengths of colored yarn from the ceiling. Each artist presents the frame as the spatial component of the overlap between the artwork and the environment where the work is presented. In this approach, the frame is part of the complete artistic concept rather than a formal edge or the mechanical bound. These artistic trends show parallels with some recent research results in empirical cognitive neuroscience and the semantics of space [48].

2.5.2 Framed Paintings of Canvases

The frame holding a painting separates our real world from the painter's imagined world. But paint a mirror reflecting our world and this neat division breaks down: many items depicted in the mirror should realistically be presented in the real world, but of course they are not. Clearly, artists can break the physics of mirrors and still display a convincing mirror. At this point, the viewer might ask, which rules are required to successfully depict mirrors and which rules are optional (see Fig. 2.19)?

Fig. 2.19 A perfect example of the artistic play between required rules and optional rules in depicting mirrors. Pierre Bonnard, *Dressing Table and Mirror*, oil on canvas (1913). Permission by courtesy of the © 2015 Artists Rights Society (ARS), New York/ADAGP, Paris

Cavanagh says "even very abstract painting can convey a striking sense of space and light, despite the remarkable deviations from realism" [49]. Cavanagh describes this characteristic by suggesting that our visual brain may use an 'alternative physics', simpler than physics, and a reduced model of reality to understand the world. A painting that gives an unhindered sense of the space and objects within it, despite physical impossibilities in the depiction, says something about the way our brain processes physics.

In paintings depicting water and glass, significant deviations from the laws of reflection and refraction are not noticed by the viewer, indicating that the visual brain only computes a small set of the possible physical properties of a transparent material when assessing whether or not a surface is transparent [49]. For example, even if there is no optical distortion of a lemon in a glass of water, the glass and the water appear convincingly transparent in Margaret Preston's *Implement Blue* (oil on canvas on paperboard, 1927, Art Gallery of New South Wales).

By following these liberal *physical licenses*, the artist can also use a frame to orient the viewer's perspective relative to the subject being depicted. The frame provides a metaphorical window into another world which is disconnected from the viewer in space and time. For example, Paul Klee, in his *Ad Marginem* (1930), paints precious, delicate vegetation, herbaceous umbels, and birds upside down on and around the frame, and by doing so, creates an unusual proliferation of objects repelling each other from the center. There is an unknown source of light, centered in his painting, but it is hidden by an eldritch central object. His technique "interrogates the modern subject through its frame", as Louis Marin would say [26]. The frame transforms a painting into a reflecting pool placed on the floor, similar to Pollock's style of drip painting, or like the framed frames in Frank Stella's *Gran Cairo* (1962), where we may represent a pyramid seen from the top. Sometimes the painting and the frame are one and the same thing, as in Middlebrook's painted maple planks (see Fig. 2.20). Middlebrook uses various approaches to consider the complex interplay between humankind and the environment. This ecological angle is achieved through various juxtapositions in natural and man-made materials, as well as in line and color.

2.5.3 Eliminating Frame Effects

The most important developments for the digital era are building faster computers, providing more memory with more portability, and creating a frameless and interactive display. Together with research on optimization of the shape of the windshield in automobiles, airplanes, and reading glasses, this topic is one of the most modern debates in visual science. In 2005, Pinhanez and Podlaseck published a paper in which they conclude from visual design theory that "a frame creates and indicates spatial disruptions" [50]. Their main hypothesis is that frameless displays connect with the surrounding environment and objects better than framed displays, by contextualizing the information presented within them.

Some impressionist painters used tricks like demolishing the illusion of perspective in order to better submerge their paintings into the same space as the viewer. Rauschenberg, Luis Sottil, and Johns enhanced the presence of their paintings by integrating physical objects into the canvas, while Kaprow (with his *Environment* or *Happenings*) or Burden (*Extreme Measures*) totally removed the visual frame, creating works that are sometimes indistinguishable from the physical reality of the viewer [50].

Fig. 2.20 Total diffusion and identification between the 'painting' and the frame: Jason Middlebrook's maple planks, *Breakthrough*, acrylic on maple plank (2012). Photograph by the author included with permission by the Artist

In one of his essays Kaprow observed that, just as visual phenomena establish a presence or embodiment, or at least the possibility of occurring in principle at the same time and place in the reality of the viewer, by frame elimination, their participatory status increases. He also notes that this is the present consensus among computer theorists.

A. Morton very clearly advocates the importance of framing for depth perception and the three-dimensionality illusion in his study *A New Frame of Mind* at the website of the *International Institute for Frame Study*, where he quotes a museum curator explaining a carved sixteenth-century Venetian frame for a fifteenth-century

Giovanni Bellini painting: "This frame actually adds to the illusion that you see Bellini's figures in three-dimensional space with the frame forming a wonderful window."

While advocating that virtual or digital frameless displays should be used in everyday life, Pinhanez and Podlaseck further suggested that frameless displays create a strong sensation of direction by projecting direction cues (arrows, footsteps) along already existing real directional channels. At the same time they argue that frameless displays suppress perspective and depth perception, because without a frame there are no more references to vanishing points, making it harder for the viewer to understand perspective.

An interesting effect of the biological vision function in humans is *boundary extension*. The viewer's memory of a possible extent of the borders tends to include additional scene information. In other words, after viewing a certain picture, subjects remember it as having a wider angle than it really had. The complementary effect, called *spatiotemporal boundary formation*, is the mental construction of illusory boundaries generated by consistent local changes. For example, if a certain detail is eliminated partially or totally, the viewer projects an imaginary and invisible wall that hides that detail.

The artist's need is to free the art work from the frame's rigidity. For example, the artwork of architect Frederick J. Kiesler is not a reaction of minimization of the importance of the boundaries, but rather the enzyme that frees the painting and dissolves it into the spectator's space.

Boundaries as Vacuum, Darkness, and Without Information

There are many examples where the boundary consists of darker regions, or shadows around the image. Such regions give 3D depth to an otherwise flat image. Of course, the region should be completely dark since otherwise the brain does not recognize it as a frontier. Good examples are still-life subjects presented in a dark room (see Fig. 2.21). Scientific studies on the perception of shadows support earlier discoveries made by painters.

Shadows below the nose, eyebrows, and chin define the depth of a face. If the painted shadows do not obey optical rules, the face loses its stable 3D appearance. Studies of lighting configurations support artists' understanding that, as long as shadows are perceived as the boundaries of a lit surface, they contribute to a realistically accurate artistic message. *Google video*, for example, has the feature of reading face shapes by adding designs and shadows in order to move the face.

Boundaries as Lines

In the real world, lines do not divide objects from their backgrounds. This raises the question as to why line drawings work so well for our visual brain? In conventional line drawings, the lines trace the contours characterizing a shape. More

Fig. 2.21 In some paintings the boundary consists of darker regions around a central image, enhancing the 3D depth effect. *Lamp Study*, watercolor (2006). Original painting ©Delia Krimmel, by courtesy of the Artist

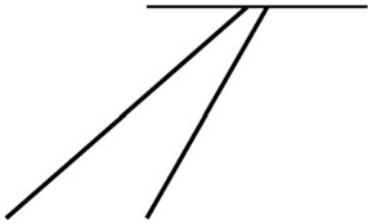


Fig. 2.22 Three lines do not necessarily create depth perception. The image can instead be interpreted as a symbol

important than simple lines are lines that cross. Figures 2.22, 2.23, and 2.24 present three situations where the same intersection of three simple straight lines provides different depth perceptions: non-framed lines (no depth perception), framed lines but line not extended to the frame (weak depth perception), and line intercepted by the frame (best depth illusions). In the study *The Art of Transparency* [51], Sayim and Cavanagh describe experiments by Metelli that have shown how these crossings or *X-junctions* are critical cues for the successful depiction of transparency. When the X-junctions are misaligned, the impression of transparency is lost (see Fig. 2.25). Actually, in this painting, Hopper masters another artistic feature of the use of boundary: the *frame-within-a-frame* technique.

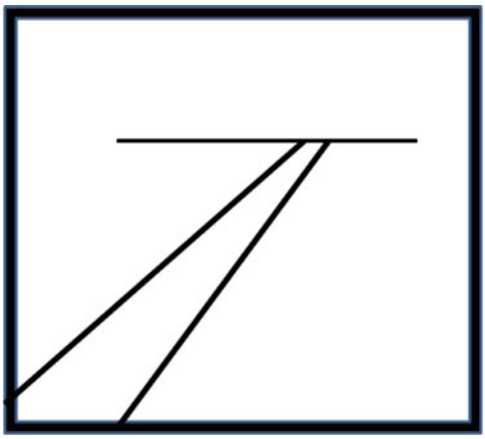


Fig. 2.23 The same line structure as in Fig. 2.22, but in a frame. There is some sense of depth

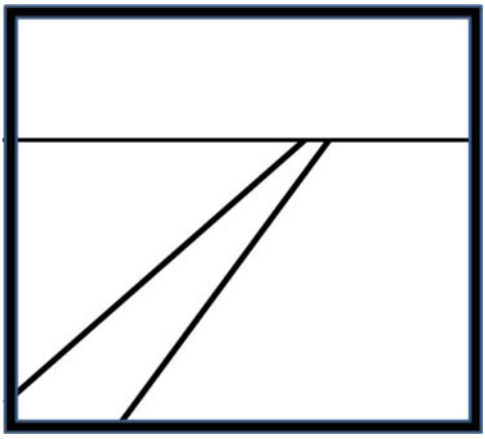


Fig. 2.24 Frame in action: the same drawing as in Fig. 2.23 with the *horizontal line* extended to the edge of the frame. We now experience the greatest depth perception

Boundaries as Smooth Continuations of the Image

Here is my experience of seeing an original impressionist painting for the first time. I was in London for collaborative work on thermonuclear plasma, and during a late and cold evening I decided to take a break and enjoy some fine art. I visited the impressionist paintings at the National Gallery of Art, and I found myself face-to-face with Monet's *Water-Lily Pond* (see Fig. 2.27). After a very long day of staring at oscilloscope screens in my lab, I found myself unable to make sense of the painting's visual information. I noticed the arc of a bridge, but under the bridge, I simply saw a random sea of colors without any correlation, shape, or realism. What I saw in these first few moments was just a combination of parts without global structure, perspective, or depth, like the detail presented in Fig. 2.26. It was only after some time that my brain suddenly triggered, and instantly processed all the information. I saw the water, the bushes and grass by the shore, the reflections of the willows, the water lilies, the ripples . . . in short everything (look at Fig. 2.27 again). It took some

Fig. 2.25 Crossings or X-junctions are critical cues for the depiction of transparency. The strength of the effect depends on the alignment of the X-junctions. An example is provided by the dark background window across the street: the blue-white triangle. It enhances the perception of a view through three panes of glass, one of Hopper's special skills. Edward Hopper , *Nighthawk*, 1942, oil on canvas. Reproduced with permission of the Art Institute of Chicago

Fig. 2.26 Isolating a certain area with constant distribution of pattern size from *Water-Lily Pound* by Claude Monet (1899, oil on canvas) prevents the conscious image-recognition centers from firing, and depth perception is weak. Only by including the whole distribution of patterns and scales does the landscape become recognizable and acquire depth, as we observe in the complete painting (see Fig. 2.27). Reproduced with permission of the National Gallery, London

time for my visual brain to integrate, compare, and extrapolate, and whatever other operations it did, until it finally found the reality matching the painting.

In modern technology, through enhanced perceptual linkage like 3DTV, the boundaries between observer space and display space become blurred, supporting an illusion of non-mediation [52]. IJsselsteijna et al. have introduced the concept of *presence* as a measurement of the perceptual linkage between the observer and the mediated environment. The weaker the illusion of non-mediation, the greater the *presence* factor. In experiments where right brain activity was observed,

Fig. 2.27 The impressionist painting connects better to emotional centers because its patchwork of brush strokes and mottled coloring distract conscious vision [49, 52]. The pattern of reflection on water does not have to match the actual scene around it, only the average properties of natural scenes [49]. This is why, in the detail from Fig. 2.26 of the same painting, it is difficult to gather what it represents. *Water-Lily Pond* by Claude Monet (1899, oil on canvas 140 × 150 cm). Reproduced with permission of the National Gallery, London

researchers found that people presented with faces expressing fear respond strongly to a blurry version of the same faces. Brain areas responsible for face recognition respond weakly to blurry faces, but strongly to detailed faces. IJsselsteijna and co-authors believe that impressionist painters connect better to emotional centers than to conscious image-recognition areas "because the unrealistic patchwork of brush strokes and mottled coloring distract conscious vision".

Elimination of Boundaries by Expanding the Field of the Image

In comparison to painting or photography, our awareness of offscreen space in the cinema exists largely through sound, because it gives offscreen events the same correlation as onscreen events [53]. In contrast to a painting or photograph, a projected image does not have a physical surface: only the screen on which the

image is projected has a surface. The movement of the projected moving image lends a particular kind of presence to that surfaceless image. Because the image is moving, objects within appear to enter and leave the frame of the image.

Offscreen space can be understood as an extension of the dramatic space of the image, and it is supported by the soundtrack, which emphasizes both the presence of the image and the continuum between onscreen and offscreen space. A precondition for projective illusion is thus established.

Cavanagh considers that almost anything can be put in a reflection as long as it is bright and curves appropriately for the reflecting surface curvature [49]. A survey of medieval, Flemish, and modern paintings reveals a number of extraneous items in reflections that should not logically be where they are painted.

We close this section by presenting Jeffry Dell's artistic solution for transcending the boundaries of a visual artwork, selected from his *Boundary Extension* solo exhibition of screen prints in the Art Palace Gallery in Houston, TX 2015. The term 'boundary extension' comes from a cognitive phenomenon in which the observer perceives the frame of an image as being stretched out, exceeding the art work, and apparently including information not derived from sensory input. This cognitive phenomenon is usually revealed by a subsequent memory error when we confidently misremember the extended scene instead of the original. Dell's work relies on the viewer's ability to coherently read the principles of occlusion, physics, and memory. When Dell applies the concept of boundary extension to an ephemeral motif like spirals, bows (Fig. 2.28), screens, and stripes (Fig. 2.29), the vivid palette of color captured by sharp edges "adopts a striking split-personality'. Dell's work teaches the viewer how perception, even lacking certainty, contains an abundance of emotional aptitude. Yet, when one sees his spirals and stripes, one cannot help but think of multi-valued complex functions and branch cuts.

Psychological experiments show [54] that a reliable boundary extension effect is expected in about 11% of adults, that it depends on the test size, and is accentuated in people suffering from PDD disorders, like Asperger's syndrome. It has also been shown that the boundary extension in memory for a picture occurs when that picture's boundary is understood as limiting or truncating a continuous view that would otherwise extend beyond the frame of the picture.

Boundary extension is believed to be an automatic brain process, because when participants are explicitly instructed about it prior to experimental trials, the effect is not diminished or eliminated [54]. This result supports the idea that frames are somehow ontogenetically and phylogenetically part of our natural habitat. It is also believed, and Dell's artwork supports this belief in a creative way, that boundary extension results from anticipatory extrapolation of what might be encountered in the next moment of time following the moment of time when the image was taken or created. In this regard, the frame breaks the original image symmetry and deforms this symmetry towards the direction of an anticipated motion, or fall, or any potential future action. For example, a photograph of a real ball taken against some simple and uniform natural background, framed by a round boundary will still be remembered, through the boundary extension process, as elongated vertically downwards where the viewer expects the ball to fall due to gravity.

Fig. 2.28 Jeffrey Dell, *Flat Bow Orange* 2015, 3, × 23 in, screenprint. Reproduced with permission of the Artist

2.5.4 Frameless: Greek Pottery, Vermeer, and Feynman

The medium of painting imposes certain inherent limits on innovation, and the artist's work depends to a large extent on the knowledge and expectations of his or her viewers and fellow painters. Painting on vases and pottery is technically a transition between painting on flat surfaces and sculpture (see Fig. 2.30). Some attempts at visual narrative appear in Mycenaean vase painting from the late Bronze Age, although human subjects are rarely depicted.

Painted Greek vases are known from the second millennium B.C. until near the end of the first century B.C. Greek vase painters were greatly influenced in their subjects by epic poetry and by oratory and lyric versions of stories, tragedy, and local folklore. Lowenstam shows [55] that the painters not only depicted newer

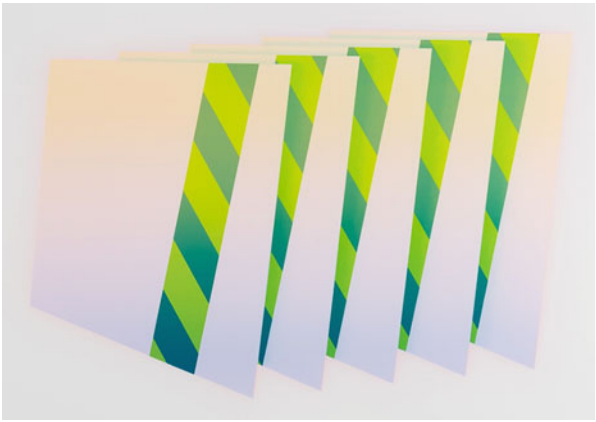


Fig. 2.29 Jeffrey Dell, *Green Stripe Array* 2015, 23 × 34, screenprint. Reproduced with permission of the Artist

Fig. 2.30 Khantaros is an ancient Greek pottery cup using the red-figure technique. Here we have an interesting combination between round surface media and sculpture. ©Ken Backer/dreamstime

versions of stories, but combined all of the various sources in their expositions. For example, Lowenstam found that the Iliad and Odyssey myths are presented in Greek

Fig. 2.31 Amphora for wine or oil, with red-figure painting, depicting the Greek goddess Aphrodite with the mirror and winged Eros. ©elmm/shutterstock

vases before the end of the sixth century B.C. in multiple and distinct versions, some of them quite different from the well known written versions.

An artist's creativity goes beyond the tribute to specific stories, creating fresh aspects or even new subjects not originally depicted. Figures 2.30 and 2.31 (sixth to third centuries B.C.) show the change of style in vase painting, and the transition from archaic Greek pottery painting to Athenian vase painting. According to von Bothmer [56], this technique gave the vase its maximum potential of expression. The curvature of the vase is enhanced by the extra 3D visual cue, namely, light. Moreover, the painted faces can be seen from greater distances. With the red-figure technique, the contour of the vase is less eroded by decoration and the vase's spherical attribute is re-established, with proper emphasis on the profile silhouette of the vase. These re-established vase paintings prove that artists acknowledged the presence of unboundedness on a vase as opposed to flat painting, and then exploited and enhanced it.

The special medium offered by the unboundedness of the vase has had a strong influence on the subjects to be painted. As Bazant says in [57], "in Athenian vase painting [...] there are numerous so called scenes of reality [...] scenes of myth,

evocations of distant past, social fantasies, political propaganda, etc. Nevertheless, the conviction that sixth century B.C. Athenians started to paint what they saw around them is holding astonishingly well." We detect here once again the same starting point for abstract vase painting and geometric discovery. We can see what would happen to a painted vase if it were broken up 'artistically'. Its geometric infiniteness and unboundedness effects can actually be enhanced in this way, as we can see in the modern ceramic art experiment of Eberle, shown in Fig. 2.32.

The ancient Greeks were able to take advantage of the spatiality of the vase in paintings because they had the concept of geometry, in contrast to the Mayas, for example, whose pottery paintings lack spatiality and geometry. The Maya didn't have the concept of dimension or understand the relations between length and width. In order to represent 3D objects in 2D media, the Maya repeatedly used artificial tricks with a form of hieroglyphical structure, rather than visual illusions [58]. Archaic Greek vase painting, and especially the red-figure technique, ranks so well because it developed simultaneously with Greek mathematics and geometry, from roughly the seventh century B.C. to the fourth century A.D.

According to Jünger's classification [57], art describes two types of memory: communicative and cultural. The first type, the 'communicative', is casual, infinitely increasing, and diversifying in time, a function of the individual relative experience.

Fig. 2.32 Enhancement of the unboundedness effects of archaic vase painting when multi-dimensionality is brought in through modern ceramic experiments. *Firehat and Fool Moon*, 2002, porcelain © Edward S. Eberle. Photo courtesy of the artist

It is mainly inspired by and related to sympotic Greek events, like a newspaper. The second type, the 'cultural', is classical, traditional, and tends to freeze into structures and principles. The two different artistic memories are two different codes with settings. Communicative memory is neither stored or communicated in a correct technical language, nor intended as an educational tool. As Bazant says, a pedagogical ideal would bring an asymmetry (student/teacher) that would cut its two-way flow of communication. Recollections cannot be taught, they are co-equal and plurally expandable. Athenian vase painting after the sixth century contains these communicative memories: the vase is infinite and bi-directional, reversible, like these recollections. An Athenian vase could be compared to a photograph taken today with your cellular phone. In the many ways a cell phone is useful, so is the Athenian vase: it is a decoration, a drinking device, an object of art, a storage vessel, and a recollection of stories. Greek pottery painting, including sympotic poetry, created a niche in the art world between sculpture, painting, poetry, drama, etc. [59]. At the end of his article, Neer [60] stresses a basic point about Athenian vase painting: it is intimately related to material life, expressed in its own complex language.

Let us take the 'frame/no frame' dilemma to the seventeenth century and study Vermeer's *Woman in Blue Reading a Letter* (see Fig. 2.33). This painting is proof of how dramatic it can be when an image is robbed of its frame's power of focus and comment. In his article *Posed Spaces: Framing in the Age of World Picture* [26], John Gillies argues that the painting's depiction of a world map is truncated by the physical frame and serves to dim and dispel the theatrical effect arising from the fact that there is a map in the painting.

The 1960s saw the development of the nouveau realism movement, minimalism (Mondrian), science fiction, American conceptualism, and graffiti, all trying to exceed or dissolve the boundaries of the classical artistic subject. We witness the relinquishment of limits by Kaprow and Beuys, who profess an anthropological practice and understanding of art by expanding the margins of framing. Attenuation, dissolution, formlessness, and relational disfiguration are expressions of the general 'breaking-the-frame' artistic trend of the modern painting, as in Fig. 2.34. Summing up Greenberg's positions on these trends, Welchman writes: "The frame is a virus in the machine of formalism, a sort of double agent functioning as a necessary part of the system, but also as the gateway to its dissolution." Mel Bochner [47], reviewing the 1966 *Primary Structures* exhibition, claims that "art need no longer pretend to be about life. Art is, after all, nothing [and] invisibility is an object" (see Welchman's essay in [26]). We may want to correlate these revolutionary trends in the visual arts of the 1960s and 70s with the great mathematical developments in understanding space itself, viz., the final proof of the four-color theorem, Donaldson's breakthrough work on the particularly unique space $\mathbb{R}^4$ (in which we live after all), knot theory, sheaf theory, graph theory, singularity theory, and chaos theory (all about space and its properties). In theoretical physics, at about the same time, we had the trigger for unification of field interactions, a deeper understanding of higher dimensional spaces and gauge theories (Wigner, Schwinger, Feynman, Gell-Mann). Maybe the artists of the 1960s did not want to move away from life and

Fig. 2.33 The painting's depiction of a world map is truncated by the physical frame and dims the theatricality effect of the map's existence in the painting. Johannes Vermeer, *Woman in Blue Reading a Letter* (about 1663–1664, oil on canvas, 18 5/16 × 15 3/8 in). Public Domain: https://commons.wikimedia.org/wiki/File:Vermeer,_Johannes_-_Woman_reading_a_letter_-_ca._1662-1663.jpg

nature painting, but began to understand the implications of pure space, the vacuum, and spontaneous breaking of symmetry, and tried to represent these concepts in their own language.

Alloway [61] mentions that "there is all the difference in the world between a compact zone, such as a painting establishes, and a boundless field, the continuous space of the world". What is this affirmation but an artistic expression of the difference between a compact set and an unbounded set, the differences in their resonance frequencies, the principle of uncertainty for Fourier series? Aesthetics figured out, finally, that the existence of a frame actually implies the compactness of the 2D artwork together with all its topological, geometrical, and functional consequences, and hence all its signal perception and neurological consequences, too.

Fig. 2.34 Innovative use of frames. Margret Hofheinz-Döring, *Landschaft mit Berg und Kirche* (1983, oil, WV-Nr.3799). Public domain: https://commons.wikimedia.org/wiki/File:Landschaft_mit_Berg_und_Kirche,_Margret_Hofheinz-D%C3%B6ring,_%C3%96l,_1983_%28WV-Nr.3799%29.JPG

2.6 Perception of Image Boundaries

The presence of boundaries and frames surrounding images influences how we perceive orientation and shape. This sounds plausible because it is known that we perceive the visual world primarily by reading barcodes [62]. When our eyes read a framed painting and move from one side to the other in different directions, our neurons detect and recognize lines, orientation, and patterns, and most importantly, they feel the break of continuity from the inside space condensed by the frame. The frame generates a very localized visual signal which, by the uncertainty principle, cannot be accommodated in our visual brain by a narrow Fourier spectrum, so it generates a wide range of extra frequencies and scale. All these reactions are gathered under the common feeling and perception of boundary.

W. Kemp says that "the frame is the necessary condition for structural perception being possible". In Christian art of the late antiquity and the Middle Ages, the role of the frame was neither to create an excerpt, as in the cinema, nor to constitute

the esthetic border of a picture, but rather to provide an organizational operator for visual material. It holds the elements together and guarantees their connectivity. Following Nicolas Poussin's ideology (1594–1665), the most important role of the frame is in the construction of meaning, namely "the rays of the eye are focused and do not become distracted by the impression of other neighboring objects". For Poussin, as Duro comments [26], the frame is both a conceptual marker of limits and an aid to representation. The frame's significance is twofold: it actively participates in the construction by linking the representations of the objects in a non-contradictory way, while on the other hand it offers the artist a way to prove competence in handling perspective and other technical principles (see Fig. 2.35). According to the principles of the French Royal Academy of Painting and Sculpture (founded 1648), academic painting should be considered deeply related to the sciences, even to mathematics, through perspective, and the frame of a painting would very much play a catalytic role. Frames become a condition of intelligibility and create a useful pictorial vocabulary when similarity or likeness between symbols can create confusion. In this context, Rico Franses [26] gives the

Fig. 2.35 Importance of frames as an organizational operator of the visual field in ceiling murals. Charles Le Brun, ca. 1860, Salon de Venus, Versailles. Public domain: https://commons.wikimedia.org/wiki/File:Salon_de_V%C3%A9nus.jpg

example represented by the framed single-episode scenes in Giotto's fresco, a ploy used more generally in the Renaissance and beyond (see, for example, Fig. 2.8).

2.6.1 Illusions and Frames

Hebert et al. [63], for example, conducted experiments on the effect of such frames on the perception of bilateral symmetry in dot patterns. The rod-and-frame illusion and the processes of pattern perception and recognition are examples of real situations where frames provide extra cues to the images. Boundaries and frames also influence the way we perceive symmetries, because they supply a visible reference that directs the selection of the symmetry axis. Hebert and his co-authors noticed that the symmetry detection effect of the frame actually slows down the process of detecting symmetry.

There is experimental evidence that shape information within a given image is processed by working from the outside, from frame to center. It is also true that the opposite effect can counterbalance: a large amount of factual background information is usually detected before a central target. If we order the detection steps, the first is detection of a general shape, followed by proportions and orientation of the frame, then a dominant background (if present), then details and their symmetry.

The visual geometric illusion known as the HV (horizontal–vertical) illusion, which is an overestimate of the vertical line relative to a horizontal line of equal length, is known to have an explanation resembling framing [64]. It relies on the observation that a line enclosed in a large frame appears to be shorter than a line of equal length in a small frame. This effect is explained theoretically through the elliptic shape of the visual field (retinal anisotropy): the ends of the vertical line are closer to the visual field boundary than the ends of the horizontal line of the same length. Mamassian also shows [65] that the HV illusion is strongly affected by the figure orientation in the image plane, and hence the orientation with respect to its prior frame. Williams and Enns actually proposed that the HV illusion is generated by multiple causes, among which the frame theory is the most important. The frame, according to Enns, plays a game of contrast/assimilation with the viewer's eye.

The interdependence between vision and touch or motion may become enhanced by the existence of a frame as partially belonging to the landscape, and partially belonging to the spectator's real and tangible world. When considering more than one sense at a time, Kennett et al. show [66] how modern research on sensory perception reveals a great interdependence between vision, touch, and body motion.

In his FACADE theory of 3D vision from a neural network point of view [67], Grossberg argues that long wavelengths (low spatial frequencies) selectively process nearby objects, while short wavelengths selectively process more remote objects. In his theory, this happens because of the so-called size–disparity correlation, viz., retinal images increase in size and disparity as the distance to the object decreases.

It has been hypothesized that short wavelength patterns may appear closer and may be fused, while long wavelength patterns appear more distant and prominent (Weisstein effect or the size–disparity correlation, Fig. 2.36). For example, painters can use luminance and chrominance, not necessarily for matching natural conditions, but just to produce special effects. Patterns equiluminant with the background, even of different color, may create difficulties in identifying the boundaries and position of the patterns, and can be used to suggest motion, for example.

The explanation for the Weisstein effect may stem from the way the visual brain perceives a spherical patterned surface (see Fig. 2.37): a pattern closer to the viewer appears larger in size, while further away from the viewer, it looks smaller. Moreover, while shrinking the widths of patterns, the visual brain also notices that this domain is close to the boundaries, so it tends to associate narrower patterns with

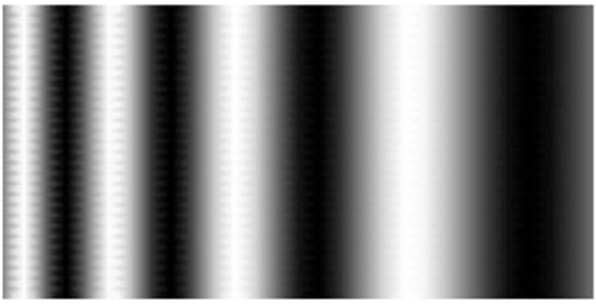


Fig. 2.36 Longer wavelengths appear closer in a 2D monocular picture, in contrast to shorter wavelengths. Density plot of a sine function with linearly increasing frequency

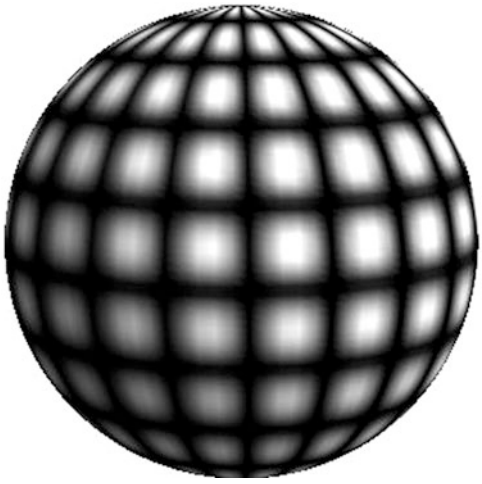


Fig. 2.37 Tiled spherical surface showing that wider patterns are closest and narrower wavelengths are related to the boundaries of the shape in our visual experience. Boundaries are chirped

Fig. 2.38 Spatial chirp generated by a density plot of a function $\sin^8\left[(1+x)x\right]$, with a linear increase in frequency from *right* to *left*. The illusion of wider patterns being closer persists

Fig. 2.39 Spatial chirp generated by the density plot for a cnoidal wave function with linearly increasing frequency. The illusion of a wider pattern becoming closer is not as powerful, but rather it is superseded by a depth perception in three dimensions. The stripes now look like white bars

boundaries of objects, and vice-versa. The effect appears to be independent of the speed of variation, the slope of the wavelength, or the type of luminosity oscillations (see Figs. 2.38 and 2.39).

In contrast to this property, another experiment demonstrates that, if regions filled with relatively short wavelength patterns are adjacent to regions containing longer wavelength patterns (see Figs. 2.40 and 2.41), the regions with the shorter wavelength appear closer in depth than those containing the longer wavelength.

Grossberg's explanation of the Weisstein effect can be related to how we perceive boundaries, especially when they are filled with patterns. The shorter wavelength patterns have higher values for the luminance gradients. Higher luminance gradients excite wider receptive fields, and the vertically oriented complex cells (which are activated by shorter wavelengths) inhibit the neighboring vertically oriented hypercomplex cells (the ones activated by long wavelengths).

As a consequence, gaps begin to form around the short wavelength stripes, giving the sensation of edges and boundaries, that is, a "boundary web of form-sensitive boundary activations".

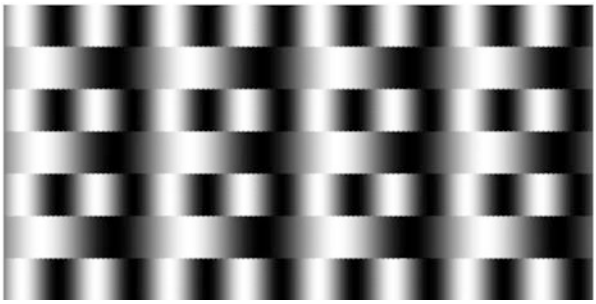


Fig. 2.40 Alternating stripes of two-wavelength patterns for a sine function. According to Grossberg's theory, stripes with higher luminance gradients should appear closer, even if the wavelength is the same. In reality, any stripe looks closer if we look directly at it

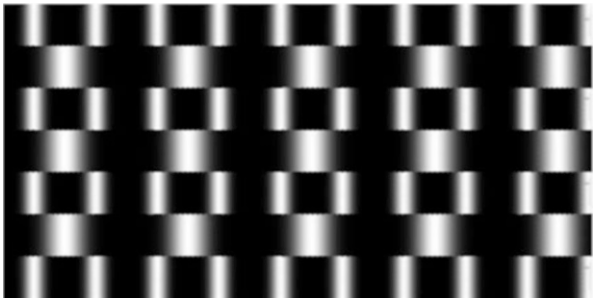


Fig. 2.41 Alternating stripes of two-wavelength patterns for a cnoidal function. None of the stripes stand out unless one looks for and focuses on them

This is the same process that makes us understand that the tiled spherical surface ends and has a boundary separating it from the background where the pattern becomes denser and richer in short wavelengths.

However, if we look carefully at the alternating two-wavelength stripes in Figs. 2.40 and 2.41, we notice a different optical illusion than the one described and explained by Grossberg. Each type of stripe appears closer to the viewer if looked at directly. In other words, independently of being shorter or longer wavelength patterns, the different parallel stripes which we focus on will always seem closer to us. This effect is an example of an older 'figure-ground' paradigm. Hence the explanation in the foveal type of vision: the lateral field always perceives larger patterns and tends to be blind to narrower patterns. In addition, from attention experiments [46], the resolution limit of attentional selection is inhomogeneous across the visual field. The limit is scaled with eccentricity, coarser in the upper visual field, and coarser along radial lines from fixation. Consequently, the visual brain receives the signal that the stripe we are looking at has the narrowest pattern, so it should be prominent in the image and closer to us, and because its boundaries have shorter wavelengths (higher frequency patterns), it appears closer. This effect

becomes more obvious in Fig. 2.41. It is also interesting to mention that while moving the eyes and attention from one type of stripes to another, and unconsciously trying to understand the images it "feels" like a discontinuity, or a sort of a phase transition in the perception process.

It is somewhat surprising that there is no mention in the literature of as simple and natural a visual occurrence of the $1/f$ pattern distribution as the surface of a sphere. Consider a spherical surface of radius R and center O, observed by a monocular vision so that the visual axis points towards the center of the sphere (see Fig. 2.42). The spherical surface is uniformly tiled with constant patterns, like those in Figs. 2.37 or 2.43, and consists of uniformly distributed patterns on the surface, all elementary cells having the same angular measure $\mathrm{d}\varphi$ on the surface. We choose one such spherical square denoted MN, as in Fig. 2.43, shifted at angle φ from the visual axis $\boldsymbol{r}_E = \boldsymbol{OE}$. In the visual field, represented in the figure as a circle of radius E (the eye being placed at distance $d = |OE|$ from the center of the sphere), this spherical rectangle has a width $\mathrm{d}\lambda = AB$. In the disk shown to the right of the figure, we present an example of a circular crown of such equal patterns of constant width $\mathrm{d}\lambda = |AB|$, as seen from the eye E. The width of the crown is the pattern width, and its radius is $\rho = R\sin\varphi$. When the angle φ scans the visible surface of the sphere from 0 to $\pi/2$, we see different scales of patterns, from the original

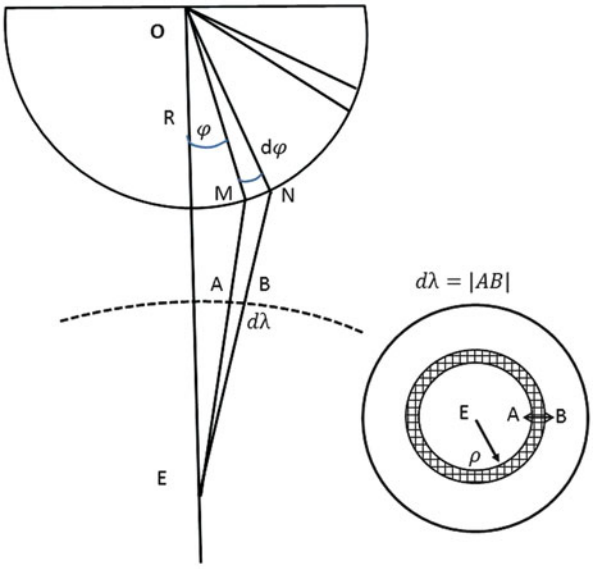


Fig. 2.42 Uniformly tiled ($\mathrm{d}\varphi = \text{const.}$, $\mathrm{d}\lambda_{\text{sphere}} = |MN| = R\mathrm{d}\varphi = \text{const.}$) sphere of center O and radius R, observed from a distant point E. The visual field is represented through a *dashed circle* of center E, and it is also shown to the *right* of the figure, where the sphere image is a disk of center E. The patterns appear to have a different wavelength function of their angular (φ) position on the sphere. For example, those perceived in the circular crown in the *right-hand* image have the same apparent wavelength $\mathrm{d}\lambda = |AB|$, where $\rho = R\sin\varphi$, and φ runs from 0 to $\pi/2$

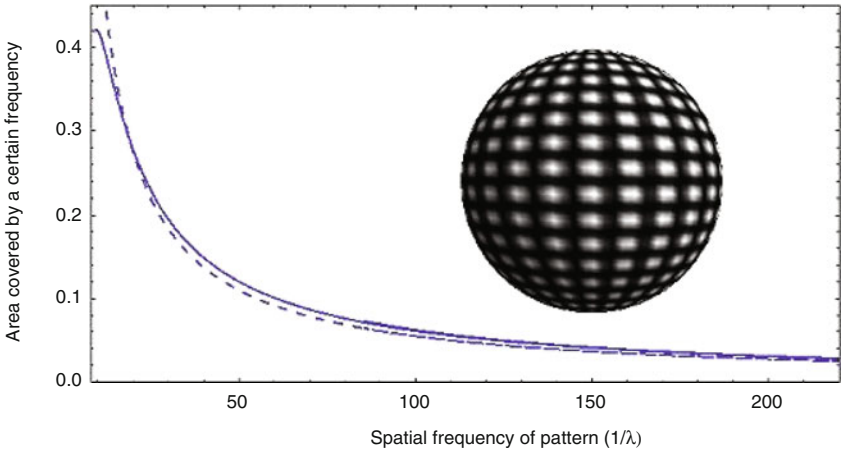


Fig. 2.43 Spectral power distribution versus frequency as perceived in monocular vision when gazing at a uniformly tiled and uniformly lit sphere (*see the inset*). For most of the spectrum, the distribution is a pure $1/f$ pink noise distribution

pattern size directly punctured by the visual axis, down to zero pattern size towards the boundary of the sphere.

We calculate the density of the distribution of patterns of different spatial frequencies f ($f = 1/\lambda$ with wavelength $\mathrm{d}\lambda$) on the apparent spherical surface as they appear to the eye. Then we plot this power spectrum distribution versus the spatial frequency, or in other words, we plot the number of patterns of a given spatial frequency f versus the spatial frequency. This plot shows the visual power spectrum of the perceived patterns and their scales as seen by an eye in monocular vision gazing centrally at the uniformly tiled sphere. We have

$$\mathrm{d}\lambda = |AB| = \Big\| (\boldsymbol{r}_N - \boldsymbol{r}_M) - \big[(\boldsymbol{r}_N - \boldsymbol{r}_M)\cdot\boldsymbol{e}\big]\boldsymbol{e} \Big\| \,,$$

where

$$\boldsymbol{e}(\varphi) = \frac{\boldsymbol{r}_M - \boldsymbol{r}_E}{\|\boldsymbol{r}_M - \boldsymbol{r}_E\|}$$

is the unit vector along the line of sight, with $\boldsymbol{E} = OE$. It follows that

$$\mathrm{d}\lambda = \frac{2R\left[R\cos\dfrac{\mathrm{d}\varphi}{2} - d\cos\left(\varphi + \dfrac{\mathrm{d}\varphi}{2}\right)\right]\sin\dfrac{\mathrm{d}\varphi}{2}}{\sqrt{d^2 + R^2 - 2dR\cos(\varphi + \mathrm{d}\varphi)}} \,. \tag{2.2}$$

All the patterns of perceived wavelength $\mathrm{d}\lambda$ form a circular crown in the visual field of area

$$\mathrm{d}A = 2\pi\rho\mathrm{d}\lambda \ ,$$

and we hypothesize that the visual power or surface density of these patterns is proportional to this area, i.e., $\mathrm{d}P \sim \mathrm{d}A$, where we do not take into account the Lambertian reflection law (the cosine law of decrease of the light intensity with the angle between the line of sight and the normal to the surface), because we are not concerned with the light intensity here, but rather with the number of patterns of a certain perceived width. In the following, we study the dependence

$$\mathrm{d}P = \mathrm{d}A(f) = 2\pi R\frac{\sin\varphi}{f} = 2\pi R\frac{\sin\varphi}{1/\mathrm{d}\lambda} \ ,$$

as a function of $f = 1/\mathrm{d}\lambda$. The graph of this function, which represents the spectral power distribution versus frequency of perceived patterns, is presented in Fig. 2.43. The result is surprising, because the distribution obtained theoretically is exactly the celebrated $1/f$ type of pink noise distribution. As proof we note that, from (2.2), in the limit $\mathrm{d}\varphi \to 0$, $d \to R$, which is the close view limit, we have the behavior

$$\sin\varphi \sim 1 - O\left(\frac{1}{f}\right) \ ,$$

and in the limit $\mathrm{d}\varphi \to 0$, $R/d \to 0$, which is the remote view limit, we also have the behavior

$$\sin\varphi \sim 1 - O\left(\frac{1}{f}\right) \ .$$

Consequently, in both limits, and hence in the whole visual range, the expression for the spectral power can be approximated very well by

$$\mathrm{d}P \sim \frac{\text{const.}}{f} \ ,$$

which is exactly the power distribution for $1/f$ pink noise.

This kind of noise naturally occurs in many physical, meteorological, astrophysical, biological (heart beats, neural activity, DNA sequences), and economic systems (long-term memory effect). It can be humanly generated in sandpile models, or it can be related to a mathematical convergence theorem for statistical processes characterized by a variance-to-mean power law. More importantly for our study, pink noise describes the statistical structure of many natural images [68]. Mamassian's explanation [9] using the scale approach to the visual brain is based on the fact that, in the visible world, the pattern scales, also called 'spatial

frequencies' by psychologists, neuroscientists, and artists, are available with a pink noise type of distribution. In this case, natural images are dominated by large spatial scales (coarse blobs), while smaller scales (finer details) are gradually less present. It is likely that the human visual system has adapted to these statistics of the natural environment and that the $1/f$ spatial frequency distribution is the observer's expected outcome when he/she looks at an image.

We also mention an artistic practice of locally altering the contrast in paintings in order to enhance depth perception and better separate objects of different depths. This is especially visible in the art of Dali, Picasso, Seurat, and Matisse. The success of such an artificial procedure (or such an artistic convention, to eliminate ambiguities induced by the flatness of the painting, as Mamassian would explain it) was recently demonstrated by Luft et al. [69] using their computational method to achieve special effects for images that contain depth information ('depth darkening'). Basically, Luft and colleagues use a depth buffer, and on top of it apply a toon shading, and a haloed contour, which in the end enhances the local contrast and the depth perception. This procedure adds a high spatial frequency component to the image near the boundaries of objects placed at different depths. By this means, the distribution of spatial frequencies in the image is closer to the pink noise power law for frequencies, which is known to play a role as a depth cue. The procedure can be seen working in a practical way in Brischler's paintings of geometric abstraction, and color fields, as shown, for example, in Fig. 2.44.

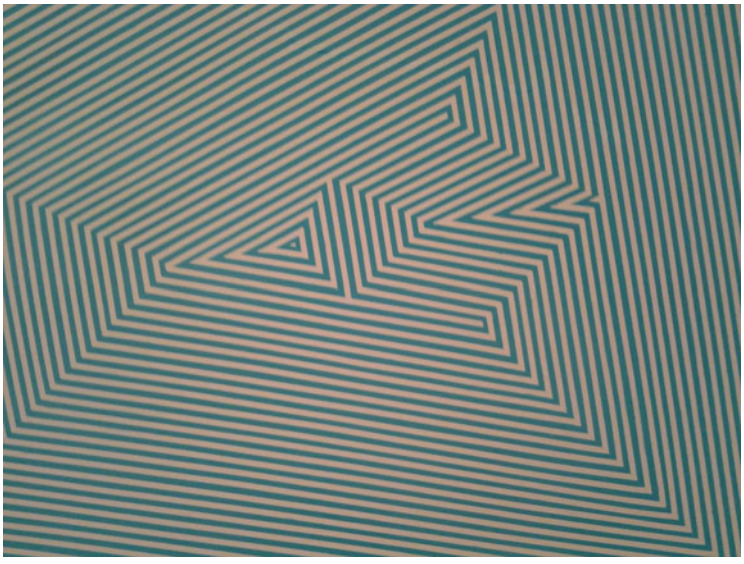


Fig. 2.44 In order to provide a strong depth cue, the spatial frequencies in the image are distributed close to a pink noise power law. Andrew Brischler, *Sacrilege*, (2013, acrylic, flash, marker, and colored pencil on canvas). Courtesy of the artist and Gavlak, Los Angeles

2.6.2 Biocybernetics

Why and how our visual perception of 2D images is influenced by frames is still an open question. Several biocybernetics models have of course been developed, along with ideas that could be used to answer to these questions [70]. A frame around a coherent image represents a discontinuity in the smooth distribution of patterns and scales in the image. Therefore, it may be useful to understand how the brain perceives pattern discontinuities. We will discuss below the two mathematical ideas which may help clarify the question.

One way to understand how our visual brain perceives framed 2D images is in terms of multi-resolution analysis. In this context there are four essential degrees of freedom in an image: two for position and size, one for orientation, and one for spatial frequency. In contrast to Fourier analysis, multi-resolution analysis uses wavelet decomposition of the signal. Wavelets are in general bases of compactly supported functions, or rapidly decreasing functions, with self-similarity properties. One very efficient wavelet procedure is the Gabor filter, and interestingly enough, it seems that the visual brain analyzes images in a similar way. Filters and wavelets are now widely modelled with Gabor functions in computing, image coding, pattern recognition, and texture analysis applications because of their mathematical convenience and their important theoretical properties concerning localization in space and in frequency [71]. A plot of the excitatory or inhibitory effect of a small flashing light or dark spot on the firing rate of a simple cell, as a function of the (x, y) location of the stimulus, was fitted with 2D Gabor functions. The residual error between the measured response profile of each cell and the theoretical Gabor filters was indistinguishable from random error [44].

The hypercomplex cells in the visual cortex are able to process inter-scale interactions. At the end of a uniform area, the cells with higher spatial frequency receptivity will have a stronger response than those with lower spatial frequency, and they will be able to excite the hypercomplex cells. At other points across the uniform signal, both high and low spatial frequency cells are equally excited, and in the process, the response of the hypercomplex cell is inhibited. In addition, the simple cells in the visual cortex are selective to four coordinates: the (x, y) retinal location in the visual field, and the two polar spectral variables for pattern orientation and spatial frequency. It is well known [44] that the first three of these variables are sampled in a systematic way by striate simple cells and there is evidence for a systematic sampling of the fourth variable as well. There is a division of labor among simple cells for the resolution of information along the different axes of the information hyperspace. Some cells, for example, favor orientation selectivity at the expense of spatial resolution in one direction, and so on.

The second way of understanding the perception of framed images is to take into account the limitations imposed by the universal (Fourier) principle of uncertainty. Simply stated, a narrow structure needs wider representation in harmonics, as compared to a more extended structure with the same pattern. This principle works for any pattern, scale fitting, interpolation, regression, Fourier decomposition, or

wavelet analysis, etc. This is why static and noise penetrates through any filter, and this is why, in a dynamical process, steep changes in a pattern generate new patterns at different scales. In musical language, in order to reproduce a pure note played for an infinitely long time, we need only one frequency. The same thing happens when reproducing mathematically a uniformly distributed pure color. However, if the same signal (short note or colored spot) is bounded by a finite region, we need more components, harmonics, and frequencies to reproduce it (see the examples in Fig. 2.46). We rely in the following development on the minimal mathematical background introduced in Sect. 2.4.

The visual brain is believed to process the light input through interrelated sequences of sampling, differentiation, and integration. Electromagnetic radiation is projected on the retina and is filtered and distributed by different types of specialized sensorial neurons (photoreceptors). The last process at this retinal level is the transmission of the signal to the optical nerve by the retinal ganglion cells. From about 100 million photoreceptors in the retina, the signals are compressed in the retinal ganglions and in the optic nerve by a factor varying from $1/2$ in the fovea, up to $1/100$ in the rest of the retina (about 100 photoreceptors for one retinal ganglion cell). Inside the ganglions, the image is processed by a type of differentiation process. The retinal ganglions analyze the signals from multiple sensorial neurons and detect phase shifts in time and space, that is, they analyze the modulation and contrast (identify the edges). What is seen by one ganglion inside its receptive field [72] triggers the ganglion to fire in a particular way: excitatory or inhibitory. The receptive field of the retinal ganglion consists in two concentric disks (antagonistic center-surround system). If the larger disk is uniformly illuminated the ganglion does not fire or it has a weak response. If the center area is irradiated with a different amount of intensity than the peripheral area, the ganglion fires a strong signal. Some ganglions fire a strong signal if the center of their receptive field has greater illumination intensity (on-center retinal ganglions), and some other ganglions fire a strong signal in the opposite case (off-center retinal ganglions). The receptive fields thus favor the analysis of boundaries, contours, and movements of the image. In addition to this dotted structure, the responses have different wavelength sensitivities for different types of cone cells. There are three important types: S, M, and L types, spectrally sensitive to short (blue), medium (yellow–green), and long (red) wavelengths.

Each ganglion cell forms a fiber in the optical nerve through which the signal is transmitted to the LNG (lateral geniculate nucleus) made of a sort of relay cells, which is a part of the thalamus. In the LNG the information is processed, then further distributed to the primary visual cortex. At this level the *simple cells*, discovered in the 1950s by Hubel and Wiesel, respond to primitive visual elements like edges, gratings, and bars, and decode information responsible for orientation of patterns. The simple cells re-transmit information to the rest of the cerebral cortex, and to the *extrastriate cortical cells*, which integrate the visual information from large receptive fields and fire when they recognize complex patterns like meshes, lattices, limbs, or buildings. Our perception of faces works in terms of first decomposing a face into smiley faces, surprised faces, emoticons, etc. [44]. This is possibly why

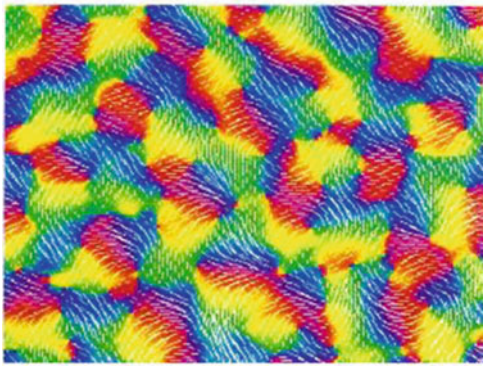


Fig. 2.45 Patches of monkey visual cortex. Colors indicate the preferred orientation of the neurons. Photo by courtesy of M. Schneider, R. Little, and M. Schneider, Pittsburgh Supercomputing Center, 1995

kids love cartoons. It is on this level that our visual brain hypothetically works in a similar way to a Gabor filter. Different such cells cover receptive fields of different sizes and different orientations, similar to the scale tuning and the sine function phase and direction modulation in the Gabor filter (see Fig. 2.45). This way, the extrastriate cortical cells perform a polar-separable decomposition in the frequency domain, thus allowing independent representation of scale and orientation.

The retina generates an 8-dimensional time-dependent scalar field with its four layers of photoreceptors (S, M, and L cone cells and rods) distributed across a 2D and almost hemispheric surface. The signals from the photoreceptors are mixed in the retinal ganglions mapping the scalar signal from each retinal point into a quaternion space of components (off-center, on-center, excitatory, inhibitory). Actually, the retinal vision system delivers a decomposition of the projected image on a 4-dimensional (L, M, S, rods) self-similarity base of circular Gaussian-like scaling functions Φ (on-center receptivity), $-\Phi$ (off-center receptivity) of minimal width given by the minimal diameter of the retinal field. When these signals are further sent to the simple, complex, and hypercomplex cells, the time dependence, space modulation, and elongation asymmetries are taken into account.

Boundary Effect

In this section we describe a possibly new physiological effect triggered by the boundary of an image. The phenomenon may originate from rapidly browsing between the framed painting and the surrounding blank ambient. While performing this motion, the visual brain records a steep change in the complexity and structure of patterns, from an excitative stimulus to an inhibitory one.

In the literature, there are several examples of psychophysical phenomena generated by such rapid variations of visual stimuli. For instance, we have the after-image, flash-drag, flash-lag, and flash-grab effects [73], as well as the very interesting and still intriguing *pattern induced flicker colors* (PIFCs, or simply the subjective color effect) [74–76]. PIFCs occur if a rapid change in an achromatic

stimulus produces the sensation of colors. The neurological interpretation is a side-effect of a neural mechanism providing color constancy under normal stimulus conditions [76]. A simple way to demonstrate this phenomenon is Fechner's pattern, or Benham's disk [74, 77]. The latter is a pattern of concentric arcs of black circles of different radii and different angles drawn on a white half-disk, rotated at 5–10 Hz in alternation with a dark half-disk. A good computer realization can be viewed under the Project LITE heading at the Boston University site (2005). This effect and the others mentioned above are basically initiated by rapid and localized changes in image patterns at a certain location on the retina (see an illustration in Fig. 2.48). Of course, through the principle of relativity, the effects must also occur if the visually localized pattern is constant in time, but the gaze is moved over the pattern fast enough [78]. In this interpretation, we expect that, by looking alternately inside and outside of the framed image, it must trigger a similar sensation which adds, in a nonlinear way, to the overall sensation produced by the image itself. In order to investigate this possible effect, we shall look more closely at some current explanations of the physiology behind the PIFC effect.

Several interpretations have been put forward to explain the PIFC effect. They all agree that the effect of color sensation is triggered rather by the temporal modulation of patterns, and not necessarily by the motion of the patterns in the visual field. One interesting explanation based on neuronal anatomic structure [76] was developed by Schramme, Tritsch, and von Campenhausen. Their model assumes a highly interconnected neural structure attached to the L, M, and S cone bases. This structure consists of 'horizontal cells' connecting laterally different types of cones in a plane parallel to and underneath the retinal plane, and 'bipolar cells' (on-center and off-center) vectoring the neural channels perpendicular to the retinal plane, and connecting the cones and the horizontal cells with the retinal ganglions. The horizontal cells mediate the phase-sensitive lateral interaction between excitations and are responsible for the PIFC phenomenon. In other words, excitation/inhibition of different cones (space resolution) and at different moments of time (phase resolution) can be measured by the horizontal cells and communicated to the retinal ganglions, which interpret the phase-shifted signals as colors.

Two very important consequences emerge from this model. On the one hand, von Campenhausen and his collaborators found [76] that the horizontal and bipolar ganglion triad placed at the base of the cones (and rods) is the site of a phase sensitive lateral interaction, and this triad acts pretty much like a system of low-pass filters retaining fundamental frequencies and lowest harmonics of the modulated patterns. This structure is present in humans (primates) and honey bees, and recent studies show the possibility of such structures in homing pigeons. It becomes obvious that this model favors the idea that retinas work like a multiple Fourier transform in the space of visual patterns. Hence, a steep change in visual stimulus, like gazing quickly over the boundary or frame of a coherent image, will generate a wide spectrum of signals in the retinal structure, and this spectrum will be recorded by the retinal ganglions and relayed forward to the LGN and the visual cortex. According to this model, we should perceive a special input when jumping over the visual boundaries. In a simple numerical experiment, one can study the

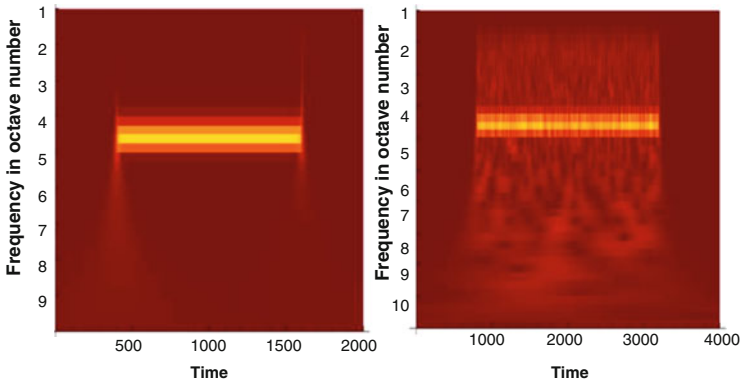


Fig. 2.46 The density plot of spectral distribution versus time (*horizontal axis*) and frequency (*vertical axis*) of a finite duration signal. The *left-hand frame* contains a pure sine signal with its frequency narrowly limited by a horizontal stripe. However, at the beginning and end of the signal, the spectrum explodes into a broader spectrum shown by the diffusion of the yellow color. This is the frame effect applied to a 1D image (*yellow rectangle*) surrounded by *dark brown*: at its ends, the spectrum is deformed and extended. In the *right-hand frame*, the sine signal is perturbed by white Gaussian noise, and consequently its spectrum is extended up and down to multiple frequencies. In this case, the frame effect is weaker

wavelet transform of a finite duration signal and measure the spectral effects of the boundaries (see Fig. 2.46). A more complete simulation of this frame effect is presented in Fig. 2.47, where the extended spectrum at the boundaries is artificially enhanced around a photography by including the wavelet decomposition of the image: yellow color bleeding out of the neat image is what a Gabor wavelet model of the visual brain would notice in a quick glance at this landscape.

On the other hand, there is always the question as to whether such a PIFC mechanism is the result of evolution, and if not, why it is present. In nature, it is rather an exception than the rule to have time-modulated visual stimuli with such a large space gradient in the phase shift. It indicates that, without selection pressure, there would be no biological purpose for the PIFCs. However, the effect may be related in an indirect way to a true evolutionarily generated ability, namely color constancy. The visual brain and retina have developed a specific ability to operate in the range of natural variations of the daylight spectrum. If we map these variations in a 3D vector space of base S, M, and L cone sensitivity, we obtain an almost flat surface, or a plane generated by the S and $M + L$ directions, the so called $S/(M + L)$ *opponent channel*. Surprisingly, the spectrum of the PICF phenomenon lies entirely in this plane [74, 76]. So according to the Schramme, Tritsch, and von Campenhausen model, the PIFC phenomenon is an indirect consequence of our evolutionary adaptability to the variation of natural daylight.

A more mathematical approach to the explanation of the PIFC effect is provided by the Grunfeld–Spitzer model [75], which accounts for the time variation as well as the space variation of the rotating patterns. According to this second model,

Fig. 2.47 Frame effect artificially enhanced around a photography by including the wavelet decomposition of the image. The *yellow color* outside of the (*blue*) frame represents a Gabor wavelet analysis made at the boundaries of the picture

the ultimate cause of the effect is the nonlinear response of the neurons (retinal ganglions). The stimuli responsible for the different colors elicited are determined by the arc lengths and the location of the black circular patterns. Grunfeld and Spitzer explained the phenomenon using a nonlinear physiological mechanism called the *rebound response*.

Under some conditions, if a neuron is maintained deactivated for a long time, for example by keeping it in a hyperpolarization state, when the inhibition is turned off suddenly, a depolarization wave of Ca^{2+} ions is triggered and travels in the form of a train of spikes. The rebound responses associated with the three color cells (cells having different response parameters) yield different responses at the color pathways and thus provide a quantitative explanation for subjective colors, such as those induced by the Benham disk. Thus, the rebound response enables cells to detect spatial and temporal edges, without a complete loss of information about the duration of the stimulation.

Along the same lines, we may mention a recent study on the nonlinear nature of the propagation of the nerve pulse as an acoustic soliton induced by a phase transition in the axon membrane [79]. The authors provide experimental evidence for the occurrence of doublet or triplet trains of spikes along the nerve. These trains of spikes are believed to be solitary wave solutions traveling along the axons through a thermomechanical mechanism that is parallel to and independent of the well known ion channel electrochemical propagation. It is well accepted, [80], that there are four types of spatial cells in the hippocampal formation: (1) place cells

responsible for sensing regions, (2) head direction cells responsible for directional moving, (3) grid cells responsible for regular patterns and (4) *boundary cells* responsible for sensing boundaries [81]. The boundary cells are really special: they are formed in the medial entorhinal cortex in the earliest days of life, in rats for example, and are related to the first tendencies of the animal to move freely around. Most recent models, [82], introduce a new input layer in the hypothalamus made of *boundary vector cells* that fire when the animal senses remote environment boundaries. These studies infer that in the earliest stable of life the very cues to location are supported only by the *boundary cells*, and only later the place and grid representations develop and interact.

Local Nonretinotopic Geometry

When we watch one object subjected to change, motion, and/or small deformations, our visual systems not only let us see and record various images of an object, but make us believe in its existence, stability, invariance, constancy, etc., once the object is identified and recognized. The visual brain is capable of building more abstract representations that allow the integration of different instances of an object into a single category and at the same time, segregate instances of different objects into different categories.

The visual brain may associate the trajectory of the object's motion with its classification, meaning, and definition. Instead of analyzing in a one-to-one mapping of the object's *retinotopic* image, higher levels of our visual brain (from V3 up) tend to deform and adapt to the otherwise local, uniform, and isotropic retinotopic space that closely surrounds the object. In analogy with general relativity, when a mass locally deforms its surrounding spacetime, in our visual brain, the occurrence of the catalogued object image induces local deformation of its surrounding retinotopic space using the repeated occurrence of the same object, and this simultaneously in all of its elements. This may explain why we cannot easily find a lost object, even if it is manifestly visible, when it is placed in a location not related to its meaning. This effect can generally be described as a sort of local spacetime causality developed in our brain when we see simultaneous events.

In a long series of research papers including [73], Cavanagh and Anstis show that, when an object moves back and forth, not only does its trajectory appear significantly shorter than it actually is, but if a light signal is emitted at the endpoint of the trajectory, this flash is mentally 'grabbed' by the object, and it is seen at the perceived endpoint of the trajectory, rather than the physical endpoint. This discovery is important because it proves that our visual brain overlaps the flash with the object at the moment of motion reversal, even if there is no causal, structural, or congruence relation between them, apart from the spacetime simultaneity of their occurrences. The occurrence of such a local nonretinoptic geometry is identified by Intriligator and Cavanagh, too [46]. They found that the visual selection mechanism acts by pointing to the spatial coordinates (or cortical coordinates) of items of interest, rather than by holding a representation of the items themselves.

There are many observations converging towards such a conclusion. A moving object viewed through a narrow slit is still perceived as a spatially extended shape moving behind the slit rather than an incoherent pattern confined to the region of the slit, a phenomenon called *anorthoscopic perception* [83]. If we watch an ellipse moving behind a narrow slit it appears compressed because the trailing edge of the stimulus is perceived to move faster than its leading edge, similarly to the flash-grab effect described in the paragraphs above. The Ternus–Pikler test (illusion of group motion versus individual motion of patterns) also demonstrates that motion establishes a reference frame according to which nonretinotopic computations take place.

It is already known from *metacontrast masking* experiments [83] that the presence of a retinotopic image is not a sufficient condition for the perception of form. This effect refers to the reduced visibility of a target stimulus due to the presence of a second stimulus, namely, a spatially non-overlapping mask. Although the target is fully visible when presented without the mask, the spatially and retinotopically non-overlapping mask can render it completely invisible.

There are more arguments in favor of the local geometry of the images in the higher level of perception than the simplistic retinotopic image mapping. For example, under normal viewing conditions, visible persistence is approximately 120 ms, so one would expect moving objects to appear highly smeared, much as in Fig. 2.48, but that is not the case. In the presence of occlusions, we do not perceive a set of fragmented parts; rather, the occluded object appears as a whole, a phenomenon known as *amodal completion*. All these effects show that at some level of perception the visual brain recognizes that points belonging to the same

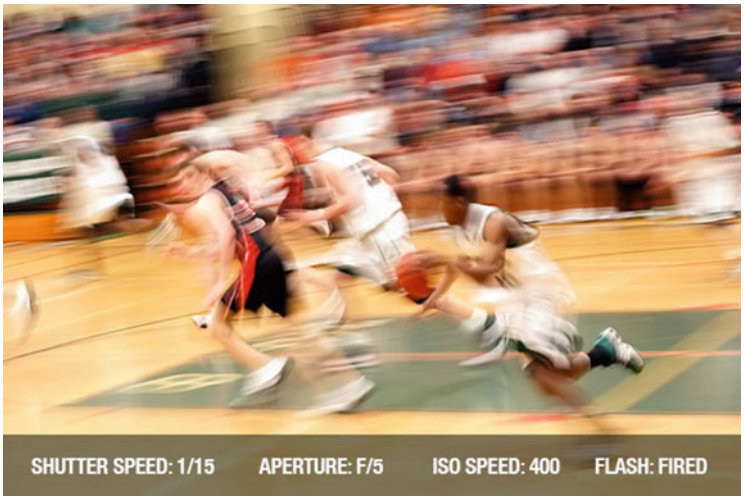


Fig. 2.48 Photographic illustration of the effects of rapid motion of patterns on our image perception. Image taken at a shutter speed of 1/15–1/60 s, that is, an exposure time of 17–66 ms. Photograph courtesy of ExposureGuide.com (2013)

entity (i.e., the image of a recognized object) move with the same velocity, and then the visual brain places the representation of the object in a relatively moving frame.

These effects can be given an explanation based on nonretinotopic representations, viz., a local geometry with a velocity-dependent metric. Öğmen and Herzog consider the existence of local manifolds created by motion segmentation as having a metric based on relative motion vectors. Such a model involves replacing the spatial Riemannian manifold model of images in the brain with a non-Riemannian spacetime manifold. A possible candidate would be a Finsler type of geometry where the metric depends on the arc length and on the tangent [84]. This geometry non-trivially generalizes the Riemannian manifolds in the sense that they are not necessarily infinitesimally Euclidean, and loosely speaking the metric depends on the curve's arc length and velocity.

Color, Form, and Frame

A familiar and very important property of visual perception is its sensitivity to context. For example, the strong influence of surrounding regions on the size or color of a target region has been known for centuries [85], as illustrated by the Ebinghaus–Titchener illusion. In their review, Shapley and Hawken present an experiment where several squares of identical wavelength spectra (red) are placed at the center of different surrounds (green, gray, reddish, black) (see Fig. 2.49). The

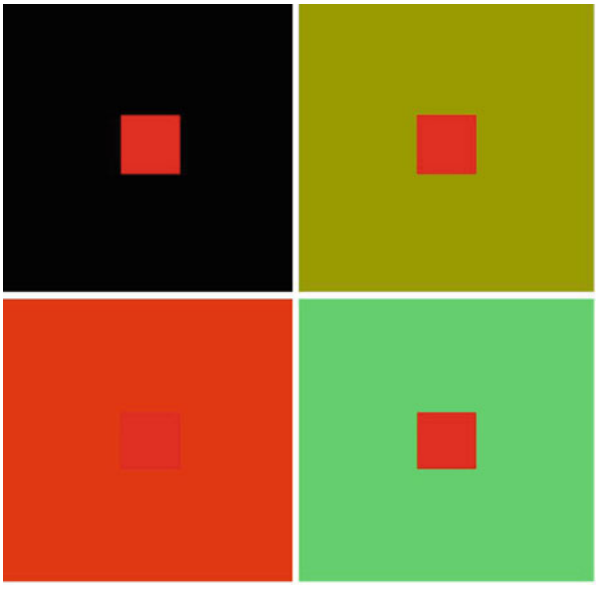


Fig. 2.49 The central square has the same color spectrum in each of the four large squares, yet it appears to have a drastically different color depending on its environment

Fig. 2.50 Artwork illustrating how the form, color, and depth of the objects link up in our visual cortex and generate special visual perception. James Little, *Juju Boogie Woogie*, (2013, mixed media, oil and wax on canvas). Reproduced with permission of the Artist

edge integration concept [86] also indicates that the perceived lightness or color of a surface cannot be determined simply by information derived from the edges of the surface itself.

Depending on the surrounding color, the different central squares appear either intensely red, pink, and even almost white [87]. Therefore, the appearance of these physically identical targets is very strongly affected by the color and brightness of the surrounding area. Psychophysical and perceptual studies have established that form, color, depth, and motion are inextricably linked as properties of objects in visual perception and in the visual cortex. This fact is known and utilized by artists, as can be seen from the example in Fig. 2.50.

Many different neuroscience laboratories have found that responses of the visual cortex strongly link multiple visual properties like depth, motion, lightness contrast, color, shape, frame, texture, and orientation. This is a consequence of a fact well known among neurophysiologists that the visual performance of the visual cortical neurons is mainly dependent on their receptive fields. The visual part of the brain is the cerebral cortex responsible for processing visual information, and it is located in the occipital lobe. It refers to the primary visual cortex (or striate cortex) V1, and extrastriate visual cortical areas: V2 (less involved in color processing but acting as a buffer between V1 and the other V's), V3 (coherent motions and complex images), V4 (integrates color and form information for perception), and V5 (motion perception). The V1 region of the brain plays an important role in color perception through the combined activity of two kinds of color-sensitive neurons, single-opponent and double-opponent cells. The single-opponent cells, which work in a

similar way to the LGN cells, respond to large areas of color. The spatio-chromatic sensitivity function of the receptive field can be expressed as the red–green $|L-M|$, and the blue–yellow $S-(L+M)$ opponent channels, respectively:

$$\mathrm{RSO}_{\mathrm{red-green}}(x,y,\lambda) = L(\lambda)r_L(\sqrt{x^2+y^2}) - M(\lambda)r_M(\sqrt{x^2+y^2}) ,$$

$$\begin{aligned}\mathrm{RSO}_{\mathrm{blue-yellow}}(x,y,\lambda) = {} & a_L L(\lambda)r_L(\sqrt{x^2+y^2}) + a_M M(\lambda)r_M(\sqrt{x^2+y^2}) \\ & -(a_L+a_M)S(\lambda)r_S(\sqrt{x^2+y^2}) . \end{aligned} \tag{2.3}$$

Here $L(\lambda)$, $M(\lambda)$, and $S(\lambda)$ are the spectral response functions of the L, M, and S cones, and $r_L(\sqrt{x^2+y^2})$, etc., are the spatial sensitivity distributions for each cone input. The single-opponent cells are circularly symmetric, hence orientation-blind.

The double-opponent neurons are affected by oppositely signed inputs from different cones (cone opponency), and also oppositely signed inputs from cone-opponent inputs at different locations in the cell's receptive field (spatial opponency). The spatio-chromatic sensitivity function of the receptive field reads

$$\mathrm{RDO}(x,y,\lambda) = a_L L(\lambda)r_L(x,y) + a_M M(\lambda)r_M(x,y) + a_S S(\lambda)r_S(x,y). \tag{2.4}$$

The double-opponent is strongly responsive to high spatial frequency color patterns, but poorly or non-responsive to color stimuli of low spatial frequency. Moreover, the double-opponent cells' receptive fields $r_L(x,y)$, etc., are non-symmetric, in fact, rather elliptic elongated in shape, so they can detect orientation (see Fig. 2.45).

The sensitivity functions (2.3) and (2.4) of the simple and the double-opponent cells are not simply products of functions depending on the wavelength and the spatial coordinates, but are linear combinations of mixed terms, each depending on both parameters. This feature amply proves the phenomenon of correlation between spatial orientation of patterns and color in our visual perception.

2.6.3 Representations of Boundaries in the Left and Right Cerebral Hemispheres

At first glance, the two cerebral hemispheres look like the mirror image of one another, both anatomically and structurally [88]. However, in spite of this apparently symmetric brain organization, the supposed fifty–fifty cross control of the left and right sides of the body, morphological asymmetries exist along with suggested functional ones. Anatomically, the brain is not symmetric: the Sylvian fissure, temporal planum, prefrontal, and venous system asymmetries evidently confirm the nonsymmetrical shape. Basically, the right frontal-parietal lobes are larger, and the left parietal-occipital lobes are larger, showing a center of mass median directed at

about 40°NE in the horizontal mid-section plane. As expected, these morphological asymmetries involve functional brain asymmetries.

It is interesting to try to discover how the brain supports recognition, space orientation, boundary sense, and artistic thinking, for example, based on the topology of the brain region where these functions are generated. In principle, it is known that the left hemisphere is more responsible for language functions, while the right hemisphere cares more for nonverbal functions. Studies show that the left brain is the logical brain, responsible for notions (concepts), symbols, words, representation of objective space geometry, distant space, sections through space, and linearity [89].

Recently, fascinating results on the dynamics of drawing abilities in humans undergoing unilateral electroconvulsive shock therapy (ECT) have been reported in [89–91]. In clinical experiments, patients were asked to draw a house, a person, a tree, a cube, a bridge over the river, or rails stretching to the horizon, before the shock, and several times subsequently, at successive intervals of 10–20 min, until the impacts of ECT were completely removed. Since the brain hemisphere subjected to the shock is temporarily inhibited and then gradually gets back to normal, the evolution of the sketches reveals the changing balance of interaction between the inhibited and active hemispheres. In particular, it was demonstrated that patients recovering from right-ECT with active left hemisphere *draw iconic representations of objects, by merely sketching their boundaries*, and connecting them in a hieroglyphic way. Later, over time, the patients gradually return to a realistic transfer of the visual images.

Studies of drawings by patients with focal lesions in the right or left hemisphere have helped us to understand how artistic thinking is supported by brain structures [89]. The role of the right hemisphere is significant at the early stage of the creative process, and it operates with images placed in a visual space as icons. Only by ulterior activation transfer from the right hemisphere to the left can the expression and artistic act be completed by a conscious effort.

The role of the right hemisphere is significant in the early stage of the creative process, in the stage of conception. Drawings made by patients with left hemisphere damage show somewhat 'impossible' postures, like those in the paintings of Picasso, Matisse, and Leget. Such drawings were also noticed in the pictures of 5- and 6-year old children, or at the dawn of human civilization, in the Stone Age, or in the representation of mythological characters and creatures, such as griffins, winged horses, and sphinxes.

For a healthy adult living in modern times, the images created in the 'inner' space by the right hemisphere need to be transferred into an essentially different language, i.e., the left hemispheric language of words, which operates with a finite number of discrete units. Somehow, the images have to be moved from the subjective geometric spaces into algebraic, finite, and discrete (words) representations. The transfer, once initiated, develops oscillating feedback between the two hemispheres. Obviously, the act of creation requires a continuous dialogue between the hemispheres.

Studies made with right hemispheric suppression show that drawing representations of the objective space geometry created by the left hemisphere were schematic

and body-less. These results suggest the hypothesis that the left hemisphere is modeling objective space in an analytic way by sectioning it into discrete parts according to the principle of opponent definitions [90].

The real geometry of material objects is a typical function of the right hemisphere. During right hemisphere suppression, the mechanisms of stereo viewing are altered and 3D objects are represented as flat. Right hemisphere suppression results in the loss of the direct visualized character of visual perception. Representations of objective space geometry created by the active left hemisphere are schematic, showing abstract simplified forms, such as boundaries like a rectangle to designate different objects. The image of a concrete object is replaced by its contours [91]. These experiments, corroborated by the results showing that the left hemisphere is orientated toward remote regions where the space is more body-less, suggest that the left hemisphere is responsible for our enhanced understanding of boundaries. That may be because of the logical and organized way the left hemisphere tries to program activities in terms of frames, schedulers, tables, hence contours, limitations, deadlines, frontiers, and boundaries.

2.7 René Magritte and Bernhard Riemann

The boundaries between the natural and art worlds are central foci in René Magritte's work [92], and so too is the frame. His obsession with the balance between reality and fiction is often found in representations, enframings, and boundary game studies. Probably the most outstanding example is his *La condition humaine* (see Fig. 2.51 left). The easel's position in front of the window opening up to a landscape, with the easel mimicking the same landscape, is a statement of disruption of reality by art. The edges almost vanish, the outside becomes inside and vice-versa. The real world and the artistic creation become one on the easel: just two different representations—and the accuracy of the painting tells us these are very faithful representations—of another truth, beyond vision and perception. The presented theme is reminiscent of the etymology of frame and boundary from the word 'echo' (*bombitire*, *budina*) in Vulgar Latin (see Chap. 1). The easel reminds us of the primordial definition of the frame as skeleton and introduction. Whatever the viewer's interpretation, it is a display of a Rilke-type of philosophy on questions that are more important than their answers.

All of these elements have a key, yet there is a detail which stands out: the vertical white dotted line to the right of the canvas. It obviously has a realistic interpretation, but why is it placed so dissonantly compared to the calm of the landscape? Almer thinks it is a motif, similar to Heidegger's *riss* or *Umriss*, standing for a contour. However, it cuts the holistic illusion of the landscape. The line "interrogates the question of what has been framed: is it the painting within the painting, is it the outline of the canvas, or is it the boundary, the frame, the limit of the landscape and its representability? This line creates a joint, a hinge on a door, which simultaneously opens and closes."

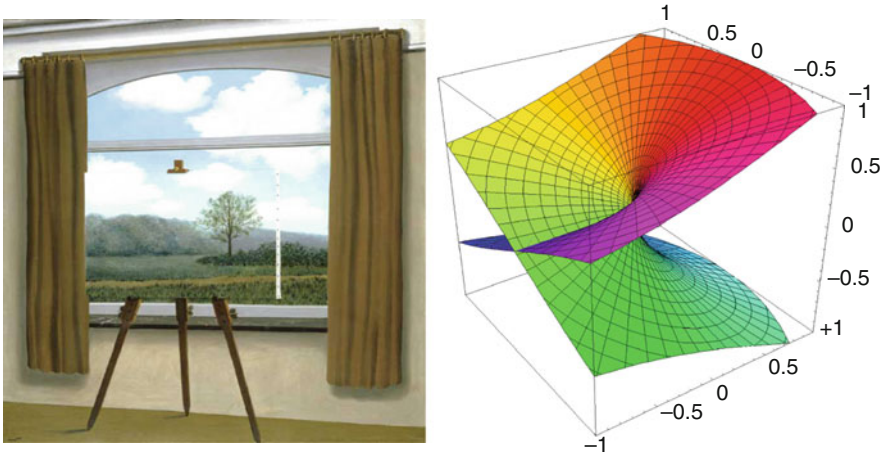


Fig. 2.51 The duality between reality and its representations presented artistically (*left*) and mathematically (*right*). *Left*: Framed reality and its framed representation. René Magritte, *La condition humaine* (1933, oil on canvas). Taken from the public domain internet pages of the National Gallery of Art, Washington, D.C. *Right*: A 3D plot of a Riemann surface $Z = \sqrt{z}$. The attempt to fit the graph of a complex function with $(2+2)$ dimensions in a 3D space representation generates undesired non-uniqueness which must be removed by branch cuts between different sheets. Note the disappearing white *vertical dotted line* in Magritte's landscape, and how similar it is to the line of intersection of the *green* and *purple* surfaces: the location in our mind (*left*) or space (*right*) is the same, but it hides a repetition of multiple realities

Inspired by Almer's words regarding the dotted line, we can see a deep connection between Magritte's persistent questions here, and the duality inherent in the double reality of the Riemann surfaces. In the complex plane (considered as a 2D Euclidean plane in which the horizontal axis is real and the vertical axis is 'imaginary'), we can define functions that overlap one another and repeat this foliation to infinity, like a spiral stairway. One example is the complex square root $Z = \sqrt{z}$. In the right-hand frame of Fig. 2.51, we show the second leaf of this complex square root function. Using Riemann surface theory, it is easy to show how the two leaves overlap.

Regardless of whether a smooth surface is self-intersecting or not in an arbitrary parameterization, if we want to plot it locally as a function $z = f(x, y)$, it must satisfy the vertical line criterion, that is, uniqueness of the value of z for each point (x, y): any vertical line should intersect the graphs $z = f(x, y)$ at one point only. Otherwise the function f must be declared multi-valued and another geometrical parameterization must be chosen if we care about uniqueness of the realization. Such situations were a little bit tricky until the 1850s, but it was Riemann's merit to solve the problem in an elegant way by introducing the concept of a 'branch cut'. Branch points are the points where various sheets of a multi-valued function intersect. For example, the function $\sqrt{x + \mathrm{i}y}$ has two branches: one where the square root has the plus sign, and the other where it has the minus sign. A branch cut is a

Fig. 2.52 In the painting *Evening falls*, the shattered glass still contains the image that you see through the window. Magritte questions the 2D representation of a framed 3D world, and the possibility of understanding a higher dimension through combinations of flat images. René Magritte, *Evening Falls* (*Le soir qui tombe*), 162 × 130 cm, oil on canvas. Courtesy of the Menil Collection, Houston

curve in the complex plane which makes it possible to define a single analytic branch of a multi-valued function on the plane when that curve is removed.

To sum up, a Riemann surface is a multi-valued 2D smooth manifold that can be represented as a union of 2D real single-valued sheets connected together by the branch cut curves. Locally, the Riemann surface is homeomorphic to the complex plane (in any of its sheets), but globally it has a complex topology, different from that of the complex plane. It represents multiple repetitions of the same type of object. The only rule that can make the difference between the connected sheets, and that can also label them correctly, is the branch cut curve (see Fig. 2.51 right). Magritte's white dotted line in Fig. 2.51 (left) may be understood as the concept of 'branch cut' adapted to his own realm, connecting (or separating) multiple landscapes. He needs to separate two identical representations, and only one cut is necessary.

In *Le soir qui tombe* (Fig. 2.52), Magritte shows the same duplication of a painted reality, this time manifestly transparent, but broken. His question may be similar to that of Matisse and Rothko, namely, after the glass is broken, do we see something

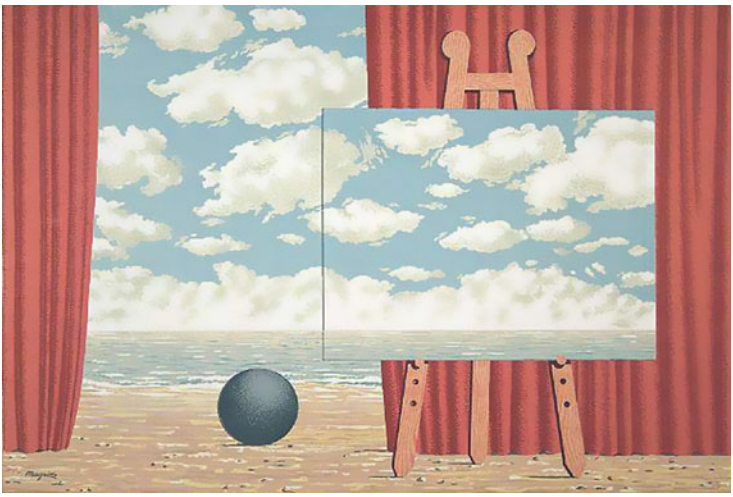


Fig. 2.53 The central theme of the frame and enframed world in René Magritte's painting *La belle captive* (1931, oil on canvas). Permission for reproduction from © 2015 C. Herscovici, Artists Rights Society (ARS), New York

new or do we just see what we have already seen? The frames multiply in the room: real frame, wall, curtain, floor/ceiling, lower and upper window, and so on. The view through the empty window is obviously flat and almost fake. Magritte approaches the frame theme in many paintings, among which we mention only a few here.

In *La condition humaine* (Fig. 2.51), *Le soir qui tombe* (Fig. 2.52), and *La belle captive* (Fig. 2.53), Magritte presents the reality/representation duality by using the motif of duplication of the landscape, and separation of the two images by the frame. In almost every such painting, the duplication is an almost perfect isometry. Especially in *Evening Falls*, this one-to-one mapping is underlined by the occurrence of the centered upper sun in the window's frame, and in the break. It is as if the force of the real, that sun, has guaranteed not only the appearance and veracity of that representation, but simultaneously ruined and frozen its apparatus. Both solar disks are perfectly circular and thick, painted with the same bright orange, like in the situation of the afterimage created by looking at an intense light source, such as the sun. The idea of duplication and repetitions of a landscape is taken further than shattering one of the images. In *Les charmes du paysage* shown in Fig. 2.54, Magritte completely empties the frame of any image, except of the floor line left there to enhance the emptiness of the frame. By replacing the eye's iris with a cloud-filled sky in *La faux miroire* shown in Fig. 2.55, Magritte questions what we see and what we think we know. Regardless of whether the sky is a reflection of what the eye is seeing, or the eye is an opening into another reality, one thing is certain: Magritte offers an invitation to look at the world only through frames.

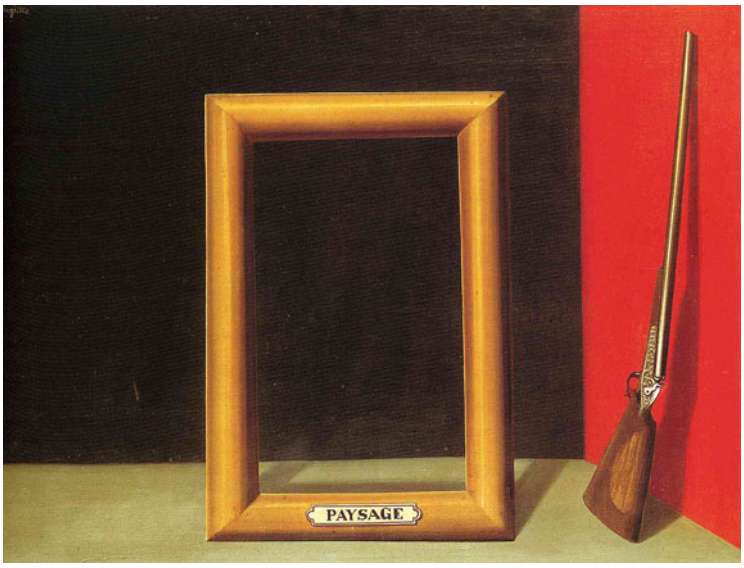


Fig. 2.54 The theme of the frame, whether filled with the image of a landscape or empty, as in this painting, is abundant in Magritte's paintings. René Magritte, *Les charmes du paysage* (1928, oil on canvas). Permission for reproduction from © 2015 C. Herscovici, Artists Rights Society (ARS), New York

Fig. 2.55 In this painting, the eye may be an opening into another reality, or a peek at an inner vision. One thing is certain: René Magritte's *The False Mirror* is an invitation to look at the world through frames of different kinds. *La faux miroire* (1928, oil on canvas). Permission for reproduction from © 2015 C. Herscovici, Brussels/Artists Rights Society (ARS), New York

Chapter 3
Boundaries in Social Systems

Over the last few years, the concept of boundary has been at the center of influential research agendas in anthropology, political science, social psychology, sociology, and economics [93]. Perhaps the most serious and pressing challenge in the social sciences is to find a working mathematical representation for social systems. Universal models for social systems must be defined in very high-dimensional spaces, with complicated structures, which have very little to do with the physical space. Nevertheless, if such systems are related to a certain physical territory, they can coexist and overlap. Studying social and economic systems in relation to the dynamics of their boundaries, and consequently considering the interactions occurring at the level of their boundaries, adds value and understanding to the research, particularly when one is concerned with the study of relational processes.

There is a lively contemporary scholarly activity in the field of social boundaries. For a review of the literature see the introduction to [94], for example. Social boundaries are far from being uniquely defined. The multi-dimensionality in collective identity formation and the diversity of social networks determine the existence of multiple, interacting boundaries. One can define social boundaries through ethno-racial characteristics, spatial and demographic distribution of institutions, job markets, immigration processes, nationalistic characteristics, or even as aesthetic boundaries, or through concepts like gender, sexuality, religion, health, or simply the dynamics of risk.

3.1 Social Science Approach to Boundaries

How do boundaries work in social relations? How do their existence and dynamics generate feedback in social relations? Social boundaries are generated and survive through a multitude of mechanisms, tangible processes, or concepts. Well known examples of such mechanisms are: cultural capital and membership, racial and

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_3

ethnic group positioning, professional jurisdictions and controversies, social and collective identity, group structure, residency and civil rights, and even cognition. In some studies, such as *Ethnic Groups and Boundaries* [95], the author embarks on a systematic classification of boundaries in three levels of abstraction: territorial, social, and categorical. Donaldson and Wood consider that *categorization* (or classification) processes are central to both the practice and understanding of control [96]. And the use of categories is connected to the existence of boundaries, because the category concept stems from the concept of boundaries that delineate one category from another.

There are two important mechanisms known to generate boundaries for a social system, and the structures of these boundaries are shaped by the balance between these mechanisms. The mechanisms are:

– conceptual distinctions (or interpretations),
– institutionalized social differences.

Each of these mechanisms can generate its own boundaries, viz., *symbolic boundaries* and *social boundaries*, respectively. These two categories have a nonempty overlap. For example, crossing a nation's frontiers, a social boundary, may sometimes be a symbolic gesture, too. The symbolic boundaries are created and agreed upon by the social actors in order to define reality. Even play or humor can be testing procedures for symbolic boundaries. They represent the boundaries between groups of people, fellowships, or similarity classes. They can be crossed by members in an effort to acquire social status or resources. A typical example of crossing symbolic boundaries is the act of prejudice [97].

In contrast, social boundaries divide social systems according to access to resources and opportunities. It is rather the exception than the rule when symbolic boundaries become social boundaries. This may occur when a high degree of coherence or collective interaction or consensus exists between the same social actors within the same symbolic domain. Sometimes symbolic boundaries enforce social boundaries, and sometimes they contest the meaning of social boundaries. Cross-cultural differences may change the way symbolic boundaries are linked to social boundaries.

In Sects. 3.1.1–3.1.4, we present the main mechanisms of social boundary formation, listed here from the micro to macro levels of analysis [93, 98].

3.1.1 Social and Collective Identity

The pressure to evaluate one's own group positively leads to the self-differentiation of social groups. As a consequence, the perception of group boundaries as impermeable will make social change more likely for low-status groups who engage in social competition. Direct exclusion, over-selection (as in very competitive educational institutions), self-exclusion, and lower level tracking are key mechanisms in the generation and stabilization of social (symbolic) boundaries.

3.1.2 Class, Ethnic, and Gender Inequality

Cultural practices have classificatory effects that shape social positions by defining class boundaries. Dominant groups generally succeed in legitimizing their own culture, exercising 'symbolic violence' by imposing a specific meaning as legitimate. In this case the symbolic distinction ends up producing boundary closure. Cultural boundaries are usually semi-permeable. While a certain cultural enclosure is valid for a social group of a determined cultural level, the same enclosure may disappear for an individual with a higher level of culture.

Mobility has always been configured by borders and boundaries composed of a multiplicity of hybrid objects, from infrastructure and technology to law and culture [99]. These boundaries are permeable to different degrees, hence creating a society that is differentiated by speed and access.

In his anthology, Fredrik Barth [95] makes a clear distinction between the *ethnic boundary* of a group, and the cultural stuff that it encloses. The ethnic boundary is a social boundary, even if it may sometimes have some territorial counterparts. Barth defines the interior of the ethnic boundary, the ethnic group, through a relation of equivalence between people. Two people belong to the same ethnic group if:

– they have the same criteria for evaluation, judgment of value, and performance,
– these criteria do not change when a member interacts with other non-members.

There is a manifest mathematical analogy for this definition. According to this analogy, an ethnic group is either a class of equivalent systems of reference sharing the same laws, or a fiber bundle where the base space is the set of members and the standard fiber is the set of criteria understood as a subset of all possible criteria of all cultures. Two members of the same ethnic group may see things differently, but they can always find a transformation with some degree of smoothness and faithfulness that maps one's judgment into the other's judgment. There are interactions between members of different ethnic groups, and the corresponding transformation functions between their different judgments and evaluations are expected to have singularities, defects, and gaps. Only members of the same ethnic group can be connected by 'smooth' transformations of criteria and judgments. It follows that the boundary of the ethnic group is actually the boundary of the set of smooth transformations preserving the quality of membership of that ethnic group.

3.1.3 Professions, Science, and Knowledge

The notion of boundary is also an essential tool to describe how models of knowledge are diffused across countries, and how they impact local institutions and identities. In this context, boundaries may become means of communication as opposed to divisions, thus becoming essential to the circulation of knowledge and information across social worlds. For example, scientists want to distinguish

themselves from amateurs and charlatans by establishing the 'boundaries of real science'.

3.1.4 Communities, National Identities, and Spatial Boundaries

These can be large-scale collectivities where members are linked primarily by common identities, but minimally by networks of directly interpersonal relationships. Examples include nations, races, classes, and genders [94]. In some studies, community borders are seen as interstitial zones, largely dominated by processes of globalization and transnationalization that have increasingly deteriorated and hybridized national identities. For example, the border between two countries induces more subdivisions than just the two nations. To the east and west of the Romanian–Hungarian border at Bors, there are Romanian communities in Hungary, Hungarian communities in Romania, and the names of some of the cities are mirrored in the two places: Santion becomes Szentianosh, and Szentpeterszeg is mirrored into Santpatru, etc. The border between these two nations generates three separate sub-borders. Immigration, migration (including members of transnational and professional elites), refugees, and displaced and stateless persons become generators in redrawing the boundaries of national identities.

3.2 Social Boundaries and Networks

We are entering a new type of society, the network society, and a new type of economy is without any doubt one of its features. Following this line, it is easy to predict that the new technological paradigm will allow the formation of new types of social organizations and social interactions through electronically based information networks. A second feature is globalization, given the technological capacity of certain systems to work as a unit in real time on a planetary scale [100]. Interactive hypertext is definitely another important feature of this new type of society. In this developing new world dominated by markets and networks, the state and the family are entering a crisis. Castells [100] believes that the key themes of traditional society (religion, nation, ethnicity, locality, nation, family) will tend to break up in favor of value-founded communes.

The new society is made up of interactive information networks, so it becomes important to determine how networks interact, and what determines their boundaries. The individual becomes secondary to its position, and is reduced to the degree of a node in several networks. It is less and less important where you live or where you graduated, and more and more important how many 'friends' you have on Facebook, or how many 'followers' you have in professional internet networks.

The spatial structure of social systems is transformed. Territorial contiguity as a condition for the existence of social practices is being transformed into totally different concepts, such as network bandwidth and internet connectivity. The intrinsic geometry of social systems changes because 'spaces of places' now transform into 'spaces of flows' [100]. The new type of city, the 'global city', for example, is a network of noncontiguous domains embedded in all the big cities of the world.

Theoretical approaches to social systems through classical network methods are amended by limitations. Although the society of networks is one of the most modern approaches in sociology, it suffers from a number of flaws. Firstly, people still live in houses, and houses involve the existence of physical space. In addition, people still travel, while the age of suburban development and freeway commuting are no longer sustainable models [101]. There are other criteria which, at least for now, remain independent of the development of information technology: basic needs of life, climate, ecology, and health. The new information technology era has irrevocably and irreversibly reconfigured our space and time. There is an urgent need to accommodate a new type of social model, in a new type of spacetime, and in particular, to allow the interaction between various networks of humans and of nonhumans, within the constraint of keeping the possibilities of human individuals constant throughout their real time existences.

In some specific social domains, there are local solutions trying to accommodate this new paradigm. For example, in social urbanism, the current response to increasing density is mixed use, allowing new residents and new jobs to maintain a harmonious balance. The mixed-use building ('pixelated urbanism', see Fig. 3.1) is part of the 'smart growth' alternative, a solution to suburban sprawl, urban

Fig. 3.1 Pixelated urbanism. A mixed-use strategy for urban density and neighborhood development, by Adrienne Watkins (2013 Thesis, University of Washington). Reproduced with permission of the artist

traffic, and inner-city decay [101]. The vast majority of such local solutions show a tendency towards poly-culturalization, diversification, niche-formation of friendly economic structures, and greater collective interaction, with the consequence that more complex types of boundaries interface such pixelated configurations and fractal networks. In modern urbanistics, the approach to such phenomena is broken down into several stages. A first stage called 'marking space' [102] involves the actant's modifications or enhancements of distinctions that are already present. The next stage in architectural actions, the 'filled space', refers to complete transformations of the physical world by human interaction with the material environment. The marked space stage occurs in a space otherwise unaffected by human presence, making use of its already existing properties, while in the filled space stage, its physical properties are continuously contained and transformed by human interaction. According to Vis' article [102], social systems *create boundaries from the inside towards their outside*, continuously processing the shape of the social world by taking place in space.

3.3 Impact of Social Boundaries in Social Relations

According to Wood and Graham [99], there are two fundamental problems in the study of societies: firstly "the way in which long-lasting social structures appear out of social interactions; secondly the method by which power can act at a distance". These questions sound familiar to a mathematical-physicist's ear. The first is similar to the problem of understanding how stable patterns can occur from local and nonlinear interactions. The second problem of social systems is closely related to the fundamental issue of grand unification of field theories.

Any theory of social systems should attempt to find plausible answers to these questions. A modern social theory should then contain equations involving more than just humans or groups of humans. Given the present context of the industrialized societies, a unifying social theory must take into account all the complex systems. For example, the recent actor–network theory (ANT) [103] adds to the interactions between humans, the interactions between humans and *inhumans* (other living beings, natural processes), and *nonhumans* (man-made systems), all of these three categories being called *actants*. The fundamental principle of ANT consists in accepting a generalized role for the actants, which is not only limited to developing and accomplishing a program of action, but also involves the enrolment of other actants in fulfilling those programs. Thus, nonhumans can be conceptualized as a social mode of ordering, not fundamentally about control of the person, but about control of information and activity. Wood and Graham therefore conclude in [99] that nonhumans (like surveillance systems for example) can be seen through a topological perspective as a technological operator for generating boundaries as a form of territoriality, and for controlling mobility through the construction of boundaries.

According to this theory, and to what we mentioned in the introduction to this chapter, complex systems whose functioning is based on software, and whose power decisions are taken on the basis of categorizations can therefore deliberately delineate between areas based on prior categorical work. Things that are allowed to pass through the boundary, and the nature of the spaces separated by the boundary, will result from the classification processes [99]. Nonhuman complex systems can therefore be conceptualized as a social mode for ordering information and activity.

Using an interesting example, Sloterdijk [104] describes how electronic road tolls allow a smoother flow of traffic by providing a bounded exclusive space for the 'kinetic elite', with slower and congested traffic outside of this boundary, where the 'cash poor–time rich' are relegated. The same process happens for the 'kinetic elite' transgressing national borders, while 'illegal' migrants and refugees find that such borders become more inaccessible.

The world is definitely moving more towards the software-sorting type of social (or at least symbolic) boundary-generation direction. There is one socio-political problem, however, when following such a direction. So far, the software is still designed by humans. Hence, the potential for discrimination, and consequently for selective mobility and permeability of boundaries, belongs to the software architects who may be able to embody their prejudices in the architecture itself, in spite of the commercially postulated infallibility of computer systems. On the other hand, there is another direction on the IT market: the trend toward self-programming software. Very soon software-based nonhumans will use learning systems that are not based on simple algorithms, but possess algorithms that enable the writing of additional new algorithms to cope with new knowledge and new situations. In this case, nonhuman boundary actants will be able to make judgements beyond the binary, moving towards heuristic boundary generation.

Inhuman systems can also affect boundaries. An extended flood following a major weather front can isolate a certain geographical community and totally change its social structures. An example is provided by hurricanes Rita and Katrina which hit the city of New Orleans between 29 August and 5 October 2005. Among several long-lasting social effects generated by these inhuman events are the permanent migration of populations to large distances, and the neat process in which new social boundaries are generated. A large part of the population whose houses were destroyed by the hurricanes in New Orleans was represented by economically challenged families, whose lives merged very well in the impoverished pre-Katrina local diversity. The hurricanes displaced similar groups of people (about 800 000) in other states, including in developed industrial areas like Houston, Baton Rouge, or Dallas. In this situation new symbolic boundaries appeared in these areas, and soon enough they became social boundaries, especially through the interaction with local (mainly illegal) immigrant populations.

Lamont and Molnár consider [98] that the concept of social boundary has become one of our most fertile thinking tools because it captures a fundamental social process, that of *relationality*. The boundaries of social systems are definitely a complex subsystem of the social system, and they form an interesting subject for social topology and socioeconomic dynamics. There is no other example, except

maybe the envelope of the biological cell, which has such transforming, vanishing, gluing, and branching dynamics as social system interfaces.

For example, the European Union (EU) is defined as an economic, social, and political system constructed as a union of states, transcending national and governmental frontiers. The EU boundary is in continuous change, depending on its development and purposes. Initially, the EU was created as a political system to defend European peace: it included only Western Europe. Later on, it became a system of coal and steel users, and its boundaries changed accordingly, including Belgium, France, Italy, Luxembourg, the Netherlands, and West Germany (see Fig. 3.2). Later on, the EU was restructured as the Eurozone, and even later it became the Schengen area. The EU went on reshaping its boundaries according to its economic or trade relations. However, other more realistic conceptual boundaries overlap these institutionalized boundaries. Initially, all EU countries offered facilities for individuals to become permanent residents after residing in the EU for a specified time interval. However, reports from the European Commission show that these rights exist more in theory than in reality. The cross-border mobility rights under the directive failed, with some countries limiting these rights to only one: no need to apply for an entry visa for the EU when setting up in a new country. Of the almost one million third-country nationals with EU rights, only around 100 per year were able to make use of the freedom of labor provision across the whole of Europe. Van Houtum and Pijpers describe this phenomenon as "hiding in a gated community

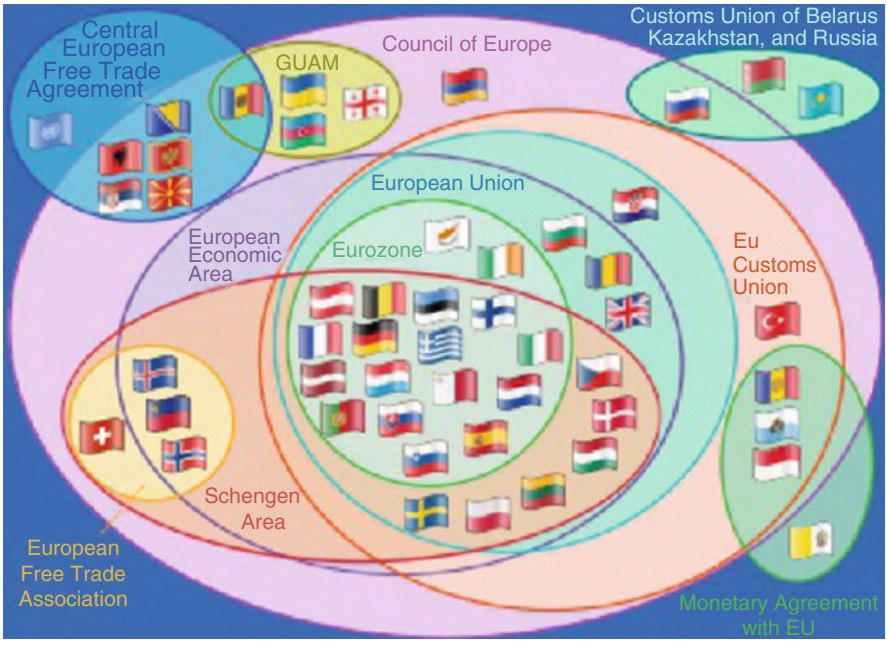


Fig. 3.2 Different substructures of the European Union with different boundaries

in order to protect this comfort zone and trying to exclude outsiders" [105]. Such situations generate changes in the definition of social boundaries, including quick metamorphoses from deterministic to chaotic, and from smooth to fractal.

The *natural identity criteria*, or equivalence relations between the actors of a social system (individuals, groups), can be generated by a long list of social partition factors: ethnic, racial, gender, profession, knowledge, cultural membership, religion, community affiliation, and national identity. The study of the nature and properties of social system boundaries reveals several interesting subjects, among which the relationship between social and symbolic boundaries, the mechanisms for the generation of boundaries, difference and hybridity balance, and group classifications. At the same time these ideas require the development of new mathematical tools studying boundaries in non-traditional formalized fields like social sciences, arts, or economics.

Last but not least, understanding social system boundaries can help to improve daily life, e.g., by designing better immunization protocols for cell phone networks or human communities against the spread of viruses [106]. In addition, knowing the nature of the interfaces between different systems and networks would facilitate early detection and prevention of joint blackouts, undesirable community and social changes, or propagation of crises between systems.

3.4 Mathematical Approaches to Social Boundaries

Mathematical models for social systems developed in mathematical sociology try to connect the data from sociology (surveys, percentages, tables, etc.) with rigorous formal analysis in terms of differential equations, invariants, and theorems. Deriving such equations from assumptions about chances directs the focus of research to the field of stochastic processes. At the present time, all the mathematical approaches to social systems are statistical/stochastic, or network and graph theory based, or approached by reduction of the dimensionality of the data set in order to present intuitive 3D images.

Recently, mathematical sociology has developed highly formalized theories like the *agent-based models*, where social life is modeled as a function of interactions among adaptive agents who influence one another in response to the influence they receive. In comparison with variable-based approaches using stochastic differential equations, agent-based simulation offers the possibility of modeling individual heterogeneity, representing agents' decision rules explicitly, and situating agents in a geographical or another type of space [107]. It allows modelers to represent in a natural way multiple scales of analysis, the emergence of structures at the macro or societal level from individual action, and various kinds of adaptation and learning, none of which is easy to do with other modeling approaches. These models allow sociologists to understand how simple and predictable local interactions between agents can generate familiar global patterns of social structure. They can also be used to perform virtual experiments that test macrosociological theories by

manipulating structural factors like network topology, social stratification, or spatial mobility.

Agent-based models consist of agents that interact within an environment. The philosophy of the model is of cellular automaton type. Agents are parts of a program that are used to represent social actors or individual people, organizations, or bodies such as nations. They are programmed to react to the computational environment in which they are located, where this environment is a model of the real environment in which the social actors operate. The most traditional models are opinion dynamics, consumer behavior, industrial networks, supply chain management, urban models, and electricity markets. The last two are *spatially explicit* computational models and generate patterns, contour levels, and boundaries on a map.

Another large class of mathematical models for social systems is the network approach. For a recent review, see for example [108]. A social network is a social structure composed of individuals (or organizations) called nodes, which are connected by one or more specific types of interdependency, such as friendship, kinship, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge, or prestige. In social network analysis, the groups are not necessarily the building blocks of society: the approach is open to studying less-bounded social systems, from non-local communities to networks of exchange. The model focuses on how the structure of ties affects and constitutes individuals and their relationships. Network analysis looks at the extent to which the structure and composition of ties affect norms, as opposed to other models which assume that socialization into norms determines the behavior. Through the network model, it is easier to implement more modern features like small world theory, fractal geometry, scale-free networks, global network analysis, or complexity theory.

Social system networks can be exemplified by the different types of infrastructures our daily lives depend on, viz., the means of transporting goods, energy, or information through communication networks, epidemics, rumors, and opinions through social networks, electrical power through the power grid, and land transportation through road and railway networks, etc. There is a strong interaction between the structure of these networks and their mobility properties (capacity, bottleneck structure, delays). One consequence of using the network model is the possibility of analyzing the propagation of the risk of failure through the coupling between systems. All the relevant infrastructures for daily life are interdependent, and failures in one network are very likely to propagate to the others, leading to large scale phenomena. An example of such a real situation is the 2003 blackout affecting Italy and Switzerland [106].

In order to detect boundaries, patterns, and non-random structures in a social system some recent studies used the method of exponential random graph modeling (or p^* modeling) [109]. Together with the multiple regression quadratic assignment procedure, the ERGM represents a well-developed statistical technique, used extensively in the social sciences, that enables examination of the underlying mechanisms of network factors and processes that generate non-random network structures. A recent version of these methods is the *deconstructing networks* (or motif analysis) method. This consists in decomposing networks into subcomponents and comparing

the relative frequencies of occurrence of these subcomponents across networks. A network can be deconstructed into sets of dyads, triads, or n-node subgraphs. The frequency of occurrence of each such type of subgraph is compared with similar frequencies from other empirical networks, or models of random networks. For example, in [109] these methods were used to study animal social structures and sociality.

Another traditional procedure for studying social systems by network theory is to use statistical models. The aim of the statistical model is to represent the main features of the data set (the network) by a small number of parameter estimates and to express the uncertainty of those estimates by standard errors, distributions, etc., which give an indication of how different these estimates might be if the researcher were to repeat them [108].

In [110], the authors use a general Markov chain model of network evolution that operationalizes the aversion aspect of the Schelling segregation model: actors within a social network "may not strictly prefer forming homogeneous networks", but such structures can emerge if actors are "subject to a small bias against interaction with partners who are dissimilar from themselves." This model predicts pretty well the segregation process and non-homogeneous clusterization mediated by the rewiring of an initially random network. Even if not discussed explicitly, the boundaries of the network change, following a clear pattern of clusterization.

Among various statistical models some particular models, of interest for our book, are the *distance models*. Such metric models, with roots in network and graph topology theories, have to face two challenges: on the one hand one should find a sociological definition for distance, and on the other, the mathematical formalization of this definition should find harbor in some rigorously defined metric space.

3.5 Social Distance: Euclidean Metric

The *social distance* describes the degree of separation, according to some stated criteria, between different groups of society, and in general has very little in common with the Euclidean distance. The social distance should include all social differences such as social class, race/ethnicity, or culture, but also the fact that the different groups have impermeable boundaries, i.e., they do not mix.

For example, Bogardus, who actively promotes the use of metric spaces in sociology, introduces the concept of social distance scale [111], defining social distance as a function of the *affective distance* between the members of two groups. For him, social distance is essentially a measure of how much or how little sympathy the members of a group feel for another group. His social distance involves the concept of affectivity in its definition.

Another metric approach views social distance as a normative category, referring to the norms in terms of 'who is an insider' and 'who is an outsider'. This *normative distance* is the social topological concept closest to a rigorous definition of a social boundary. Normative social distance, sometimes called *psychological distance*,

generally differs from affective social distance. For example, the psychological distance is large between individuals influenced by different cultures, especially when there is no meeting point between the two. The consequence is, in general, the generation of prejudices that various cultural groups assume to be true for other social groups.

Another definition of social distance is related to the frequency and intensity of interactions between two groups: the more the members interact, the closer they are socially. This *communication distance* becomes useful in sociological network theory, where the frequency of interaction can determine the 'strength' of the edge between two nodes.

Route planning exercises have also hinted at a conceptual link between social distance and physical distance. When asked to draw a route on a map, people tend to draw routes closer to friends they pass along the way and further away from strangers. This distance effect is robust, even after controlling for how easy it is for the people passing one another to communicate. Here, social relationships influence the way participants reason about physical distance, and it supports the notion that social distance, defined here as friendship, and physical distance are, again, conceptually linked. There is some evidence that reasoning about social distance and physical distance draw on shared processing resources in the human parietal cortex.

With the growing popularity of social networking sites, social network visualization has become a tool to improve user experience or to condense information into a friendly and intuitive format. The main procedure used in such mapping of data is inspired by network theory, namely the social distance and the frequency of communications between the social actors. A successful application of this type of visualization is *sociomapping* [112]. This procedure uses the landscape metaphor to display complex multi-dimensional data in a 3D map, where actors are localized in such a way that their Euclidean distance on the map corresponds to their social distance inferred from the original data, and their elevation corresponds to their social status or average frequency of communication.

Originally developed as a tool to prevent conflicts within teams of military personnel, the use of sociomapping was extended to long duration spaceflight simulations, and later on it was successfully used in business environments to analyze relationships within management teams. The basic principle of sociomapping (see an example of a sociomap in Fig. 3.3) is to map the data collected from a number of social actors into an artificial 3D landscape map. Transformation of the data is a matter of choosing some multi-dimensional social metric that could be reasonably interpreted as distance, and mapping it into a 2D coordinate system through an optimization procedure.

The algorithm for data transformation is based on a nonlinear dimensionality reduction technique. For example, nonlinear principal component analysis (NLPCA) uses backward propagation of errors (back-propagation procedures) to train a multi-layer artificial neural network (a perceptron) to fit a manifold. In general, the back-propagation method calculates the gradient of a loss function with respect to all the weights in the network. Then the gradient becomes input for the

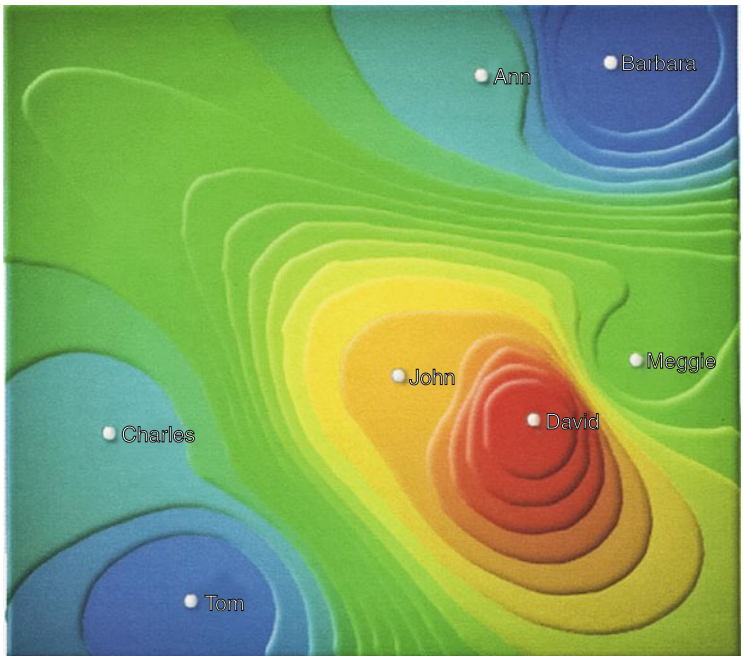


Fig. 3.3 An example of a sociomap. The elevation shows on average how actively each team member communicates, while the distance between individuals illustrates the intensity of mutual communication

optimization method, which in turn uses it to update the weights, in an attempt to minimize the loss function. Consequently, the NLPCA method updates both the weights and the inputs, which are thus considered latent values. After training, the latent inputs are a low-dimensional representation of the observed vectors, and the multi-layer neural network maps from that low-dimensional representation to the high-dimensional observation space.

Besides the distances between the group members, the sociomap method shows additional variables coded in the height and color of the subject (see Fig. 3.3). The height may represent the social status, performance indicators of the subjects, and usually the average communication frequency. For large systems and populations, sociomapping is a data mining approach based on visual pattern recognition. The social data associates a preference vector to each social actor and hence determines a position on the map. The weight for a social actor is found by calculating a sort of norm of its vector of preferences.

There are other similar methods providing visually interpretable model-based representations of network relationships. In [113], for example, the authors describe the probability of initiating relations between the social actors through their positions in an ‘unobserved social space’.

Apparently, the connection between the physical space (or Euclidean geometry) and social network geometry and topology (especially in the case of networks describing linguistics, psychology, or behavior relationships) is rather a convenient metaphor, or a highly visual means of presenting data. On the other hand, non-Euclidean geometries are more successful in the mathematical modeling of social networks, as we will show in the following section. In a highly intriguing paper, a group of computational neuroscientists from Japan [114] demonstrated that "neuronal activity in the human parietal cortex (see Fig. 3.4), which is involved in the spatial processing of self-referential physical distance, seems to be associated with the evaluation of social distance between self and others." Their study proved that the human parietal cortex is a member of the social brain network. The authors performed experiments which involved arranging dolls on a stage. Subjects showed a tendency to think of social compatibility as a Euclidean distance from a "self-representing doll that brought their egocentric viewpoints." This research suggests that we have an ability to judge human relationships in terms of spatial relations. Another positive identification of how the way our ontological brain used to process the physical world has evolved into an extended function in human social cognition.

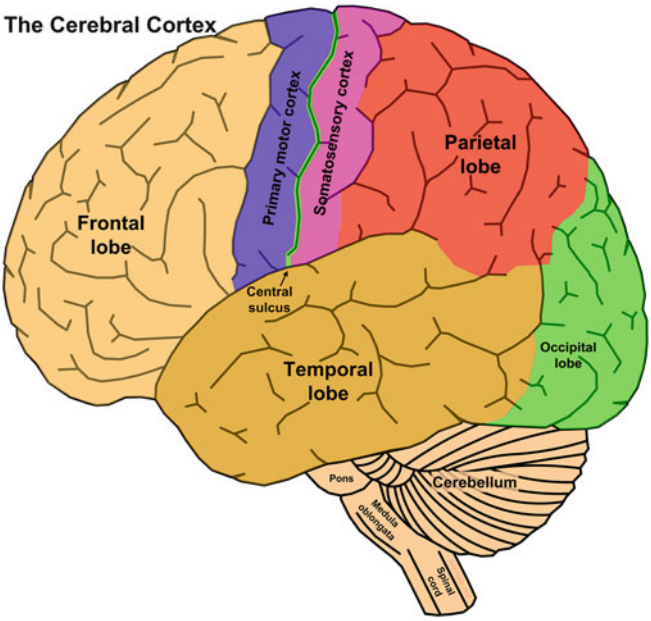


Fig. 3.4 The domain of the brain called the posterior parietal cortex, the *light blue area* in the figure, seems to be responsible not only for spatial sense, navigation, and touch, but also for social networking

3.6 Social Distance: Ultrametric

In spite of the unprecedent degree of connectivity of long-distance events through multiple parallel modern means of travel and sophisticated communication media, Expert et al. consider that the 'death of distance' has been greatly exaggerated [115]. The entangling of physical and virtual worlds, which happens also in economic and biological networks, does not exclude a spatial embedding which strongly influences their topological organization. Consequently, a metric- or distance-based social network theory should have some roots in the physical space. It is possible that a Euclidean metric may not work very well for social systems. At the same time, a social distance must not be defined in terms too remote from traditional Euclidean types of metric [114].

The axioms of metric spaces are too general to describe the complexity of social systems. Szell et al. show in [116] that networks with a positive connotation (friendship, private messages, trades) are strongly reciprocal, and their node pairs rather form bidirectional connections, whereas networks with a negative connotation (enmity, attack, bounty) all show significantly lower reciprocity.

In order to create more realistic models, Freeman proposed [108, 117] the use of non-Archimedean spaces, and consequently the use of an *ultrametric* instead of a metric based on the triangle inequality. The ultrametric space has a distance $u(x, y)$ defined by the properties

$$u(x, y) \geq 0 \ , \qquad u(x, y) = u(y, x) \ , \qquad u(x, y) \leq \max_z \left\{ u(x, z), u(z, y) \right\} \ ,$$

where the last relation is the so-called strong triangle inequality. The benefit of using ultrametric distances is that, for every given k, the graph with edge set $E_k = \{(x, y) | 0 < u(x, y) \leq k\}$ is a perfectly transitive graph, meaning that it consists of a number of mutually disconnected cliques (the same advantage for which the Markov models are also preferred statistical social models). Snijders [108] notes that ultrametrics are useful structures for representing the transitivity of social networks.

In an ultrametric space, several interesting things can happen. Triangles are always isosceles with the unequal side being shortest, and every point in a given 'disk' is a center of that disk. Two disks can intersect only by having one completely contained in the other (actually a 'disk' in this space is the closed subgraph emerging from one given node). Holly [117] imagines an ultrametric space as having its points on a tree. The ultrametric distance between two points x and y (see Fig. 3.5) is defined by the differences between their heights in the (inverted) tree and the height of their common generic node. The distance $u(x, y)$ is greater if they originate from farther nodes. For example, in Fig. 3.5, x and y are at level 4, and their closest common node is at level 1, so $u(x, y) = 3$. Indeed, if we choose for z any of the other nodes, a, b, c, d, e, or f, we can check directly that the strong triangle inequality holds.

Apparently, the ultrametric property of social distance arises from its binary tree graph structure. However, from a topological perspective, the points representing the actors are still in the Euclidean plane, except that they are not allowed to be

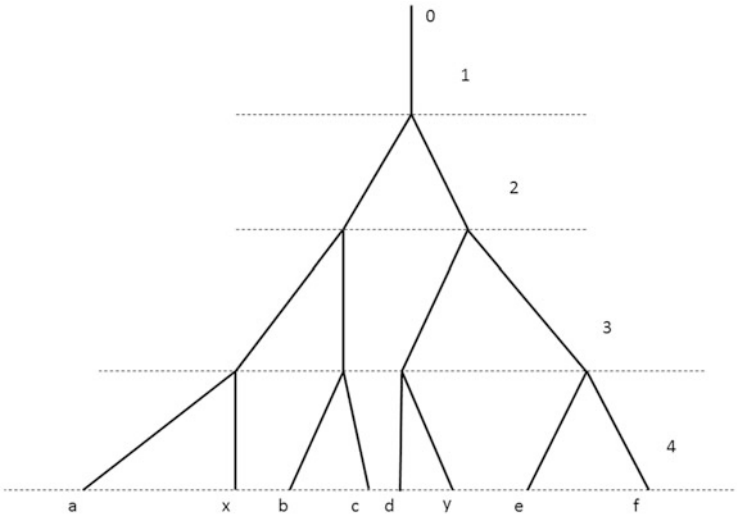


Fig. 3.5 Example of a binary tree. The set of the end nodes $a, \ldots, y$ is structured like an ultrametric space

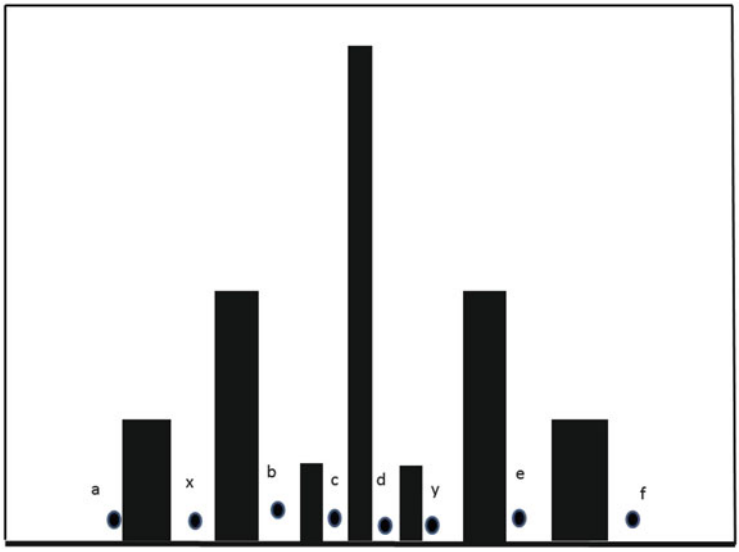


Fig. 3.6 The same ultrametric space as in Fig. 3.5, except that it is not now generated by a tree graph, but as a multiply connected domain

connected by the shortest Euclidean distances, but only through the graph paths. The same structure can be realized in another way. We can neglect the mandatory tree paths, and instead carry out the construction as in Fig. 3.6. In this case, the

points $a, \ldots, y$ can be connected by any smooth paths lying anywhere in the plane, while the strong triangle inequality is still satisfied. Moreover, the former binary tree graph generates an obstacle pattern which is similar to a Cantor set, itself a generator of multi-fractal structures. We consider this geometric intuition to be closer to the essence of the social system topology, and also related to the concept of social boundaries. The ultrametric distance occurs in this way as a consequence of the topology of the social space itself, and not through a forced graph-like relational constraint.

3.7 Social Topological Boundaries

The working topological space for social systems with boundary cannot be the set of actors or actants, because even if we include large populations, we still have a finite set, and any topology ends up being trivial. Very few social models approach the concept of social boundary from a mathematical perspective, and while some models involve this concept, the way it is represented is through interface migration processes. Some social models use the term *social periphery* in conjunction with social distance, whatever definition is used for distance. In general, it refers to people's distance with regard to social relations, especially in urban sociology. For example, it is often implied that the social distance is measured from the dominant city elite. In such a context, the social periphery of a city is rather its center. Another often used term in social urbanistics, especially true for global cities, is the *locational periphery*, which describes places physically distant from the heart of the city. These places often include suburbs and are socially close to the core of the city. It is even noted that, for practical purposes, the centers of two cities are often closer to each other than to their own peripheries.

In order to understand the function of the social boundary in a social mathematical model, the best approach is the topological one. In Chap. 4, and in particular in Sects. 4.2 and 4.3, we define the boundary concept from the mathematical point of view, in terms of topology and geometry, respectively. The geometrical definition of the boundary involves the concept of smooth manifold, and consequently needs to relate the working social space to a real n-dimensional space. Such an assimilation is not always possible, and it would induce limitations on the mathematical social model. The topological boundary, on the other hand, does not necessarily relate to the property of finite dimensionality of a space: it can be equally well defined for infinite-dimensional spaces. This constraint immediately eliminates spaces of actors from such a definition of the boundary, because the number of actors in a social system will always be finite, and we elaborated on this above. Let us recall the definition of the topological boundary:

Definition 1 For a set A in a topological space X, the point $a \in A$ of this set belongs to its boundary, $a \in \mathrm{Fr}\, A$, if a is not an interior point for A.

Definition 2 A point $a \in A$ of the set A is interior to A if it has a neighborhood included in A, that is, $\exists V(a) \subset A$.

In order to use the concept of topological boundary in social systems, we have to introduce rigorous definitions for topology and neighborhoods (or open sets). Consider a social system and take s to be a very well defined and without any doubt social and human entity. For example s can be a social act, a social event, some group of people, a judgment, social constraint, planning, a social actor, etc. Let X be the set of all these possible entities s, $s \in X$, and let us call this set a social space. We assume that a social space satisfies the basic Boolean rules from set theory: existence of an empty set, belonging relationship, inclusion, union, intersection, and complement of a set with respect to the social space. In order to define a topological structure on a social space X, we need to associate to each of its points s a system (a collection) of subsets called neighborhoods of s, denoted $N(s)$, having the following properties:

1. The element s belongs to each of its neighborhoods, i.e., $s \in N(s)$.
2. Any set which includes a neighborhood of s is a neighborhood of s itself.
3. The intersection of two neighborhoods of s is also a neighborhood of s.
4. For any neighborhood $N(s)$ of s, there is a proper subset $M \subset N(s)$ containing s, but distinct from $\{s\}$, such that $N(s)$ is a neighborhood for all points of M.

The first three constraints are trivial for almost any system of sets in a social space, while the last, the most difficult to satisfy, involves the transcendence to infinity of the collection of neighborhoods.

Let us present an example of the construction of such a social space and its topology defined by a process of contraction. For example, we choose a certain society, observe it during a given time interval, and note every single social act related to it, realistic, hypothetical, historical, planned, possible, etc. These entities form the social space X. Choose an element of it, let us say $a =$'school'. We define a neighborhood $N(a)$ by the set of all social entities related to any possible or real school in that society. Obviously, the set of real events related to a school, that actually happened in that society, form a subset of $N(s)$, and s belongs to this subset, being itself the concept of school. But any school built in that society is itself a social event of type 'school', so $N(s)$ is also its neighborhood. By applying any type of contraction on the general concept of school we generate neighborhoods of a.

Let us practise the concept of topological social boundary. We build DB as the set of all social events and social entities related to a city, for example Daytona Beach in Florida, USA. Let us denote by **bw** the social event known under the name 'Bike Week', which is a ten-day motorcycle event and rally held annually in Daytona Beach, when approximately 500 000 motorcyclists gather from almost everywhere. During this week the motorcyclists gather day and night, generally along Main Street. Let BW be the set of all possible social events in Daytona Beach related to this Bike Week element, $\mathbf{bw} \in BW \subset DB$. Food preparation and management for this week are examples of neighborhoods of **bw**. Very close to Main Street there is 'Papa John's Pizza' restaurant. Of course, during Bike Week, this restaurant is

open, and the restaurant is deeply related to the event through its many motorcyclist customers, so we can consider the element **pjp** =‘eating at Papa John’s Pizza’ as belonging to the set BW, i.e., **pjp** $\in BW$: “while going to the Bike Week we will eat every evening, and have a good time at Papa John’s Pizza.”

At the same time, the social life at Papa John’s Pizza is always at the center of other social entities, like supplies, banks, financial events, cleaning teams, paychecks, photographs, kitchen accidents, birthday parties, etc. There are always customers, waitresses, and drivers deeply related socially to this restaurant who do not join the Bike Week. None of the above-listed social entities, points of the space DB, belong to the Bike Week. Any neighborhood of Papa John’s Pizza social life contains events and entities not related to Bike Week, so any neighborhood of Papa John’s Pizza has a part completely disjoint from the set BW. In other words, there is no neighborhood of the **pjp** element which is completely in the set BW. So **pjp** is not an interior point of BW, and it is thus on the topological boundary of BW. All restaurants involved geographically in the Bike Week event are actually at the boundary of this event from the topological point of view. Such restaurants, as well as other similar organizations, companies, associations or institutions form the topological boundary of the BW set. We thus see that social systems may have their boundaries geographically densely embedded in them.

The same reasoning is applied, for example, by Castells [100] when he talks about the *global city*. A global city extends to spaces located in many cities around the world, and it is made up of territories from different cities connected socially in global information networks: “Thus, a few blocks in Manhattan are part of the global city, but most of New York, in fact most of Manhattan, is very local, not global. These globalized segments of Manhattan are linked to other spaces around the world, which are connected in networks of global management, while being loosely connected to their territorial hinterlands.” The global city is a union of disconnected territories.

3.8 Social Topological Patterns

Social or economic boundaries involve patterns in the social landscape. Economic and sociological spatiotemporal patterns belong to the class of phenomena far from thermodynamic equilibrium, among other ubiquitous natural processes like turbulence in fluids, interface and growth problems, chemical reactions, and biological systems [118].

3.8.1 Growth Models

In order to model the dynamics of a social boundary we can apply the *growth model*. Such a model is $(1 + 1)$-dimensional, involving one space and one time dimension.

In this mathematical model, the domain consists of a pile of identical rigid blocks, all stacked in columns. The boundary is represented by the upper line delimiting the last blocks added. In other words, it is the height profile of the pile of blocks at a given moment. The growth is driven by an independent Poisson process for each column of blocks falling from above and accumulating on top of the growing stacks of blocks (as in a simple 'Tetris' game with identical, non-rotating blocks). We apply the growth model by assuming that a certain social group can be represented by a connected 2D bounded domain in the plane, divided into equal regular cells, i.e., squares, and that the curve describing its boundary is an interface that can grow by addition of new cells. Each such cell represents a member of the social group. Assume that new people are approaching the group from a given direction and try to join it. These people will be represented by a set of new cells coming from outside the domain and approaching it along different, yet unchanging directions, all directions pointing towards the center of gravity of the domain. From the way we model the procedure of affiliation and acceptance of new members into the social group, we have three main types of social boundaries:

1. **Superficial Social Boundaries.** It is enough for an outside person to contact any member of the group only once in order to be accepted. The rest of the group relations adapt and the new member's connections propagate by transitivity. In this case, we can apply the *ballistic deposition growth model*. In this model, a new block sticks to the first block it touches: in full frontal edge-to-edge contact, through a corner touch to its left, or through a corner touch to its right. For example, a block can stick to the corner of a tall column even if it was falling directly over a low column immediately underneath it. A sudden modification of height can occur like that, and it breaks the independence of the column heights, so it introduces spatial correlation. The social boundary grows with inclusion of holes, but grows about the same way everywhere. The social group is more fragile, yet grows faster and more uniformly (see Fig. 3.7 middle row). Large irregularities in the group boundary are rapidly filled and the interfaces smooth out quickly. This type of boundary is spacetime uncorrelated.
2. **Non-Interactive Social Boundaries.** When a new member arrives along a certain direction with respect to the group, it continues advancing along this direction until it reaches the boundary of the social group. The new member does not deviate from its initial direction, i.e., it does not diffuse. However, it stops only at a terminal stop, and not at the first contact. This process can be modeled by the *random deposition growth model*. Such a process is realized when the blocks fall vertically, and stop only on top of the first frontally encountered column (see Fig. 3.7 upper row). This boundary is highly irregular, nonuniform, and generates fractal patterns [119].
3. **Structured Social Boundaries.** A new member can arrive from a given direction and intersect the group at some point. However, it will not stop at the first intersection. Instead, it will try to find a close niche and advance as far as possible towards the center of the group, without changing its incoming direction, and without jumping or crossing other members. The assimilation of each new

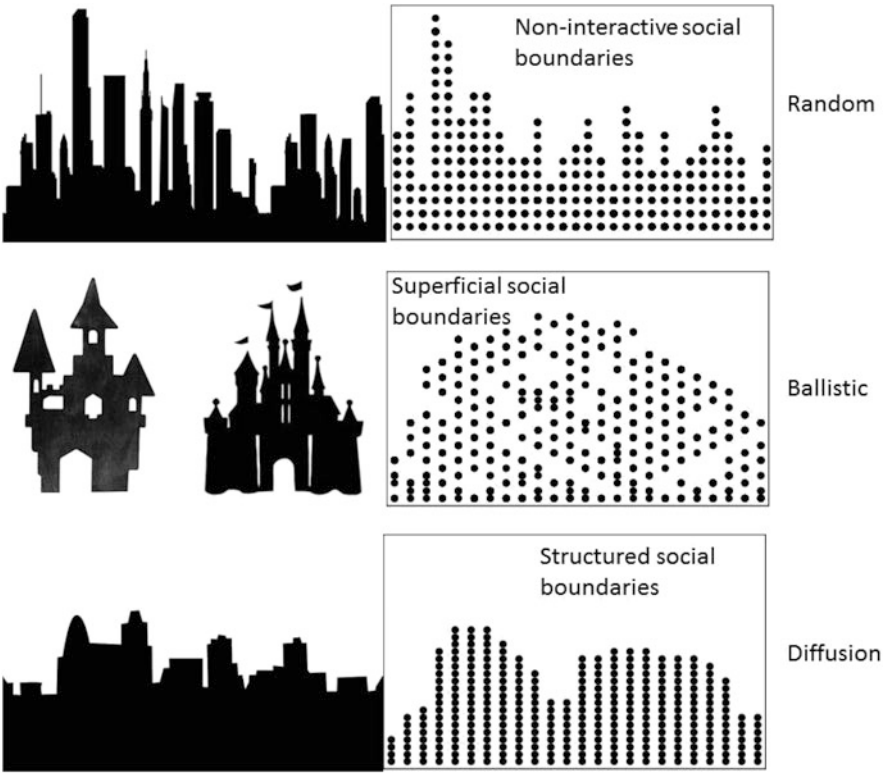


Fig. 3.7 *Right*: Examples of social boundaries described using growth models. *Left*: City silhouettes similar to the corresponding growth model on the right

member is a long time-scale process during which the new member gradually diffuses into the group until it finds the most appropriate, or compatible position for itself, depending on the local balance between individual identity, motivation, and resistance. This is related to the *random deposition with relaxation growth model*: blocks fall vertically, temporarily stop on top of the first frontally encountered column, but then keep trying to advance by diffusion towards the center of gravity of the group, if the neighboring column has a lower height. The new member keeps advancing like this until it reaches the deepest available position inside the group. In terms of our free fall example, that means until the new member reaches the lowest possible height (see Fig. 3.7 bottom row). This is the most stable social group. It is very homogeneous and its boundary is smooth.

3.8.2 Cooperation and Patterns

Complex systems in life and human sciences are collections of many living individuals interacting in a nonlinear manner. Examples are vehicular traffic, crowds, swarms, clews, flocking phenomena, animal epidemics, complex biological phenomena, and social dynamics. Cooperation is a key force in evolution, present at all scales of organization, from unicellular organisms to complex modern human societies. The presence of structure means that each individual interacts not only with every neighbor, but also with a small subset of the population which constitutes its neighborhood, and it is connected according to an underlying network of relationships [120]. The current view on the influence of spatial structure, as a particular case of population structure, is that it usually promotes cooperation.

It is generally thought that any form of associative interactions which include spatial structure would favor the evolution of cooperation. Such an association can lead to the formation of clusters of cooperators that can maintain cooperation against defecting invaders at the cluster boundaries. Evolutionary spatial games assume randomly interacting populations. These types of game are played on a 2D square lattice. Each position is occupied by either a cooperator or a defector. In each generation, the payoff of a certain individual is the sum over all interactions with the eight nearest neighbors and with its own site. The game is deterministic, and its outcome depends on the initial configuration and the parameters of the interaction matrix. Defectors can invade a population of cooperators or vice versa, and this competition generates surprising patterns. An interesting sequence of patterns emerges if a single defector invades a world of cooperators. Spatial chaos, dynamic fractals, kaleidoscopes, royalty, crowns, crosses, and *lys de France* are observed (see Fig. 3.8).

Of course, in real life the cooperation dynamics is more complex, and less ideal laws act upon populations and systems. The dynamics of collective systems of living individuals often involves clear boundaries. For example, flocking birds establish a swarming behavior with well-defined boundaries holding the birds in. The same behavior is noticed in schools of fish, where the boundary is even more structured into different layers. The dynamics of flocks or schools can be analyzed either in terms of flow equations for fluid-like behavior, or (as it has been shown recently) in terms of shattered brittle solids. An interesting real example of cooperation in terms of geometry and the laws of physics is provided by a clew of living tubifex (or red) worms. In the aquarium department of a pet shop, one may find live food for aquarium fish. Among a wide variety of products, they may offer little piles of tubifex worms. Normally, especially in winter time, the worms are entangled in a clew. The worms are relaxed, or 'sleeping', and the clew assumes a Gauss bell shape due to gravity, water adhesion, and entanglement. If someone wants to buy such a clew, the rule is that they must knock the table in order to 'wake up' the worms and check whether they are alive. What is interesting is that, when the worms are woken up like that, they immediately adopt the shape of a perfect sphere, sitting on the table like a ball, with the plane of the table as tangent (see Fig. 3.9).

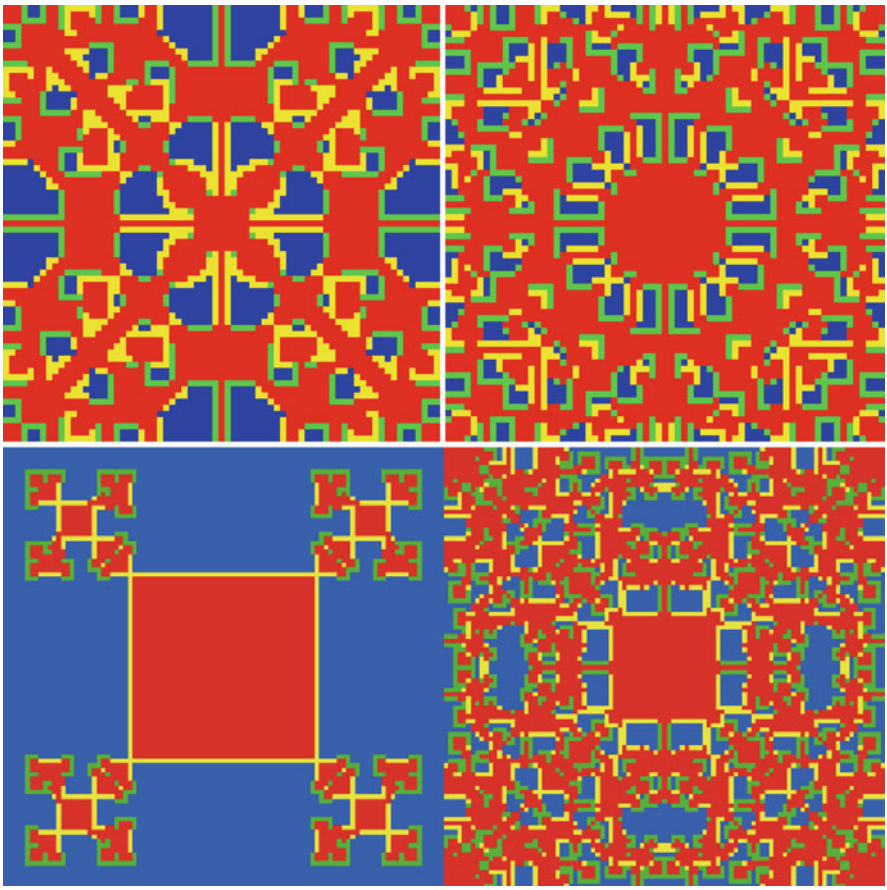


Fig. 3.8 Games played on a 2D lattice where cells are occupied by a cooperator or by a defector, and the payoff is the sum over the nearest neighbors. This competition generates surprising patterns

The explanation may be this: when woken up, the worms feel insecure and want to shrink and reduce their interface with the environment. Through an interesting collective behavior, the whole clew adopts a minimum surface area and so becomes spherical. It is interesting to model the propagation of the minimal area constraint of an individual, through entanglement interaction, towards the external shape of the system.

3.8.3 Multivariate Networks

Complex social systems (like human societies) are characterized by the superposition of their relations defined on the same set of social actors (nodes). Each type of

Fig. 3.9 A clew of living red worms. They are relaxed, or 'sleeping'. If their bed is vibrated, the worms 'wake up' and act in defense: shrink and reduce their area, their interface with the environment. This individual behavior has a spectacular collective effect on the geometry of the clew. The worm clew adopts a spherical shape, as though they 'know' that this is the minimal exposure area

relation spans a different social sub-network of its own, and all these individual sub-networks co-construct each other. Their superposition, called a *multiplex* or *multivariate network*, determines the final topology of the whole network, for the shape of each sub-network influences the topologies of the others by acting as a constraint, inhibitor, or catalyst [116]. This multi-dimensionality of human relationships makes the multivariate social network into a structure embedded in a very high-dimensional geometric space.

When the number of dimensions of the space increases greatly, and if the shape under discussion remains bounded, like for example the phase space of a thermodynamic system in equilibrium, the points in the bounded domain begin to display a peculiar behavior, viz., they stick to a hypersurface. In some sense, the higher the dimension of the space, the closer the points are to their topological boundary. It is easy to exemplify this behavior by using the sphere example in a Euclidean space of dimension n. The volume and the area of such a sphere of radius R (see, for example, Sect. 3.2 in [121]) satisfy the equations

$$V_n = \frac{\pi^{n/2} R^n}{\Gamma\left(\frac{n}{2}+1\right)}\,, \qquad A_n = \frac{2\pi^{n/2} R^{n-1}}{\Gamma\left(\frac{n}{2}\right)}\,. \tag{3.1}$$

If we consider the points embedded in a spherical shell of thickness $\mathrm{d}R$ at the boundary of the sphere, their volume $\mathrm{d}V_n$ is given by

$$\frac{\mathrm{d}V_n}{V_n} = n\frac{\mathrm{d}R}{R}\,. \tag{3.2}$$

At the same time, if we sample the sphere of radius R into smaller spheres of radius $r \ll R$, the ratio between the number $N_{\Sigma,n}$ of small spheres lying on the outermost

layer of the big sphere and the total number N_n of small spheres lying inside the big sphere is

$$\frac{N_{\Sigma,n}}{N_n} \simeq \frac{r}{R} = \text{constant}\,. \tag{3.3}$$

Consequently, for very high dimensions, the volume of the outer layer of a sphere approaches the total volume of the sphere [see (3.2)]. We conclude that almost all inside points tend to be at the surface. At the same time, using another way of counting the number of external participants [see (3.3)], this ratio is independent of the number of dimensions.

These observations point towards the same conclusion as the results from the theory of clustering of high-dimensional data with the indefinite growth of the number of dimensions of the data set. It has been proved [122, 123] that the concept of distance becomes less precise as the number of dimensions d increases, since the distance between any two points in a given data set converges. The discrimination of the nearest and farthest point becomes meaningless:

$$\lim_{d\to\infty} \frac{\max D - \min D}{\min D} = 0\,,$$

where D is the distance between any two points in the data set.

Landherr et al. in [124] performed a series of numerical tests on various social networks, and the results are in agreement with the conclusions inferred from our simple evaluations in (3.2) and (3.3). These authors analyzed five different centrality measures and showed that four of the studied centrality measures have these properties, namely, the value of centrality decreases on average with the addition of supplementary edges/relationships. Only one out of four types of centrality measures has a different (monotonic) behavior.

This result, confirmed in several numerical experiments by Landherr et al. [124], and confirmed by our simple asymptotical evaluations in (3.2) and (3.3), is rather surprising since intuitively one might expect centrality to increase when supplementary relationships are added. In other words, the hypothesis that any centrality measure increases monotonically with the increase in the number of extra paths and relationships, for almost all types of social network, has been disproved by several systematic numerical experiments. The behavior of common geometrical network properties from very intuitive at low dimensionality to counter-intuitive at higher dimensionality is of great interest. As Landherr et al. conclude, "decision-makers should not uncritically rely on intuitively obvious statements from the application of centrality measures."

3.8.4 Pattern Formation in Unstable Social Systems

At large population levels, e.g., nations, the contemporary processes of globalization and integration are partly responsible for unleashing new types of social struggle, between centrifugal and centripetal forces that are far from stabilizing the social system. Such situations happen all over the world. This is exemplified by certain European countries which have become "post-modern states, freely pooling part of their sovereignty" [125]. The European Union is a complex system defined by a large number of mutual channels of interference in each others' domestic affairs. For example, the Scottish and Catalonian independence movements are predominantly driven by centripetal forces, towards unification, because European unity could provide an escape from their controlling national governments, while equal representation in the EU would cut out the 'middle men' (formation of NGOs). This is a process of fragmentation within unification which is typical in pattern formation structures.

Social morphogenesis and pattern selection in such complex systems involve the existence of at least two opposite interactions. On the one hand, for example, we have a limitative threshold against further fragmentation. The continuous fragmentation of nation states would increase the tensions within the unified system's decision-making system, risking sclerosis. On the other hand, we have the fragmentation tendency, driven by citizens' increasing demands for more democratic control at the sub-national level, which has the effect of suppressing the importance of national borders. When these two main tendencies, together with many smaller driving forces, fail to reach equilibrium, phase transitions occur, causing new patterns and new boundaries to form. This instability model can contribute to our modern understanding of pattern formation in such large and complex social systems.

Morphogenesis models can be borrowed from nonequilibrium physics and applied to complex social systems, with varying degrees of success. For example, the dynamical systems theory predicts that, if some control parameters change, a steady solution of the equation of motion around an attractor (a fixed point) may lose its stability, and one or more new stable attractors may appear. Thus, the system can pass through bifurcations to different symmetries and different patterns and cells with opposite symmetries. Another possible mechanism of morphogenesis is 'dendritic growth' controlled by diffusion and shape instabilities, the so-called Mullins–Sekerka instability which is a typical trigger for pattern formation. This involves the formation of bumps on the boundaries and solidification fronts that grow into fingers and dendrites.

Another prevalent theory of pattern formation is 'solvability theory'. This asserts that patterns are generated when a diffusion type of process contains a singular perturbation which completely changes the mathematical nature of the problem, no matter how infinitesimally weak it may be.

One major cause of pattern generation is also related to situations where the amplitudes of some disturbances increase to a level where nonlinear effects take

over, leading to chaotic dynamics in which many degrees of freedom are active. After an instability has produced such a growing disturbance, there must be an intrinsically nonlinear mechanism by which the system moves towards the new state. The system then evolves in entirely new directions determined by its nonlinear dynamics. For example, the economic crisis that began in 2008 has accelerated the opposing forces of centralization and regionalism in the EU, sharpened the conflicts between richer and poorer regions, and generated new social patterns and boundaries, e.g., the north–south divide in Italy, the Flemish and Walloons, Hungarians and Romanians.

The instability and nonlinearity effects are responsible not only for pattern generation, but for even larger scale phenomena. For example, increased stress in granular materials forms *localized stress chains*, where forces are carried primarily by a small fraction of the total number of constituents. The equivalent in social system dynamics would be a situation where a widely available and affordable technology has broken the government monopoly on the management of information. In this way, the number of players who matter is increased, while the number of individuals commanding great authority is reduced. Another typical effect occurring in highly nonlinear systems is the 'jamming' effect, where constituents are locked into local configurations from which they are temporarily unable to escape.

Given such a variety of effects that can trigger variability and pattern formation, a legitimate question is: how is it possible that nations, countries, and large organizations with so many degrees of freedom can still be ruled, especially since nonlinear effects ultimately lead to chaotic dynamics, where more and more degrees of freedom interact and become active? The theory of bifurcations in dynamical systems can help us to understand such an intriguing situation. The *center manifold theorem* indicates that, when a bifurcation occurs, the unstable trajectories move away from the originally stable solution only within a low-dimensional subspace. In other words, the relevant degrees of freedom of the bifurcation are rather few. It is enough to describe pattern formation processes through only a small number of dynamical variables. In sociological language, this important property of nonlinear systems can be expressed through the 'law of conservation of social energy', which says that nothing disappears in social (and political) life until its replacement has already been discovered and is functioning effectively.

Part II
Mathematical Language

The difficulties are almost always at the boundary

Gilbert Strang

In the second part of this book, having explored the concept of boundary and its importance for the human experience from the perspectives of art and society, we move from this humanistic point of view toward the mathematical concept of boundary and the associated mathematical language. However, in order to maintain an interdisciplinary approach throughout the book, we shall as far as possible use non-deductive reasoning rather than deductive proofs. This approach will not only be able to render the ideas more amenable to self-experimentation, but following the explosion of the computational era, we believe that present-day mathematics relies more and more on the use of 'mathematical experiments', visualization, and Bayesian approaches. Van Bendegem argues in his essay [126] that mathematical explanation is 'as nasty as any part of *la condition humaine*'. Here, by 'nasty' or 'ugly truth', he means that mathematics should be understood in the context of general science, and not as being orthogonal to the difficulties and disappointments generated by everyday experimentation. In the spirit of Martin Heidegger and Constantin Noica, who would classify the dispute between considering mathematical concepts as real-life or ideal as a third type of adversity, where neither side can claim to uphold the truth, we will present the mathematics of the boundary from an empirical mathematical point of view.

The aim with this approach is not to restrict the importance of the concept of boundary to a unilateral 'Platonist' view, considering boundary as a real existing thing, separate from any physical world features. We rather want to present boundary as a distillate of human thinking, as a response to experience of the real world, indeed an approach in which the concept of boundary is rather a socio-cultural construct, as Hersh[1] would identify it [127].

According to 'Intuitionists' our mind is ontogenetically and phylogenetically tributary to our senses, and mainly to the senses of sight (continuum: geometrical, topological) and hearing (discrete: algebraic, numeric). The presence of an internal

[1] According to Hersh, mathematics lies somewhere between a human construction and the real world: the number π is not only in our imagination, but transcends the individual mind, while remaining in a human world.

structure hosting algebraic-geometric judgments in our brain is suggested by several experimental results, like the spatial-numerical association of response codes (SNARC) effect and its relation to the spatial and verbal working memory. Namely, the occurrence of a faster reaction for left-side placed small numbers, and right-side place large numbers, and vice versa. For example, Brouwer, who is regarded as the creator of intuitionism [128], considers mathematical approaches to be languageless creations of the mind. The intuitionist approach to mathematics thus justifies the existence of only two types of proof: geometric and algebraic. Even if this is not quite correct, such a perspective will always help us to understand better the mathematical questions and answers. It should not be forgotten that seeing and hearing are the only long range senses, so the geometric and algebraic approaches must be able to offer an integral, global picture of each problem. This observation reminds us of the Indian story of the blind men and the elephant, in which a group of blind men (or men in the dark) touch an elephant to find out what it is like. When they compare their observations later on, they disagree completely, and this happens because each had a local (non-visual, non-acoustic, hence not long range integral) approach to the problem.

Following this idea, we divide this part of the book into two chapters, dealing with the elements of continuous and discrete mathematics as they relate to the analysis of boundaries. It would be difficult to produce a comprehensive list of all the questions and general areas relevant to the concept of boundary, and which should be covered in a book such as this. In order to present a fair number of interesting mathematical ideas relating to the definitions and properties of boundaries, in the 16 sections of Part II we have gathered together a selection of definitions, theorems, and examples from pure and applied mathematics that we feel are important in their own right when read together, and also useful in providing an adequate introduction to Part III, which contains supporting examples of the importance of the boundary concept in the sciences.

The division of this part into two chapters is a natural consequence of the two major approaches in mathematics: topology, analysis, and geometry in the first part, and discrete mathematics in the second. Of course, an overlap of these two subjects is unavoidable in applications, so we have added a final section connecting some of the concepts from discrete and continuous mathematics. The concept of continuity is related in one's cognition with an intuition of an unbroken or uninterrupted whole, like the sky or a sheet of metal. In a way, opposed to continuity is discreteness: to be separated, like grains of sand, chairs in a room, or the leaves on a tree. Continuity connotes unity; discreteness, plurality.

Continuous mathematics is classical and well established, while discrete mathematics is often identified with specific application areas like combinatorics, graphs, networks, or computer science. The key structures and methods of these two faces of mathematics are basically different (neighborhoods and limits in continuous mathematics and induction in discrete mathematics). The present state of mathematics is the product of a strong intrinsic logic, but also of historical events [129].

In Chap. 4, we present the elements of topology, analysis, and differential geometry. Section 4.1 introduces basic topology using simple examples, and we then discuss in more detail the main topological properties which are invariant under homeomorphisms: separation, compactness, and connectedness. Since the concept of boundary embodies a multitude of connections, we have developed some of these interesting relationships in the following sections.

Section 4.2 is dedicated to the definition of boundary from the point of view of topology, without considering a particular distance, metric, or norm. Then in Sect. 4.3, we present an analytic and geometrical definition of the boundary using the theory of manifolds. The central part of this chapter is unfolded in Sect. 4.4, where we present the elements of analysis in terms of differential forms, defining and explaining the Lie derivative and presenting some of its properties.

Section 4.5 reviews the theory of fiber bundles and the definition of the covariant derivative. A summary of the main properties and formulas which will be useful in subsequent applications are provided in this section. In Sect. 4.6, we present a detailed example inspired by theoretical hydrodynamics, of the differences, the similarities, and the specific uses of the Lie and covariant derivatives. Section 4.7 is devoted to the effect of perturbations of the boundaries in boundary-value problems, while Sect. 4.8 discusses a few interesting aspects of differential topology and cobordism theory.

Chapter 5 provides some of the elements of graph theory, especially those concerning boundary. In Sect. 5.1, we introduce graph structures relating to concepts of discrete mathematics. As pointed out in the introduction, a strong emphasis is placed on the relation between a graph and its boundary. This topic is continued in Sect. 5.2 with a presentation of some of the main results and theorems. In Sect. 5.3, we combine the discrete approach with the continuous one in the study of graphs. The graph content of this chapter is closed by the comprehensive Sect. 5.4, where we enumerate and discuss the main results from graph theory which have connections with the properties of their boundaries. There we discuss graph properties in terms of the number of nodes and links, the volume, the diameter, the girth, and so on.

Section 5.5 introduces some elements of algebraic topology, namely simplex theory and homology. Algebraic topology is a twentieth century field of mathematics, but it can trace its origins and connections back to the ancient beginnings of mathematics. Its content brings together continuous geometric phenomena as understood by discrete invariants. This topic is continued in Sect. 5.6, where we discuss more advanced aspects of homology and cohomology, and in Sect. 5.7, where we introduce triangulations and CW complexes. The chapter concludes with Sect. 5.8, in which we summarise the continuous and discrete mathematical approaches with a discussion of their interconnections and applications.

This part of the book can be completed with additional literature in order to deepen some of the concepts or to obtain a better grasp of the methods. As a good motivational introduction, we recommend supplementing the concepts presented in the first sections of this part by some of the literature in the philosophy of mathematics, such as [130], or an introduction to topology and geometry for physicists like [131]. The important topological definitions and theorems can be

found in greater detail in [132], and many of their applications in the field of functional analysis can be found in [133] or [134]. For current technical formulas and relations, we recommend [135]. A good introduction to three-dimensional differential geometry would be [136] or [137]. For extensive developments in differential geometry, one can refer to [138], and to [139] for differential geometry as applied to surfaces and curves. For specific topics in the theory of fiber bundles, we recommend [140]. For applications to theoretical physics, we recommend [141], while for traditional applications of geometry in mechanics and fluid dynamics, we recommend [142, 143]. These mathematical approaches can be completed with an intuitive work on hydrodynamics such as [144]. For a mathematical approach to the kinematics and dynamics of free fluid interfaces, we recommend [145] and also [146]. The topics presented in this chapter can be completed with the more recent ideas to be found in [121, 147, 148].

Chapter 4
Continuous Mathematics

A boundary is that which is an extremity of anything

Euclid

4.1 Intuitive Introduction to Topology

Besides its essential role in the development of calculus, analysis, geometry, and algebraic geometry, topology has major direct applications and contributions in the study of image reconstruction and recognition, modeling, graphs, and networks, fluid mechanics, protein folding, robotics, and fundamental physics. Topology is the part of mathematics that investigates space from a qualitative point of view, roughly speaking without using the concept of distance. It may seem counter-intuitive to do geometry in a metric-free world, but the habitual dualities small/large, or near/far, etc., are more complicated than they appear because they need a comparison relation with a unit of measurement, and the order relation in the set of positive real numbers. Topological intuition does not need these concepts. It is to geometry what logic is to algebra. In his book on topology, Pavel Alexandrov says [149]: "The specific attraction and in large part the significance of topology lies in the fact that its most important questions and theorems have an immediate intuitive content and thus teach us in a direct way about space which appears as the place in which continuous processes occur." An example is given in Fig. 4.1, where a mug is smoothly deformed into a zero volume surface, the Klein bottle.

The reader is considered to have a *sensus communis* of the following concepts: set of points, element of ($\in$), and inclusion ($\subset$), as well as the basic operations with sets, viz., union ($\cup$), intersection ($\cap$), Cartesian product ($\times$), and complement ($\complement$). There will be no need to use axiomatic set theory. We also consider known the concepts of function, domain of definition, and range, denoting all this by $f : D \to R$. A useful set operation for topology is the *disjoint union* of two sets A, B, defined by

$$A \sqcup B = \Big\{(x, i) | i = 1 \text{ if } x \in A,\ i = 2 \text{ if } x \in B\Big\} \subset (A \cup B) \times \{1, 2\} \,. \tag{4.1}$$

Topology begins with the introduction of an abstract set X of points x, and we write this as $x \in X$. Inside X, we can define subsets of points $A, B, \ldots$, grouped by various

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_4

Fig. 4.1 Example of topological deformation and homeomorphism. Courtesy of AKME Klein Bottle at www.kleinbottle.com/index.htm

criteria, and we write $A \subset X$, $B \subset X$. So far, the only property we have from this subset structure is that we can enumerate elements or subsets which are all parts of X.

The boolean algebra of the subsets can be further enriched by introducing more structures. This is possible in two ways: either by defining relationships between the elements (algebraic structure), or by defining relationships between the subsets (topological structure). The set X endowed with such a structure becomes a *space* (algebraic structure or topological space, respectively). We can even simultaneously define both types of structure on the set, but in this case we need to satisfy compatibility relations between the structures.

Working with spaces instead of sets is a mathematical habit that proves valuable for the investigation of more complicated structures. For example, one can compare a newly defined space Y with a model space X by defining a *morphism* $M : X \to Y$ from X to Y, which induces in Y a structure similar to the model one in X, i.e., the morphism M preserves the structure of X. If the structure is topology, then the map is a *homeomorphism*, and if the structure is algebraic, then the map is a *homomorphism*. A homeomorphism between two topological spaces is expected to preserve all intrinsic topological properties.

Topologists often say that there is no difference between a pretzel and a doughnut, or between a coffee mug and a doughnut (see again Fig. 4.1). Under the procedure of building homeomorphisms, one can study properties that are independent of distance, size, or shape. Among different types of topological structures defined on a given set, the most useful is the *point set* topology. We build a topological space from a set X by arbitrarily choosing a family of subsets τ of X, including the empty set and the set itself, which is stable under finite intersection and finite or infinite union. The set X becomes a topological space (X, τ) and the subsets in τ are called *open sets*. The complement of any open set is called a *closed set*. The total space and the empty set are the only sets that are open and closed simultaneously. In particular, if the set X is finite, like a graph or a network, all unions and intersections are finite which makes its topology even simpler: no need for infinite unions. Then it is easier to count how many distinct topologies can be constructed on a network.

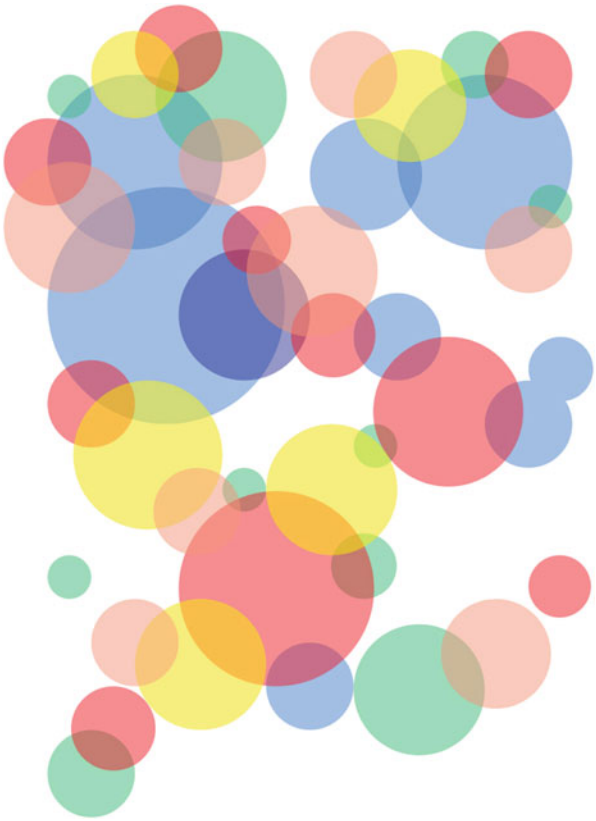


Fig. 4.2 Example of different degrees of separation in topology. *Overlapping Circles* by Jonathan Butler. Courtesy of the artist www.j-butler.com

An artist's view of a topological space might look like Fig. 4.2. The circles and their intersections are all open sets. Any set containing an open set which contains the point x is called a *neighborhood* of x, and this is denoted by $\mathcal{V}(x)$:

$$A \in \tau \ , \quad x \in X \ , \quad x \in A \subset \mathcal{V}(x) \ .$$

If the topological space also has a distance defined on it, whence it is a metric space, we can always introduce a distance-based system of neighborhoods called ϵ-ball neighborhoods. We define such a *ball neighborhood* for the point x to be the set of points placed within a distance ϵ from x.

A point is said to be *adherent*, or a limiting point, to a set A if all its neighborhoods have nonempty intersection with A. A closed set contains all its adherent points. An adherent point is the generalization of the concept of limit, and a point which is not adherent to a set is said to be *isolated*. The *interior* Å of a set $A \subset X$ is the largest (in the sense of inclusion) open set still contained in A,

while its *closure*, denoted $\bar{A}$, is the smallest closed set containing A, so that

$$\mathring{A} \subseteq A \subseteq \bar{A} \, .$$

If the left inclusion is an identity, A is open, and if the right inclusion is an identity, A is closed. The *topological boundary* of the set A is

$$\partial A = \bar{A} - \mathring{A} \, .$$

The boundary of a disk in the plane is its circumference, and the boundary of a finite set of points is the set itself. If we have two sets $A \subset B \subset X$ with the property $\overline{A} = B$, we say that A is *dense* in B. The set of rational numbers is dense in the set of real numbers. A function defined on X with values in Y is *continuous* if the inverse image of any open set in Y is open in X. A one-to-one (bijective) continuous function is called a *homeomorphism*.

4.1.1 Separation

A topological structure (X, τ) provides a way to 'measure' the degree of separation of $x, y \in X$, because there are in fact many degrees of separation apart from just having $x \neq y$. There is a very large set of different axioms of separation, each useful for a certain application in functional analysis. Below, we list only five separation criteria as a minimal meaningful set for our purpose:

- A space is T_0 or *Kolmogorov* if, for any two of its points, one of them has a neighborhood which does not contain the other one.
- A space is T_1 or *Fréchet* if, for any two of its points, we can find distinct neighborhoods.
- A space is T_2 or *Hausdorff* if, for any two of its points, we can find disjoint neighborhoods.
- A space is T_3 or *regular* if, for any point and any closed set not containing the point, we can find disjoint neighborhoods for both of them.
- A space is T_4 or *normal* if any two disjoint closed sets can be included in two disjoint neighborhoods.

In Fig. 4.2, let us define a topology in the figure frame by considering all colored circles, all their intersections, and all their unions as open sets. The points belonging to only one circle are not separated. Let us choose the two, green and pink, intersecting circles in the lower left corner, and let p be the center of the pink circle, g the center of the green circle, and i a point included in their intersection. We see that p and i are T_0 separated, and not T_1 separated, or indeed separated in any of the other ways. However, p and g are T_1 separated, and p and the center of any blue

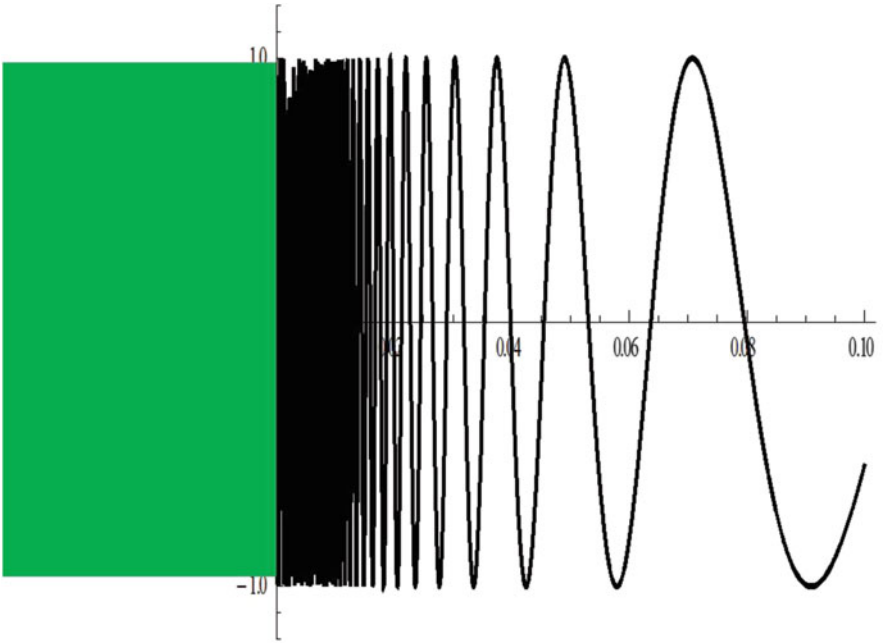


Fig. 4.3 The function $\sin(1/x)$ and the rectangle $(-0.05, 0]\times[-1, 1]$ are disjoint, but not separated

circle are even T_2 separated. In Fig. 4.3, we present an example of two sets that are disjoint but not separated.

The concept of separation provides a way to study problems in which one cannot build or write down an explicit solution of an equation, but may be able to discuss the qualitative properties of any solution. The separation of a topological space tells how many distinct solutions may exist for a given equation, and so the separation criteria are essential in identifying uniqueness of solutions. The most trivial example concerns the uniqueness of the limit of a convergent sequence, e.g.,

$$1\,,\ \frac{1}{2}\,,\ \frac{1}{3}\,,\ \ldots\,,\ \frac{1}{n},\ldots\,.$$

The numbers in the sequence get smaller and smaller, but they remain positive as they approach zero. In the limit when n is very large, zero is a good candidate for the limit of the sequence. Indeed, any open interval centered at zero contains all the numbers in the sequence except for a finite number. This observation makes zero an adherent point of the sequence, hence a limiting point. But this is not enough to declare that zero is the only limiting point of this sequence, and it is only the type of topology which dictates whether the limit is unique.

Here is another example of a topological space: My neighbor has the most replete mechanics workshop in the neighborhood. Let T be the set of all his tools,

and construct a basis of open sets of tools τ from sets of tools performing only specialized types of operations. My neighbor calls them 'tool kits', but they are the open sets of his workshop's topology. The set of all screwdrivers S, for example, is an open set and the set of electrical tools ET is also an open set. The intersection of these two sets constitutes the set of screwdrivers with insulated handles, those that can be used only on screws in electrical installations, so $S \bigcap ET \in \tau$. Unions of open sets of tools are also open sets because they represent larger tool kits, for more complex operations. My neighbor's garage with his tools organized in open sets is a topological space.

My neighbor would not keep two identical tools. He decides that two tools A and B are redundant, that is 'not separated', if any tool kit that contains A also contains B, and conversely. Take any two tools from his garage, say two hammers, and then pick a rubber mallet and a splitting maul. My neighbor confirms that the mallet belongs to the auto repair kit, and he has another kit of lumber tools that contains the splitting maul, but these two kits have nothing in common. So he has two disjoint open sets that each contain only one of the two hammers. Consequently, the two hammers are separated, and he can keep them both. His garage is a Hausdorff topological space, or a T_2 separated space.

My neighbor is working on a 'universal tool', i.e., one that can carry out any repair. The question is whether he can patent it, so that in essence the tool is unique. Let us pretend he has built this tool UT. It thus belongs to all open sets in T by definition. It follows that any sequence of tools converges to the limit UT, no matter how my neighbor labels his tools in sequences. Pretend by *reductio ad absurdum* that this universal tool is not unique and that there is another universal tool $UT' \neq UT$. If these two universal tools are distinct according to my neighbor's rule, he should find two disjoint open sets of tool kits each containing only one of the universals. But each universal tool is contained in every open set so the above assumption is false. There is no second distinct universal tool in his garage, so we have exemplified a version of the theorem according to which *the limit of any convergent sequence in a Hausdorff space is unique*.

The following is an example of a non-Hausdorff space. In the real plane $\mathbb{R}^2$, consider the union of three semiaxes: $x \geq 0$ and $y = -1, x \geq 0$ and $y = 1$, and $x < 0$ and $y = 0$. Let us induce a topology on this space as follows: a set of points $(x, y) \subset \mathbb{R} \times \{-1, 0, 1\}$ is open if x belongs to an open interval of the real axis. We note that the points $(0, -1)$ and $(0, 1)$ are distinct but they have no disjoint neighborhoods. Indeed, any open set of these two points will contain negative x numbers, which means numbers on the third semiaxis $x < 0, y = 0$, so their intersection is never empty.

Here is another example of a non-Hausdorff space. Let (X, τ) be a space such that each point has a countable neighborhood basis. For each point $x \in X$ there exists a sequence $V_1, V_2, \ldots$, of neighborhoods of x such that, for any neighborhood $V(x)$ of x, there exists an integer i with V_i contained in V. We call such a space *first-countable*. Let us have two distinct points $x \neq y$ in X that cannot be separated by open neighborhoods. Each such point has its own basis of open neighborhoods $V_1(x), V_2(x), \ldots$, and, $U_1(y), U_2(y), \ldots$, respectively. For each i, we have $V_i(x) \cap$

$U_i(y) \neq \emptyset$, because the points are not Hausdorff separated. We choose a sequence $z_i \in V_i(x) \cap U_i(y)$. Such a sequence converges to both x and y.

4.1.2 Compactness

The simplest functions are the integer powers and their linear combinations, the polynomials. It is easy to calculate polynomials and computers can do this quickly. Nevertheless, polynomials contain enough complexity to be able to fit almost any complicated function, under some topological restrictions. The Weierstrass theorem and its generalizations [132, 133] state that every continuous function defined on a closed and bounded real interval can be approximated as closely as desired by polynomial functions:

Theorem 1 *For any continuous real-valued function f defined on the real interval $[a, b]$ and for every $\epsilon > 0$, there exists a polynomial $P(x)$ such that, for all $x \in [a, b]$, we have $|f(x) - P(x)| < \epsilon$.*

The real axis with topology defined by open intervals is Hausdorff separated. What are the essential ingredients that contribute to this powerful approximation? There are two: continuity of the function and closure and boundedness of the interval. Boundedness, as one of the key concepts, is related to distance for the real axis, but in topology one can use a more general concept called *compactness*. A set $C \subset X$ is compact if, from any open covering of it, i.e., any set of open sets whose union includes C, we can extract a finite sub-covering.

It is easy to provide an example of a noncompact set. The real axis is noncompact because, if we cover it with open intervals of length 1, it is impossible to select a finite subset of such intervals which still cover all the reals.

4.1.3 Connectedness and Connectivity

A topological space X is *connected* if it is not the disjoint union of two or more nonempty open sets. Otherwise the space is disconnected. Connected spaces have a very interesting property: the only sets with empty boundaries are the total space and the empty set. If any closed continuous curve mapping the unit circle $\gamma : [0, 1] \times S^1 \to X$ into a topological space X can be continuously deformed to a point, then the space is called *simply connected* or 1-connected. Otherwise, the space is multiply connected. The algebraic properties of the sets describing multiple connectivity of a space in terms of classifying curves (or generalized curves and surfaces embedded in the topological space) that are contractible to a point are the subject of *algebraic topology* and *homotopy theory*.

Connectivity and connectedness are important for practical applications, such as the numerical geometrical analysis of patterns. In the past decade, radar systems

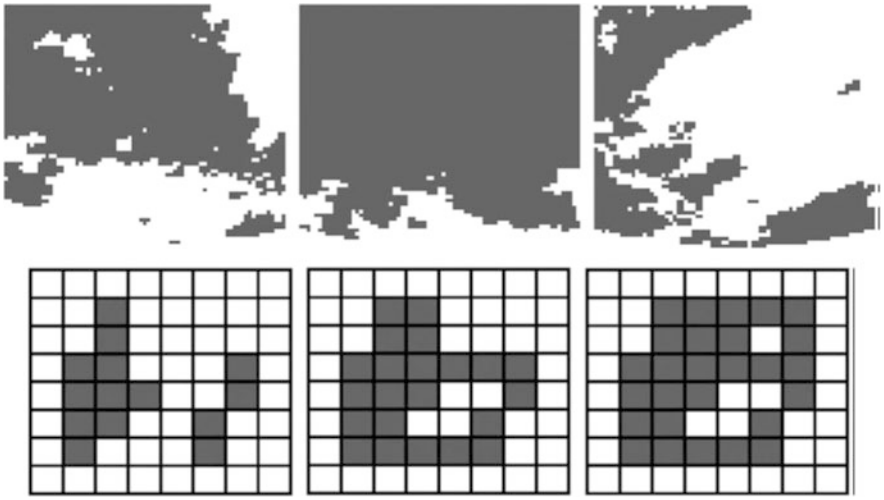


Fig. 4.4 Application of topology to weather prediction. *Upper*: Spatial patterns in precipitation fields and the use of topological connectivity (*left to right* Euler characteristics $-14, -1, -1$). *Lower*: Details of connectivity in precipitation fields, describing the occurrence of clusters and isolated structures (*left to right* Euler characteristics $2, 0, -1$). Courtesy of the Journal of Hydrometeorology [150], image obtained from Memphis Next Generation Weather Radar station

and satellites have provided detailed information on spatial patterns of precipitation, e.g., for the prediction of rainfall fields, and the methods of digital topology give fairly accurate results. Figure 4.4 shows images of digitized rainfall patterns which can be classified according to the number of connected components n and the degree of multiple connectivity (i.e., the number of holes h), through the *Euler characteristic* for the plane, which is calculated as $E = n - h$. In hydrometeorology studies that use digital topology, it may happen that very different images have the same Euler characteristics (see, for example, the last two frames in the upper row of Fig. 4.4). Hydrometeorologists prefer to use another figure of merit describing a 2D rainfall distribution digitized in $\{0, 1\}$, namely the *connectivity index*

$$C_{\text{index}} = 1 - \frac{nh - 1}{\sqrt{A} + nh} \ ,$$

where A is the percentage of dark area in the image. Figure 4.5 presents a series of abstract distributions of 0s and 1s with different topological characteristics, mentioning the Euler characteristic and the C_{index} under each case in order to understand to what extent one or the another description can be useful in classifying connectivity and connectedness.

Homeomorphisms between sets and spaces conserve openess/closure, separation, compactness, and connectedness properties.

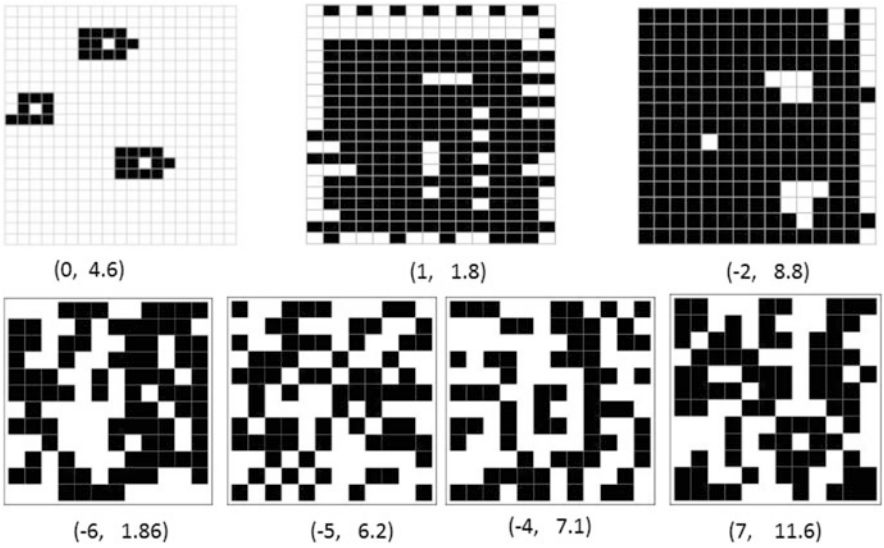


Fig. 4.5 Digitized abstract patterns with different topological aspects, described in brackets under each frame by (E, C_{index}). In the *first row*, the three figures have completely different structures, and both numbers are indeed different. In the *lower row*, while the Euler number is almost the same for the first three frames, the connectivity index makes a better separation of pattern structures

4.2 Topological Boundary

In a topology (X, τ), the *topological boundary* $\mathrm{Fr}(M)$ of a set $M \subset X$ is the set of points in the closure of M, but not belonging to its interior:

$$\mathrm{Fr}(M) = \overline{M} - \overset{\circ}{M} , \quad \text{or } \overline{M} = M \cup \mathrm{Fr}(M) . \tag{4.2}$$

An element of the topological boundary is called a boundary point of M. The topological boundary of a set is closed. Furthermore, x is a boundary point of a set if and only if every neighborhood of x contains at least one point in the set and at least one point not in the set. A set is closed if and only if it contains its boundary, and open if and only if it is disjoint from its topological boundary. One of the most peculiar properties of the topological boundary is expressed by the theorem:

Theorem 2 *The topological boundary of a topological boundary of M is included in (equal to) the original topological boundary:*

$$FrFr(M) \subseteq Fr(M) , \quad FrFr(M) = FrFrFr(M) = \cdots = Fr \ldots Fr(M) .$$

In other words, the topological boundary operator is weakly idempotent.

In the following we will observe a big difference between the topological boundary $\mathrm{Fr}(M)$ and the manifold boundary ∂M, to be defined later. While the former obeys Theorem 2, the latter has an immediate nilpotent property $\partial\partial M = \emptyset$. The reason for this major difference can be found in the definition (4.2), because once a set no longer has an interior, its chain of topological boundaries generates the same set: *the topological boundary of a set without interior is the set itself.*

For example, the intersection between an open set $A \subset (X, \tau)$ and a closed set $B \subset X$ is open in the topology (B, τ'). In the plane, the intersection between an open disk and a curve is an open set in the line topology. The topological boundary of the disk D_2 is the circle $S_1 = \mathrm{Fr}(D_2)$ (its circumference), and as a topological space, the circle is both closed and open. The topological boundary of the circle $\mathrm{FrFr}(D_2) = \mathrm{Fr}(S_1) \sim \mathbb{Q}$ is not the empty set. It is a countable set of points dense in the circle, just as the rational numbers are dense in the real axis. The double topological boundary, the circle, has no interior, and according to the definition in (4.2), no matter how many times we apply the topological boundary operator, the result will be the same.

4.3 Manifold Boundary

In order to study *manifold boundaries*, we follow the definitions and notations from the books by Kosinski [151] and Milnor [152]. We define an n-dimensional *differential manifold of class* C^k *with boundary* to be a set M with three properties:

1. It is a Hausdorff separated topological space.
2. It has a countable basis of open sets, i.e., it is a *second countable* topological space.
3. It has a C^k structure defined on it, as follows:

The C^k structure is a maximal C^k *atlas* on M, that is a maximal (against set union and inclusion) set of *charts* $\{(U_\alpha, h_\alpha)\}$, where $h_\alpha : U_\alpha \to \mathbb{R}^n$ are homeomorphisms defined each on a set from the open covering $\{U_\alpha\}$ of M onto open sets in $\mathbb{R}^n$ or $\mathbb{R}^n_+$, where $\mathbb{R}^n_+ = \{(x_1, \ldots, x_n) | x_n \geq 0\}$ is a real n-dimensional half-space. In addition, the *transition maps* $h_\beta h_\alpha^{-1}$ are C^k on $h_\alpha(U_\alpha \cap U_\beta)$. If the transition maps are defined on $\mathbb{R}^n_+$, the differentiability property is understood in the sense of a local restriction of a differential map on $\mathbb{R}^n$.

In simple terms, an n-dimensional differential manifold with boundary is a 'well-behaved' topological space which is locally homeomorphic to some finite-dimensional real space or half space, such that we can assign local real coordinates, and all coordinate charts in the covering atlas are compatible through differentiable maps. The differential structure so defined on M is borrowed from $\mathbb{R}^n$ and $\mathbb{R}^n_+$. If $k \to \infty$, the differential structure is said to be *smooth*.

There is a minimal type of reasoning for why a smooth manifold (with or without boundary) should be defined as above. Here we try to list the important topological properties associated with a smooth n-dimensional manifold:

1. The $\mathbb{R}^n$ local charts endow M with the property of *local compactness*. The model $\mathbb{R}^n$ is locally compact by construction, and this induces the property on any neighborhood of M. Local compactness means that any point of M has a compact neighborhood, or that M behaves locally like a compact space. In general, a topological space is locally compact if it is Hausdorff and compact. This property is a good enough balance between local separation properties and global compactification properties of a space. Simpler situations include the case where the space is only compact (and then any closed set is compact), or if the space is only Hausdorff (and then any compact set is closed). Local compactness is very useful. It allows compactification of the space with just one point, as in the case of the stereographic projection. This one-point compactification property allows the construction of the Riesz representation theorem, which lies at the heart of functional analysis and dual space theory. It also allows the possibility of integrating functions on groups, just as the definition of the Lebesgue integral on $\mathbb{R}$ helps us to construct harmonic analysis.
2. M is a second-countable space. In other words M has a countable basis of open sets. It follows that M is separable and any set dense in M is countable.
3. M is *paracompact*. This follows automatically from the properties of Hausdorffness and second countability [151]. It means that every open cover has an open refinement that is locally finite. The most important feature of paracompact Hausdorff spaces is that they admit differentiable partitions of unity subordinate to any open cover. Such spaces can be separated not only in the Hausdorff sense, but also by continuous functions. Partitions of unity are useful because they often allow one to extend local constructions to the whole space. For instance, the integral of differential forms on paracompact manifolds is first defined locally (where the manifold looks like Euclidean space and the integral is well known), and this definition is then extended to the whole space via a partition of unity. There are other consequences of paracompactness, e.g., it allows a Riemannian metric structure, and every vector bundle on M is isomorphic to its dual bundle.
4. M is a normal topological space. The main significance of normal spaces lies in the fact that they admit 'enough' continuous real-valued functions: for any two disjoint closed subsets, there exists a continuous real function which is zero on one of the sets and one on the other set. Disjoint closed sets are not only separated by neighborhoods, but also separated by functions.
5. The Stone–Weierstrass theorem applies to any closed subset of M. Any continuous function $f : M \to \mathbb{R}^n$, smooth on a closed subset K, can be arbitrarily well approximated with smooth functions $g : M \to \mathbb{R}^n$, which agree with f on K, $g|_K \equiv f|_K$.
6. If M is connected, any 2 of its points can be joined by a smooth curve. By connectedness, any two points can be joined by piecewise smooth curves that can be completely smoothed out through item 4 above.

If the determinant of the Jacobian of the transition maps is positive everywhere, then we have an *oriented differential structure* on M and the atlas is oriented. An oriented manifold with boundary also has an oriented boundary.

For two differential manifolds M, N, the map $f : M \to N$ is smooth if there are atlases (U_α, h_α) on M and (V_β, g_β) on N, such that $g_\beta f h_\alpha$ are smooth for all α, β wherever they are defined, and f is a *diffeomorphism* if it is smooth and has an inverse. Similarly, with the concept of oriented manifolds, the map f is *orientation preserving* if the determinant of the Jacobian of any of the maps $g_\beta f h_\alpha$ is positive everywhere.

We define the *manifold boundary* ∂M of M to be the closed subset of M made up of points belonging to charts modeled on $\mathbb{R}^n_+$. Obviously, all the intersections $U_\alpha \cap \partial M$ and the restrictions of h_α on these intersections form a maximal C^k atlas on ∂M. The boundary of an n-dimensional manifold inherits from M the structure of an $(n-1)$-dimensional differential manifold without boundary. If $\partial M = \emptyset$, we say that M is *closed*.

4.4 Forms and the Lie Derivative

A collection of intersecting differentiable curves at any point of a smooth manifold X defines a linear space. Here, we use the same convention of calling class C^∞ objects smooth. Let $\gamma : I \to X$ be a smooth curve from an open $I \subset \mathbb{R}$ on the n-dimensional manifold X. In local coordinates, the curve is parameterized by n smooth functions $\gamma(t) = (x^1(t), \ldots, x^n(t))$. At each point $x = \gamma(t)$, the curve has an n-dimensional unit tangent vector defined by the derivative $\gamma'(t)$, viz.,

$$\boldsymbol{v} = \gamma'(t) = \sum_{i=1}^{n} \frac{\mathrm{d}x^i}{\mathrm{d}t} \frac{\partial}{\partial x^i} ,$$

in local coordinates. We use the symbols $\partial/\partial x^i$ to represent the local basis for the components of this tangent vector at x [153]. The collection of all tangent vectors to all possible parameterized curves passing through a given point $x \in X$ is the *tangent space* to X at x, denoted by $T_x M$. The collection of all tangent spaces corresponding to all points of X is the *tangent bundle*, and it is denoted by

$$TX = \bigcup_{x \in X} T_x X .$$

A differentiable function $\boldsymbol{v} : X \to T_x X$ is a smooth *vector field* on the smooth manifold X. A vector field is nothing but a differential operator acting on functions on the manifold. In the geometry of surfaces, this operator is called the *directional derivative* [137]. We can show the action of a vector field $\boldsymbol{v} = (\xi^i)$ on a

differentiable function $f : X \to \mathbb{R}^m$ by

$$\boldsymbol{v}(f)(x) = \xi^i(x)\frac{\partial f(x)}{\partial x^i} \ .$$

This formula shows the action in local coordinates. Geometers prefer to use coordinate-free formulas. In this case, it would read

$$D_{\boldsymbol{v}} f(x) = \boldsymbol{v} \cdot \nabla f(x) \ .$$

The map between two manifolds would create a similar action at the level of their tangent spaces. For example, a homeomorphism between two open sets would generate an isomorphism at the tangent space level. We can express this in a more rigorous language by saying that, for any smooth map $f : X \to Y$, there is a linear map

$$\mathrm{d}f : T_x X \to T_{f(x)} Y \ , \tag{4.3}$$

called the *tangent map* (or the *differential map*), defined in terms of its action on tangent vectors $\boldsymbol{v} = (\xi^i) \in T_x X$ in the local coordinates:

$$\mathrm{d}f(\boldsymbol{v}) = \xi^i \frac{\partial f^j}{\partial x^i}\frac{\partial}{\partial y^j} \in T_{y=f(x)} Y \ , \tag{4.4}$$

where (x^i) and (y^j) are local coordinates in X and Y, respectively. In this context, the tangent map is the Jacobian matrix $J(f)$ of the map f at x, acting as a linear transformation on the tangent vectors. Any canonical local basis in $T_x X$ is mapped by $\mathrm{d}f$ into the basis $\{\partial f/\partial x^1, \ldots, \partial f/\partial x^n\}$ in $T_{f(x)} Y$. The relation between the tangent map of a map f, the action of a vector field $\boldsymbol{v}$, and its directional derivative can be expressed by

$$\mathrm{d}f(\boldsymbol{v}) = D_{\boldsymbol{v}} f(x) = \boldsymbol{v}(f) \ . \tag{4.5}$$

At each point of the manifold $x \in X$, we can choose an orientation of the basis of vectors in $T_x X$, that is, a mapping $\mathrm{d}f : T_x X \to \mathbb{R}^n$ which locally (in a chart) preserves the orientation and maps it into the standard orientation of the canonical basis in $\mathbb{R}^n$. If we do this for any point x, the manifold is an *oriented manifold*.

If X is smooth, orientable, and connected, it has precisely two orientations. Moreover, an orientation in X can induce an orientation in its boundary ∂X (if it has a boundary) as follows: at any point on the boundary, we choose a positively oriented basis $\{\boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ for $T_x X$ in such a way that $\{\boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ are tangent to the boundary and $\boldsymbol{v}_1$ is an outward vector. Then $\{\boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ determines the required orientation for ∂X (see Fig. 4.6).

Now consider two oriented n-dimensional manifolds, X compact and Y connected, both without boundary, and a smooth map $f : M \to N$. Let $x \in X$ be a regular point of f, that is, $\mathrm{d}f_x : T_x X \to T_{f(x)} Y$ is a linear isomorphism between the

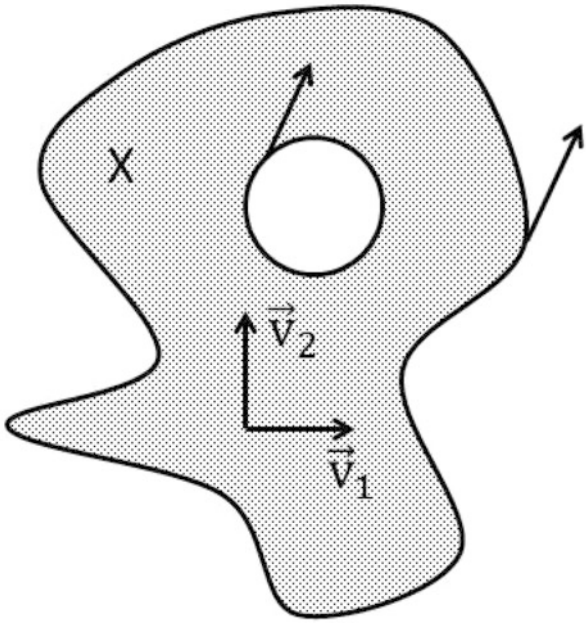


Fig. 4.6 How to orient a boundary

two oriented tangent vector spaces. Define the sign of $\mathrm{d}f_x$ to be $+1$ or -1 according as $\mathrm{d}f_x$ preserves or reverses the orientation. For any regular value $y \in Y$, we define

$$\deg(f; y) = \sum_{x=f^{-1}(y)} \operatorname{sign} \mathrm{d}f_x \,. \tag{4.6}$$

Since X is orientable, the integer $\deg(f; y)$ is a locally constant function of y. Moreover, it can be shown that it does not depend on the specific choice of y and it represents only the properties of f. Consequently, it is called the *degree* of f.

At any $x \in X$, we can define the dual of the tangent space, the cotangent space T_x^*X. The space of skew-symmetric covariant tensors of rank 1 on X is a linear subspace $\Omega^1 T_x^*X \subset T_x^*X$ of the cotangent space. Its elements $\omega(x)$ are called 1-forms. In local coordinates (x^j), the 1-form is denoted by $\omega = \omega_i \mathrm{d}x^i$, where the $\mathrm{d}x^i$ form an abstract skew-symmetric local basis for the cotangent space. The 1-form is precisely defined by its action on differential vector fields:

$$(\omega; \boldsymbol{v}) = \left(\omega_j \mathrm{d}x^j; \xi^i \frac{\partial}{\partial x^i}\right) = \sum_{i=1}^{n} \omega_i \xi^i \,.$$

This definition can be generalized to *differentiable k-forms*, that is, skew-symmetric covariant tensor fields of rank $0 \leqslant k \leqslant n$ defined on X:

$$\omega = \omega_{i_1 i_2 \ldots i_k} \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \ldots \wedge \mathrm{d}x^k \,.$$

Here the wedge symbol $\wedge$ represents the skew-symmetric exterior product. A 0-form is a differentiable function on X, a 1-form is a covariant vector field, and a maximal dimension n-form is the volume element

$$\omega = \omega_{i_1,\ldots,i_n} \mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_n} = \omega(x) \mathcal{E}_{i_1,\ldots,i_n} \mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_n} \in \Omega^n T_x^* \ ,$$

where $\mathcal{E}_{i_1,\ldots,i_n}$ is the Levi-Civita permutation symbol.

One of the main reasons the cotangent bundle rather than the tangent bundle is used in the constructions of differential geometry (like the Lie derivative) is that differential forms can be *pulled back* by smooth maps, while vector fields cannot be *pushed forward* by smooth maps unless the map is a diffeomorphism. So differential forms can be moved from one manifold to another using a smooth map. If we have a smooth map between two manifolds $f : X \to Y$, we can capture in X the behavior of the forms on Y relative to f. For an arbitrary smooth map f, it is not usually possible to push forward a vector field on X using f to obtain a vector field on Y, unless the map is a diffeomorphism. If the map is not surjective, there is no natural way to define such a push-forward outside the image of f, and if the map is not injective, there is no unique choice to define a push-forward at a given point.

We define the dual of the tangent map at x, denoted $\mathrm{d}f^* : T_{f(x)}Y \to T_x X$, and called it the *pull-back* (or codifferential) of f. The action of the pull-back on k-forms is given by

$$\Phi^* : \Omega^k T^*_{f(x)} Y \to \Omega^k T^*_x X \ . \tag{4.7}$$

The pull-back is related to the tangent map between X and Y by the following expression:

$$\big(\omega; \mathrm{d}f(\boldsymbol{v})\big) = \big(f^*(\omega); \boldsymbol{v}\big) \ , \tag{4.8}$$

meaning that k-forms in Y act on the derivative $\mathrm{d}f(\boldsymbol{v})$ of the vector field $\boldsymbol{v}$ on X in the same way as the pull-back $f^*(\omega)$ of the form in X acts on vector fields $\boldsymbol{v}$ on X.

The generalization of a function, or of an infinitesimal surface or volume element, is the differential k-form defined as a differential skew-symmetric covariant tensor field on X, with entries in the k-times exterior product of the cotangent space of the n-dimensional manifold X [131].

A differential form can couple many vectors together. This is the geometric way to construct area elements from arc length, volume elements from infinitesimal areas, and so on. For a given set of k vector fields on X, we have the action of the k-form on these fields given by

$$(\omega; \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) = \omega_{i_1 i_2 \ldots i_k} \boldsymbol{v}_1^{i_1} \ldots \boldsymbol{v}_k^{i_k} \ . \tag{4.9}$$

A k-form ω and an r-form θ can be combined into a new $(k+l)$-form by the exterior product, through the $\wedge$ operation. For example, if $\omega, \theta \in \Omega^1 = T^*X$ are 1-forms on

the n-dimensional differentiable manifold X, we have

$$\omega \wedge \theta = (\omega_i \theta_j - \omega_j \theta_i) \mathrm{d}x^i \wedge \mathrm{d}x^j \ , \quad i,j = 1,2,\ldots,n \ .$$

Here is another example, for ω a 2-form and θ a 1-form:

$$\omega \wedge \theta = (\omega_{12}\theta_3 - \omega_{13}\theta_2 + \omega_{23}\theta_1) \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \ .$$

We can also define the interior product between a vector field and a k-form ω, this operation yielding a $(k-1)$-form. For example, $\partial_x \perp \mathrm{d}x \wedge \mathrm{d}y = \mathrm{d}y$. The action of the interior product is given by

$$(\boldsymbol{v} \perp \omega; \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}) = (\omega; \boldsymbol{v}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}) \ . \tag{4.10}$$

The last operator we need for our purposes is the *exterior derivative*.

This is perhaps the most intriguing property of differentiable forms. Here we differentiate a differential form and obtain a higher order form, but the key point is that we can do this only once! Every time we differentiate a form twice in this way, we obtain zero. This is a consequence of the algebra: the skew-symmetry of the form combines with the symmetry of the derivative operator in a destructive way. More precisely, for a k-form ω, we define the linear external derivative operator $\mathrm{d} : \Omega^k T_x^* X \to \Omega^{k+1} T_x^* X$ acting on ω to produce a $(k+1)$-form:

$$\mathrm{d}\omega = \sum_{i,I} \frac{\partial \omega_I}{\partial x^i} \mathrm{d}x^i \wedge \mathrm{d}x^I \ ,$$

where I is the increasing ordered multilabel specifying the components of ω.

The exterior derivative is linear, commutes with the pull-back map, and most importantly, has the closure property

$$\mathrm{d}(\mathrm{d}\omega) = \mathrm{d}^2\omega = 0 \ . \tag{4.11}$$

For a given vector field $\boldsymbol{v}(x)$ on X and a certain geometrical object $\omega(x)$ defined on X (like another vector field or a k-form), it is natural to ask how ω changes along the integral curves of $\boldsymbol{v}$. Since at different points $\mathrm{e}^{\lambda \boldsymbol{v}} x$ the quantity $\omega(x)$ takes values in different spaces of a fiber bundle ΩX over X (e.g., the tangent bundle TX, cotangent bundle T^*X, tensor bundle $T_k^j X$, etc.), we have to compare the values of $\omega(x) \in \Omega_x X$ with the pulled-back values of $\omega(\mathrm{e}^{\lambda \boldsymbol{v}} x) \in \Omega_{\mathrm{e}^{\lambda v} x}$. This technique leads to the *Lie derivative*.

Let $\mathcal{T}$ be a smooth tensor field and $\boldsymbol{v}$ a smooth vector field (i.e., a smooth section of the tangent bundle TM). Then we can define the Lie derivative of $\mathcal{T}$ along $\boldsymbol{v}$ as follows. Let $\phi : X \times \mathbb{R} \to X$ be the one-parameter semigroup of local diffeomorphisms of X induced by the vector flow of $\boldsymbol{v}$. For each sufficiently small real $|t| \ll 1$, $\phi(x,t)$ is a diffeomorphism from a neighborhood in X to another

neighborhood in X, and $\phi(x,0)$ is the identity diffeomorphism. The *Lie derivative* of $\mathcal{T}$ is defined at a point x by

$$\boldsymbol{v}(\mathcal{T})_x = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \left[\phi(x,-t)^*(\mathcal{T}_{\phi(x,t)})\right] ,$$

where $\phi(x,-t)^*$ is the pull-back of ϕ between the corresponding tensor spaces, i.e., at $\phi(x,t)$ and at x. There are two rigorous definitions of a Lie derivative, based on the concepts of affine connection and geodesic, or on Lie groups and Lie algebras, respectively, that can be extended to all the tangent and cotangent bundles over X. However, the definition above, although it is local, serves the goals of the following formulas very well.

The Lie derivative of a function is

$$\boldsymbol{v}(f)(x) = \xi^i \frac{\partial f}{\partial x^i} .$$

The Lie derivative of a vector field $\boldsymbol{w}$ is

$$\boldsymbol{v}(\boldsymbol{w}) = [\boldsymbol{v},\boldsymbol{w}] = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(v^j\frac{\partial w^i}{\partial x^j} - w^j\frac{\partial v^i}{\partial x^j}\right)\frac{\partial}{\partial x^i} . \tag{4.12}$$

The Lie derivative of a k-form $\omega = \omega_{i_1,i_2,\ldots,i_k}\mathrm{d}x^{i_1}\wedge\mathrm{d}x^{i_2}\wedge\cdots\wedge\mathrm{d}x^{i_k} \in \Omega_k X$ is

$$\begin{aligned}\boldsymbol{v}(\omega) &= \boldsymbol{v}(\omega_{i_1,i_2,\ldots,i_k})\mathrm{d}x^{i_1}\wedge\mathrm{d}x^{i_2}\wedge\cdots\wedge\mathrm{d}x^{i_k} \\ &\quad + \sum_{j=1}^{k}\omega_{i_1,i_2,\ldots,i_j,\ldots,i_k}\mathrm{d}x^{i_1}\wedge\cdots\wedge\boldsymbol{v}(\mathrm{d}x^{i_j})\wedge\ldots\mathrm{d}x^{i_k} ,\end{aligned} \tag{4.13}$$

where we can use the formula $\boldsymbol{v}(\mathrm{d}x^{i_j}) = \mathrm{d}v^{i_j} = (\partial v^{i_j}/\partial x^k)\mathrm{d}x^k$. For example, if $k=2$, we find the Lie derivative of a 1-form by knowing its action on vector fields, viz.,

$$(\boldsymbol{v}(\omega);\boldsymbol{w}) = \boldsymbol{v}(\omega;\boldsymbol{w}) - (\omega;[\boldsymbol{v},\boldsymbol{w}]) . \tag{4.14}$$

In components, given a vector field

$$\boldsymbol{v}(x,y) = \xi(x,y)\frac{\partial}{\partial x} + \eta(x,y)\frac{\partial}{\partial y} ,$$

the Lie derivative of a 2-form $\omega = \omega_{12}\mathrm{d}x\wedge\mathrm{d}y$ is

$$\boldsymbol{v}(\omega) = \left[\xi\frac{\partial\omega_{12}}{\partial x} + \eta\frac{\partial\omega_{12}}{\partial y} + \omega_{12}\left(\frac{\partial\xi}{\partial x} + \frac{\partial\eta}{\partial y}\right)\right]\mathrm{d}x\wedge\mathrm{d}y .$$

The following relation is called Cartan's formula. Let $\boldsymbol{v}, \omega$ be a smooth vector field and form, respectively. Then,

$$\boldsymbol{v}(\omega) = \boldsymbol{v} \perp \mathrm{d}\omega + \mathrm{d}(\boldsymbol{v}; \omega) \; . \tag{4.15}$$

As a corollary, we have

$$\mathrm{d}\boldsymbol{v}(\omega) = \boldsymbol{v}(\mathrm{d}\omega) \; .$$

One important geometric consequence of the properties of differentiable forms is the Maurer–Cartan equation

$$\mathrm{d}\omega + \frac{1}{2}[\omega, \omega] = 0 \; , \tag{4.16}$$

where the bracket is the Lie bracket of the Lie algebra. In the second case, the form $\mathrm{d}\omega + \omega \wedge \omega$ represents the curvature 2-form of a linear connection, also called the first Cartan structure equation.

A space is contractible if there is a deformation homeomorphism that contracts m to one of its points. A *closed form* ω is a differential form with the property that $\mathrm{d}\omega = 0$. An exact p-form $\phi \in \Omega^p$ has the property that there exists a $(p-1)$-form $\psi \in \Omega^{p-1}$ such that $\mathrm{d}\psi = \phi$. With these definitions, we have the famous Poincaré lemma: if a manifold M is contractible to a point, then all closed forms on M are exact. This lemma is a generalization of the fact that, on simply connected domains, a total exact differential is integrable and its path integral does not depend on the path. For example, in $\mathbb{R}^3$, the Poincaré lemma is the underlying cause for the following important vector analysis relations:

$$\begin{gathered}
\nabla \times \nabla \Phi = 0 \; , \qquad \nabla \cdot (\nabla \times \boldsymbol{V}) = 0 \; , \\
\nabla \times \boldsymbol{V} = 0 \implies \exists \, \Phi \;\; \text{such that} \;\; \boldsymbol{V} = \nabla \Phi \; , \\
\nabla \cdot \boldsymbol{V} = 0 \implies \exists \, \boldsymbol{W} \;\; \text{such that} \;\; \boldsymbol{V} = \nabla \times \boldsymbol{W} \; .
\end{gathered}$$

Next, we present a property of forms which becomes useful in the hydrodynamics of incompressible fluids. Consider an n-dimensional smooth manifold M, and let $\omega \in \Omega^n M$ be a *volume form*. In particular, if $M = \mathbb{R}^n$, we have $\omega = \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \ldots \wedge \mathrm{d}x^n$. It can be easily shown that the Lie derivative of the volume form is

$$\boldsymbol{v}(\omega) = \sum_{k=1}^{n} \left(\frac{\partial v^k}{\partial x^k} \right) \mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n = (\mathrm{div}\, \boldsymbol{v}) \mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n \; .$$

It follows that the Lie derivative along $\boldsymbol{v}$ of a volume form is the divergence of the vector field times the volume form:

$$\boldsymbol{v}(\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n) = (\mathrm{div}\, \boldsymbol{v}) \mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n \; . \tag{4.17}$$

It is natural now to remark that, if we want a smooth map $f : M \to M$ to preserve the volume form $\omega \in \Lambda^n(T^*M)$, that is $f^*\omega = \omega$, this will be true if and only if $\det J(f) = Df = 1$, that is, if and only if the Jacobian of f is unity. Indeed,

$$(f^*\omega; \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) = \big(\omega; \mathrm{d}f(\boldsymbol{v}_1), \ldots, \mathrm{d}f(\boldsymbol{v}_n)\big) = \det J(f) \ .$$

4.5 Fiber Bundles and Covariant Derivative

In the mathematical literature, the motivation for introducing fiber bundles consists usually in mentioning the existence of manifolds that are only locally a Cartesian product, not globally, like the Möbius band or the Klein bottle. In this section, we motivate with some examples chosen from everyday life.

Consider a driver who lives in small rural community A in continental Europe and drives his car only in his town, where there are only narrow one-way roads, so that he is not aware of the existence of two-lane roads or of driving on a specific side of the road. Then one day, this person decides to visit the UK. He puts his car on a ferry and has it delivered directly to a British harbor B where his friend lives, a place which also happens to have a one-way road. But the visitor from the initial town doesn't realize this. A conflict would occur only if he were to drive along a two-lane road. Then he would see that drivers steer onto a specific side of the road. We label the drivers on the right side in a chart containing A, and on the left side on a chart containing B. In order to drive correctly from A to B, he has to adapt his charts, so now he is driving along a non-trivial fiber bundle. His town and his friend's town trivialize the fiber bundle. The moral is that as long as he moves locally, he is in a Cartesian product conserving orientation, but when he travels, the global orientation of the charts is no longer preserved. What made the roads become a fiber bundle was the introduction of an extra dimension, the width of the road. Or rather the edge, the boundary of the road, for a set has no center unless it has boundaries.

A *fiber bundle* $E(X, \pi, F, G)$ is topological space E that can be projected onto another topological *base space* X by a *canonical projection* $\pi : E \to X$. For any base point $x \in X$, the sets $\pi^{-1}(x) + E_x \sim F$ are all homeomorphic and the representative of their equivalence class is called a *standard fiber* F. There is an open covering U_α of the base space and a family of coordinate functions $\Phi_\alpha : U_\alpha \to E$ such that the inverse image $\pi^{-1}(U_\alpha)$ is homeomorphic to $U_\alpha \times F$ and $\pi \circ \Phi_\alpha = \mathrm{Id}\, X$ [138, 140].

The way the coordinates are assigned to a fiber F at a point $x \in X$ is handled by the structure group of homeomorphisms of F. The maps $U_\alpha \times F$ are glued together (where the coordinate neighborhoods overlap) in different ways across X, and the structure group G controls the gluing operations between local parts of the total space of the fiber bundle. A cross-section in a bundle is a differentiable injective map $\phi : X \to E$ such that $\pi\phi = \mathrm{Id}\, X$. A cross-section is in a way a generalization of the graph of a function defined on the base space with values in the fiber bundle. A cross-section is a geometric object similar to the graph of a real function $f : \mathbf{R} \to \mathbf{R}$, that is, $G = \{(x, f(x))\}$, except for one big difference. While in the real graph case

Table 4.1 Typical examples of fiber bundles

E	X	F	G	Application
Möbius band	S^1	$[0, 1]$	$O(2, \mathbb{R})$	
Tangent bundle	Σ surface	$T_0\Sigma$	$GL(n, \mathbb{R})$	$\phi(x)$ are vector fields
Frame bundle	A Riemannian vector bundle E	Orbits of G	$GL(n, \mathbb{R})$	Ordered bases
Orthonormal frame bundle	A vector bundle E	$O(n, \mathbb{R})$	$GL(n, \mathbb{R})$	Orthonormal bases

the points $y = f(x) \in \mathbf{R}$ belong to the same space $\mathbf{R}$, for all x, in the case of a cross-section Φ, the second coordinate of the pair $\{(x, \Phi(x))\}$, namely $\Phi(x)$, takes values in different spaces for different values of $x \in X$. Indeed, $\Phi(x)$ belongs consecutively to different homeomorphisms of the standard fiber.

Let us clarify this observation with a simple example. Think about the motion of a football during a long ball down the field. At different moments of time, the ball can be watched by other spectators placed at appropriate points along its path. The record of the football's motion through one TV video camera would be a real graph, but a description of this motion through the opinions of all the individual spectators in the tribunes would require a cross-section in the stadium fiber bundle (where the standard fiber would be a row of seats, for example). Standard examples of fiber bundles are presented in Table 4.1.

Many differential geometry objects originate directly from the theory of Lie groups and algebras. Let $\mathfrak{g}$ be an n-dimensional Lie algebra associated with the Lie group G, and $\boldsymbol{A}, \boldsymbol{B}, \ldots \in \mathfrak{g}$. A Lie group is simultaneously a group and a differential manifold, with the property that the group operations are differentiable in the manifold structure. The Lie algebra is nothing but the tangent bundle to the manifold of the Lie group. A function defined on a Lie group is said to be *left invariant* if it commutes with the left group translations, or with their adjoint representation. A Lie group G can act on other manifolds than itself, and induce orbits. However, the Lie algebra $\mathfrak{g}$ is 'local', i.e., it cannot act at different points on a manifold, as G does, except on G itself. In order to repair this frozen action, we need to enrich the structure of the manifold, and make it into a fiber bundle. In a fiber bundle there is more 'freedom', and we will introduce vertical and horizontal displacements by using the covariant derivative and the connection form, respectively.

Consider a base manifold X and a Lie group G acting upon it. For any element $\boldsymbol{A} \in \mathfrak{g}$, we can construct a *fundamental vector field* $\boldsymbol{A}^* : X \to TX$, defined by

$$x_0 \in X \longrightarrow \boldsymbol{A}^* = \frac{\mathrm{d}(\mathrm{e}^{t\boldsymbol{A}} x_0)}{\mathrm{d}t} \in T_{x_0} X \,.$$

The vector field is thus tangent to the one-parameter Lie subgroups generated by $\boldsymbol{A}$ and the fundamental vector field is tangent to each fiber at each point of P.

The vector–frame duality can be understood in the best way by using a 1-form called the *canonical form* $\theta \in \Omega^1(\mathrm{F}X)$ on the bundle of frames FX with values in the standard fibre F. The action of the canonical form on a vector $\boldsymbol{X} \in T\mathrm{F}X$ is $(\theta; \boldsymbol{X}) = u^{-1} \circ \mathrm{d}\pi(\boldsymbol{X}) \in F$.

If X is an n-dimensional affine space, then a point $x \in X$ is represented by a position vector $\boldsymbol{r} = x^i \boldsymbol{e}_i$, whose components are given in a certain frame $\{\boldsymbol{e}_i\}_{i=1,\ldots,n} = u \in \pi^{-1}(x) \in \mathrm{F}X$. The question is: how does this position vector change with d$\boldsymbol{r}$ under an infinitesimal motion of the frame? The answer is given by the canonical form, or by

$$\mathrm{d}\boldsymbol{r} = (\theta; \boldsymbol{X}) = (\theta^i, \boldsymbol{X})\boldsymbol{e}_i \,, \tag{4.18}$$

where $\boldsymbol{X} \in T_u\mathrm{F}X$ describes this infinitesimal motion of the frame in the tangent space to the bundle of frames.

The bundle of frames does not provide a recipe for the way frames transform when the base point moves through the base space. In order to provide such a law, we need a further construction, which is the connection on X. A connection should provide the infinitesimal transformation of a point in the vector bundle when we perform an infinitesimal move in the base. Since the infinitesimal transformations are described by vectors in the tangent space, a connection should map a point (to be moved) in the vector bundle to a vector in the tangent bundle to the vector bundle (how this point transforms), a map depending on a vector in the tangent space of the base (the direction of the movement).

A connection Γ in a fiber bundle E is the assignment of a G-invariant subspace $H_p \lhd T_pE$, for any $p \in E$ and depending differentiably on p, called the horizontal subspace. The orthogonal complement of H_p is called the *vertical subspace*, denoted by V_p, and we have

$$T_p = V_p \oplus H_p \,.$$

Any vector $\boldsymbol{V} \in T_pE$ can be uniquely decomposed into two orthogonal components, viz., $\boldsymbol{V} = v\boldsymbol{V} + h\boldsymbol{V}$, each in the corresponding subspace $v\boldsymbol{V} \in V_p$, $h\boldsymbol{V} \in H_p$. A horizontal lift of a vector field on X is the unique horizontal vector field on P such that the differential of the canonical projection $\mathrm{d}\pi : TE \to TX$ maps it to the initial vector field. Any parameterized curve in X, and any point $p \in E$, provide a *lift* of this curve to a unique horizontal curve in E (with horizontal tangent vectors), to which it canonically projects.

The existence of a connection allows us to 'flag' elements of E and watch their evolution according to a certain law imposed by this connection, when we move along some curve in the base space. This law is called *parallel displacement* along a certain curve in the base space. We consider the starting point x_0 of a parameterized curve $\gamma \subset X$, and its local fiber $\pi^{-1}(x_0) \subset E$. Through any point p_0 in this fiber, we can build a unique horizontal lift of γ which canonically maps back onto γ. When we move to a different point on γ, the intersection between the fiber over this new point and the horizontal lift of γ through p_0 is a unique point of this new fiber. Doing

this transport now for various $p_0 \in \pi^{-1}(x_0)$ is like mapping all the points p_0 of a fiber into all points of another fiber following the curve. This mapping is actually a fiber isomorphism, and it is called the *parallel displacement* of the fibers along the curve.

Later, we will need the definition of a *normal bundle*. Let M be a compact and smooth manifold with or without boundary, and $V \subset M$ a compact submanifold whose boundary is contained in the boundary of M in such a way that V meets the boundary of M *transversely*, that is, the tangent spaces to V and to ∂M at the same point generate together the tangent space of the ambient manifold. The normal bundle of $V \subset M$ is a particular kind of vector bundle, complementary to the tangent bundle, and coming from the canonical embedding $i : V \to M$. In other words, the normal bundle for V is constructed by taking the quotient space of the tangent space on M by the tangent space on V, that is $i^*(TM)/TV$. The notation for the normal bundle of $V \subset M$ is

$$i : V \hookrightarrow M \ , \quad \text{or} \ \ \nu(V \hookrightarrow M) \ . \tag{4.19}$$

One important result in differential geometry is that, for each connection Γ, we can associate a $\mathfrak{g}$-valued 1-form on E called the *connection form* ω such that, for each $\boldsymbol{V} \in T_pE$, we have

$$(\omega; \boldsymbol{V}) = \left\{ \boldsymbol{A} \in \mathfrak{g} \mid \boldsymbol{A}^* = \upsilon \boldsymbol{X} \right\} .$$

In other words, a connection form maps a vector field $\boldsymbol{V}$ on E to a Lie algebra vector whose fundamental vector field is exactly the vertical component of $\boldsymbol{V}$. In the language of physics a connection form is a vector field defined on a bundle of frames such that its directional derivatives in any direction provide one-dimensional Lie algebras of symmetry (flows) in the vertical component of those directions.

For a differentiable r-form ϕ on E, we can introduce the *exterior covariant derivative* as an $(r+1)$-form $D\phi$ whose action on vector fields in E is

$$\mathrm{D}\phi = (\mathrm{d}\phi)\mathrm{Pr}_H \ ,$$

where d is the exterior derivative and Pr_H is the projection on the horizontal space of the vector fields. The exterior covariant derivative $\mathrm{d}\omega = \Omega$ of the connection form is called the *curvature form*, and we have the *structure equation*

$$\mathrm{d}\omega = -\frac{1}{2}[\omega, \omega] + \Omega \ , \tag{4.20}$$

acting on any pair of vector fields on E. The proof is immediate and it is based on the vertical/horizontal direct sum properties, and can be found in [138]. A connection is flat if and only if its curvature form is null.

In order to build the *covariant derivative* of a cross-section $\varphi : X \to TX$ in the $\boldsymbol{X} \in TX$ direction, we have to lift this last vector to its horizontal component $\boldsymbol{X}^* \in H \subset T\mathrm{F}X$. Following the projections, we have $\mathrm{F}X \ni u \to x = \pi(u) \to \varphi(x)$, which actually defines a cross-section in FX. We can thus apply the directional derivative $\boldsymbol{X}^*(\varphi(x(u))) = \nabla_{\boldsymbol{X}}\varphi$, and this is the desired covariant derivative. Basically, it is the horizontal component of the directional derivative.

By expressing the connection form in coordinates, we obtain the explicit action of the covariant derivative on the basis of covariant vectors:

$$\nabla_{\partial/\partial x^j} \frac{\partial}{\partial x^i} = \Gamma^k_{ji} \frac{\partial}{\partial x^k} \,. \tag{4.21}$$

Equation (4.21) and the linearity of the covariant derivative yield the coordinate expression for the covariant derivative of a vector field $\boldsymbol{V} = V_i \partial/\partial x^i$ defined on X with respect to the directions of the local frame:

$$\nabla_j V_i = \frac{\partial V_i}{\partial x^j} - \Gamma^k_{ij} V_k \,. \tag{4.22}$$

We illustrate these constructions with an example. Consider a unit radius spherical surface $X = S_2$ embedded in $\mathbb{R}^3$ with coordinates $x^1 = \theta \in [0, \pi]$, $x^2 = \phi \in [0, 2\pi)$. The tangent space is TS_2, generated by the basis vectors $\{\boldsymbol{e}_\theta, \boldsymbol{e}_\phi\}$. The bundle of orthonormal frames OS_2 has coordinates $(\theta, \phi, \hat{R}(\alpha))$, where the last one represents an element of the Lie structure group $O(2, \mathbb{R})$, i.e., a rotation of the tangent frame through an angle α around the normal to the sphere. The covariant derivatives have the form

$$\nabla_{\boldsymbol{e}_\theta} \boldsymbol{e}_\theta = 0 \,, \quad \nabla_{\boldsymbol{e}_\phi} \boldsymbol{e}_\theta = \boldsymbol{e}_\phi \cot\theta \,, \quad \nabla_{\boldsymbol{e}_\phi} \boldsymbol{e}_\phi = \boldsymbol{e}_\theta \sin\theta \cos\theta \,,$$

and the horizontal lift of the basis vectors is

$$\boldsymbol{e}^*_\theta = \boldsymbol{e}_\theta - \boldsymbol{n} \cos\theta \,, \quad \boldsymbol{e}^*_\phi = \boldsymbol{e}_\phi - \boldsymbol{n} \sin\theta \cos\theta \,.$$

We can check by noticing that, at $\theta = \pi/2$, the covariant derivatives cancel, as do the vertical projections, which is correct since this equatorial circle is actually a geodesic and its tangent vectors are parallel transported along it. If we want to find how a tangent vector field is parallel transported [137], we can choose a vector which is $\boldsymbol{e}_\phi$ at an initial point and transport it along a parallel to the sphere at $\theta = \theta_0$, parameterized by $t \in [0, 2\pi)$. The resulting parallel-translated vector is

$$\boldsymbol{V}(t) = \sin(\theta_0) \sin(t \cos\theta_0) \boldsymbol{e}_\theta + \cos(t \cos\theta_0) \boldsymbol{e}_\phi \,, \quad \nabla_{\boldsymbol{e}_\phi} \boldsymbol{V} = 0 \,.$$

In the next section, we draw attention to a few similarities and differences between the Lie derivative and the covariant derivative.

4.6 Is the Lagrangian Derivative a Lie or a Covariant Derivative?

The answer is: both. Consider a volume of 'fluid' material comprising a very large number of small moving particles, 'densely' enough distributed to allow us to model a differentiable manifold structure on the set of their positions. We imagine the fluid enclosed in a rigid boundary container. Let us take a snapshot of these particles showing, not the particles themselves, but only their instantaneous velocity vectors. The picture is taken with a camera placed in the container's system of reference, and it is taken in total darkness so that nothing else shows in it (like walls, surrounding objects, etc.). Using only this picture, it is impossible to predict the future state of motion of the fluid. We cannot predict the evolution of any of the vectors from the snapshot, and we do not know what vector is attached to which particle. This snapshot is (a representation of) the tangent bundle of the fluid manifold.

To find out more about the system's states, we take more such snapshots at different times. Even by building this album of snapshots, we are still unable to compare velocities represented by different vectors in different snapshots. Are two vectors in two snapshots the velocities of the same particle at different moments, or do they represent the velocities of two different particles (see Fig. 4.7)?

In order to try to understand the 'dynamics' of such a fluid, we can carry out two other operations. One procedure (called LALI from Lagrange–Lie) is to collect sets of snapshots taken at different moments of time, but separated by very small

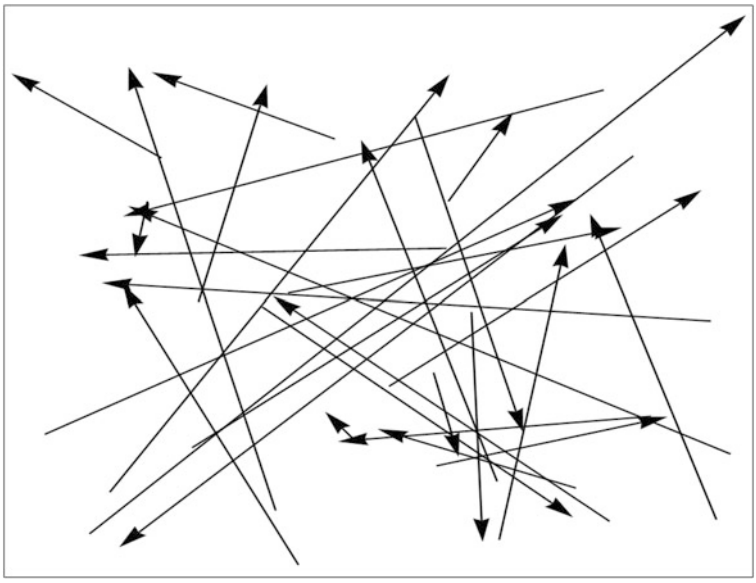


Fig. 4.7 Vectors belonging to different tangent planes to the same manifold without a connection. It is impossible to compare any two vectors, or carry out any operation on them

time intervals, so that we can follow the motion of each vector frame by frame, in other words, such that the shift in the origin of each vector from one snapshot to the next is always much smaller than the average minimum distance between any two particles. In this way, we can associate any point moving in space with a certain curve, i.e., a path of the flow defined to be tangent to all of the vectors belonging to the same particle. Then we can compare vectors along the same (almost) smooth curve. This comparison will represent the change in time of the velocity of one labeled particle, so what we record is the Lagrangian derivative of its velocity. The comparison of vectors in this procedure is then generalized by considering all the paths, over the whole flow, and over the whole manifold. This is known as the Lie derivative approach.

The Lie derivative of the velocity vector in the fluid motion is the story of one particle in the movie of the flow. Actually, such procedures are quite possible experimentally, using so-called *particle imaging velocimetry* (PIV) systems. These seed the fluid with highly reflecting microscopic particles that move everywhere with the fluid elements, then apply two consecutive laser flashes while recording with a rapid-photography camera. A computer software creates a 'particle' for each pair of consecutive vectors with origins close enough, and builds the velocity vector field. We do not need to see any exterior structure (walls, container, lab frame, etc.) in order to know the dynamics of this system. The complete set of path lines represents the fluid, and the most isolated of the paths describe the fluid boundaries (see Fig. 4.8).

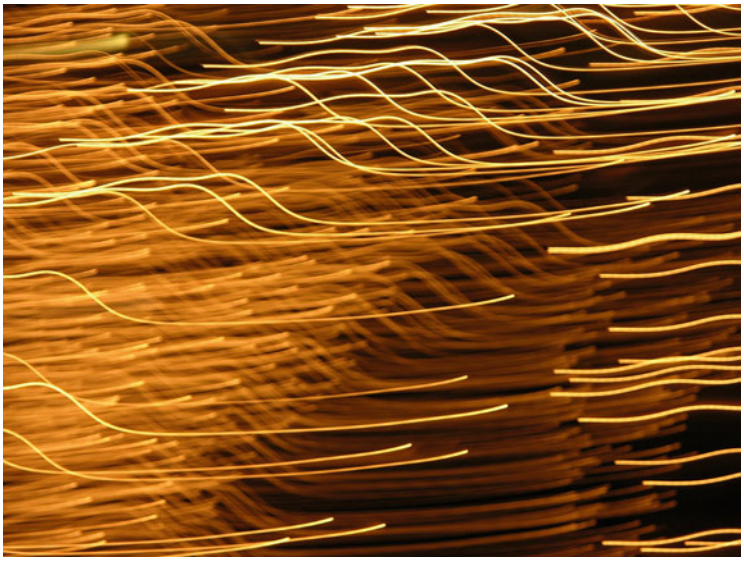


Fig. 4.8 Example of the action a Lie derivative. Path lines in a fluid motion, built by joining vectors with closest origins. We can evaluate the rate of change of a vector associated with one particle by following the vector along one path

There is a second procedure to analyze (and predict) the dynamics of the system, which can be called ECO (from Eulerian–covariant). In order to proceed, the snapshot needs to include elements from the fixed boundaries for the fluid, or elements of the container, or geometrical objects in the environment that do not move with the fluid. We build the geometry of the fixed environmental elements and their laws of transformation from location to location. For example, we collect information about the tank shape, the geometry of the walls, etc. Then we measure the vectors associated with each local subset of fluid particles using the geometry borrowed from the fixed surrounding environment. We measure any particle's vector with respect to its local environment. Knowing how to connect and transform the results of measurements performed at different locations from one to the other, we induce a law of transformation for each vector at any point in the fluid. In other words, we place all vectors in the same system of coordinates by considering their transformations from point to point with respect to the transformation of the environment. Only in this way can we compare any two vectors anywhere in the fluid (see Fig. 4.9 for an example).

This second approach is the covariant derivative or Euler approach. We do not need the flow of the fluid, we do not need the paths, but we need to select one universal system of reference which is called the *reference fiber*, or to know how to refer the vectors at different points to such a model system of reference, which is the procedure called *connection*. The covariant derivative sees different tangent planes at different points, and it can compare them, even if these planes do not represent the same particle at different times in its motion. In principle, both the Lie derivative and the covariant derivative can describe the Lagrangian time derivative [154].

Moving fluids have a very particular property, namely an interesting time dependence of scalars and tensor fields on the observer's point of view, in a

Fig. 4.9 Example of a Cartan connection: we know how strongly the wind blows at different locations because we compare the bending of different stems with an out-of-the-wheat-field fixed frame provided by the sky, the earth, the vertical direction, etc., everywhere

different way than a general global Galilean or Lorentz transformation. When we change from a laboratory frame of reference (Euler) to a moving-together-with-the-fluid observer's point of view (Lagrange), extra nonlinear terms arise in the time variation. These are terms containing the velocity, something which does not happen in the case of Galilean or Lorentz transformations represented by the group of transformations $GL(3,\mathbb{R})$. A detailed discussion can be found in Chaps. 4, 8, and 9 of [121]. This peculiar transformation from the Euler to Lagrange point of view is known for the particle velocity field in its traditional form

$$\frac{\mathrm{D}\boldsymbol{v}}{\mathrm{D}t} = \frac{\mathrm{d_c}\boldsymbol{v}}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v}\ .$$

From left to right, these expressions are the *Lagrangian derivative*, also called the total derivative, the convective, convected, advective, substantial, material, particle, or covariant time derivative. The Lagrangian derivative can act on any scalars, vectors, and tensors that can be associated with the flowing particles, or with the fluid flow in general (see [121, 155] and references therein):

$$\frac{\partial \omega}{\partial t} = \frac{\mathrm{d_c}\Omega(\sigma,t)}{\mathrm{d}t} = \frac{\partial \Omega}{\partial t} + \boldsymbol{v}_{\mathrm{E}}(\Omega)\ ,$$

where $\boldsymbol{v}_{\mathrm{E}}$ is the Eulerian velocity vector, Ω is any tensorial quantity in the Eulerian frame, ω is the same quantity observed in the Lagrangian frame, (σ,t) are the Eulerian coordinates, and the last term on the right-hand side is the Lie derivative of Ω.

It is natural to use a Lie derivative for the rate of change of a geometric object carried along by the flow, because the Lie derivative is a method for computing a directional derivative with respect to the flow of the vector field. Let the totality of fluid particles describe a smooth manifold smoothly embedded in $\mathbb{R}^3$. We know already that the directional derivative of real-valued functions on a manifold already gives meaning to measurements of the change of objects in a direction specified by a vector. Any tangent vector $\boldsymbol{v} \in T_pM$ to a given manifold M at $p \in M$ (e.g., the fluid volume manifold) is by definition an operator that acts on a smooth function f to give a number $\boldsymbol{v}\cdot\nabla f(p)$ that we interpret as a directional derivative of f at p. This number can also be interpreted as the ordinary derivative of f along any curve whose initial tangent vector is $\boldsymbol{v}$ at p.

The generalization of this rate of change measurement to directional derivatives along a given vector field $\boldsymbol{w}(p)$ on M requires a little more caution than the same problem in a Euclidean space, since values cannot be compared directly. By replacing the vector $\boldsymbol{v} \in T_pM$ with a vector field, we can use the flow of the vector field to push values of $\boldsymbol{w}$ back to p and then differentiate. The result is the Lie derivative of f with respect to the given vector field $\boldsymbol{w}$.

Let us consider the position manifold for some fluid to be a 3D Euclidean space, and an absolute Cartesian frame of reference in it labeled by (x^i,t). We introduce a curvilinear orthogonal system of coordinates, also time dependent, parameterized

by coordinates (ξ^i, t), $i = 1, 2, 3$, and express the coordinate transformation through the diffeomorphisms $\xi^i = \xi^i(\boldsymbol{x}, t)$ with Jacobian

$$\bar{c}_i^j(\boldsymbol{x}, t) = \frac{\partial \xi^j}{\partial x^i} \,. \tag{4.23}$$

A particle moving with the fluid will have velocity $\boldsymbol{v} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$ in the Cartesian system, and the velocity $\boldsymbol{u} = \mathrm{d}\boldsymbol{\xi}/\mathrm{d}t$ in the curvilinear system. The transformation of the velocity components from one frame to another obeys the transformation law of a contravariant vector:

$$u^i = \bar{c}_j^i v^j \,, \quad v^i = c_j^i u^j \,, \quad \text{and } c_k^i \bar{c}_l^k = \delta_l^i \,. \tag{4.24}$$

This can be proved using (4.23) and the definitions of the velocities. We are now looking for a derivative with respect to t, the so-called *intrinsic derivative* $\mathrm{D}/\mathrm{D}t$ [145, 156], which measures the total variation of a tensor quantity along the curve due to an infinitesimal change in t. This should preserve the tensorial character when the coordinate system is changed. In Cartesian coordinates, the derivative $\mathrm{D}/\mathrm{D}t$ reduces to the material derivative.

As a test, let us choose an arbitrary parallel covariant vector field A_i with constant components in the Cartesian coordinates (x, t), and let $\bar{A}_i$ represent its transformation to the curvilinear space (ξ, t). Consider a curve describing the path of a particle in the fluid parameterized by $x^i(t)$ and by $\xi^i(t)$ in the two coordinate systems, respectively. From the requirement of having constant components along the t curve in the curvilinear system, we must have

$$\frac{\mathrm{D}\bar{A}_i}{\mathrm{D}t} = \frac{\mathrm{d}\bar{A}_i}{\mathrm{d}t} = 0 \,,$$

because its covariant derivative (along the curve), and also its total time derivative, should vanish, since physically, $\boldsymbol{A}$ represents the same constant parallel vector field as $\bar{\boldsymbol{A}}$ [156]. It follows that

$$\frac{\mathrm{d}\bar{A}_i}{\mathrm{d}t} = \frac{\mathrm{d}\bar{c}_i^j}{\mathrm{d}t} A_j + \bar{c}_i^j \frac{\mathrm{d}A_j}{\mathrm{d}t} = 0 \,. \tag{4.25}$$

Multiplying (4.25) on the left by c_p^i and considering the orthogonality relation in (4.24), we have

$$c_p^i \dot{\bar{c}}_i^j A_j + \frac{\mathrm{d}A_p}{\mathrm{d}t} = 0 \,, \tag{4.26}$$

and according to the arguments presented above, we can write

$$\frac{\mathrm{D}A_p}{\mathrm{D}t} = \frac{\mathrm{d}A_p}{\mathrm{d}t} + A_j c_p^i \frac{\mathrm{d}\bar{c}_i^j}{\mathrm{d}t} = \frac{\partial A_p}{\mathrm{d}t} + A_j c_p^i \frac{\mathrm{d}\bar{c}_i^j}{\mathrm{d}t} \,, \tag{4.27}$$

since $A_p(x^i, t)$ and x^i do not depend on t. The components A_i represent the vector field in the Cartesian system, and since the particle follows the t path in this space, (4.27) must represent the covariant derivative of A_i. Using

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\mathrm{d}x^j}{\mathrm{d}t}\frac{\partial}{\partial x^j} ,$$

for the derivative along the t path, we can apply (4.27) to write the covariant derivative acting on this covariant vector field in the form

$$\nabla_j A_p = \frac{\partial A_p}{\partial x^j} + A_k c^i_p \frac{\partial \bar{c}^k_i}{\partial x^j} , \tag{4.28}$$

from which we obtain the Christoffel symbols for this transformation of coordinates in the form

$$\Gamma^k_{pj} = -c^i_p \frac{\partial \bar{c}^k_i}{\partial x^j} . \tag{4.29}$$

We can express the covariant derivative in (4.27) and (4.28) in terms of the physical parameters, rather then the transformation matrices. In order to do this, we will choose as specific vector field the covariant components of the velocity of the fluid $A_p = v_p$. From (4.27), we have

$$\frac{\mathrm{D}v_p}{\mathrm{D}t} = \frac{\partial v_p}{\mathrm{d}t} + v_j c^i_p \frac{\mathrm{d}\bar{c}^j_i}{\mathrm{d}t} . \tag{4.30}$$

Using the definition in (4.23), the coefficient of the velocity in the last term of (4.30) can be rewritten in the form

$$c^i_p \frac{\mathrm{d}\bar{c}^j_i}{\mathrm{d}t} = c^i_p \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \xi^j}{\partial x^i} = c^i_p \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \xi_i}{\partial x_j} , \tag{4.31}$$

where we simultaneously switched the covariant/contravariant labels and denominator and numerator, and by orthogonality, nothing changes. Because ξ_i depends only on the independent variable x_j and t, the derivatives commute, and we have

$$c^i_p \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \xi_i}{\partial x_j} = c^i_p \frac{\partial}{\partial x_j}\frac{\mathrm{d}\xi_i}{\mathrm{d}t} = c^i_p \frac{\partial u_i}{\partial x_j} . \tag{4.32}$$

The matrix c^i_p represents the transformation from the curvilinear coordinates back to the Cartesian ones, and it does not depend on x_i. In consequence, we can write

$$c^i_p \frac{\partial u_i}{\partial x_j} = \frac{\partial}{\partial x_j} c^i_p u_i = \frac{\partial v_p}{\partial x_j} . \tag{4.33}$$

In the following, we combine this last result with (4.27) for $A_p = v_p$, and we obtain

$$\frac{\mathrm{D}v_p}{\mathrm{D}t} = \frac{\partial v_p}{\mathrm{d}t} + v_j c_p^i \frac{\mathrm{d}\bar{c}_i^j}{\mathrm{d}t} = \frac{\partial v_p}{\mathrm{d}t} + v_j \frac{\partial v_p}{\partial x_j} \,. \tag{4.34}$$

That is,

$$\frac{\mathrm{D}v_p}{\mathrm{D}t} = \frac{\partial v_p}{\mathrm{d}t} + \left(v_j \frac{\partial}{\partial x_j} \right) v_p \,, \tag{4.35}$$

which is nothing but the well known convective derivative in the Navier–Stokes equation (for the covariant components, in this case):

$$\frac{\mathrm{D}\boldsymbol{v}}{\mathrm{D}t} = \frac{\partial v_p}{\mathrm{d}t} + v_j c_p^i \frac{\mathrm{d}\bar{c}_i^j}{\mathrm{d}t} = \frac{\partial \boldsymbol{v}}{\mathrm{d}t} + \nabla_{\boldsymbol{v}} \boldsymbol{v} \,. \tag{4.36}$$

In this way, we have shown that the Lagrangian time derivative can be understood as a covariant derivative, too.

4.7 Deformation of the Boundary

A deeper mathematical understanding of problems related to boundaries must involve and relate different approaches, because of the special nature of this system. In many situations, for example, the combined use of geometric and algebraic approaches helps to provide the required intuition. The properties of a surface become better understood if we can classify it according to algebraic criteria, just as some algebraic problems can be better understood when their solutions are represented in a graphic mode. There are several ways to pair dual approaches between geometry and algebra in the study of boundaries: algebraic topology, algebraic geometry, differential algebra, etc. There is even more advantage in using such dual approaches when a problem is very well understood within one formalism and parts of this conceptual understanding can be transferred to another formalism.

We exemplify this transfer in the following. One very well studied field in mathematical physics concerns boundary value problems. In principle, such a problem is formulated by giving a certain space region D with smooth boundary $\Sigma = \partial D$, a differential or integral operator defined on this region, an equation involving the operator, an unknown function and possibly other source functions, and time dependent or independent boundary conditions for the unknown function on Σ. Solutions for such problems are very well developed thanks to the gallery of integral formulas generated by the Poincaré lemma, and various representation formulas. The behavior of the solutions is known within specific classes of

operators, boundaries, and source functions. Technically, the problems are solved by reducing them to Sturm–Liouville type problems (eigenvalues and eigenvector problems) with homogeneous boundary conditions (unknown function and/or its derivatives required to be zero on the boundary) and are classified in terms of the type of operator (elliptic, parabolic, hyperbolic, nonlinear, etc.). For such classes of solutions, called Green functions or fundamental solutions, the only variability is determined by the geometry of the boundary. Through this dependence, the final output of the analytic formalism, the eigenfunctions and eigenvalues, can be used to characterize the geometry of the boundaries. The operator spectrum resulting from the Dirichlet problem for a specific boundary depends to some extent on that boundary.

For example, if we choose a Dirichlet type of problem for the Laplace operator on a test region with surface, the eigenvalue problem is represented by a Helmholtz equation $\triangle \Phi = \lambda \Phi$. For Dirichlet (homogeneous) boundary conditions, we obtain different spectra for different geometries of the surface: sum of squares of two integers for a square, Bessel function roots for a disk, while the spectrum is an algebraic expression involving powers and square roots in the case of a 3D sphere.

In spite of the differences generated by the geometry and topology of the boundary, the spectra have some common features. The Laplace operator spectra for any 2D convex region are discrete, monotonic, and divergent ($\lambda_1 < \lambda_2 < \ldots \lambda_n < \ldots$), and the eigenvalues satisfy special algebraic relations like the Payne–Pólya–Weinberger inequalities, viz.,

$$\frac{\lambda_{n+1}}{\lambda_n} \leq 3 \ , \tag{4.37}$$

or the Hile–Protter or Yang inequalities [157], viz.,

$$\lambda_{n+1} \leq \frac{3}{n} \sum_{k=1}^{n} \lambda_k \ . \tag{4.38}$$

With this type of rational inequality, it makes sense to start a comparative study of the geometry of the boundary by evaluating the relative changes of the eigenvalues when the boundary deforms from one known surface to another, less well known geometry [158]. In addition to (4.37) and (4.38), we have available the famous Faber–Krahn inequality, which relates the smallest Laplace eigenvalue to the volume of the region:

$$\lambda_1 \geq \frac{\pi^2 \alpha_{0,1}^2}{|D|} \ ,$$

where $\alpha_{0,1} \sim 2.4048$ is the first positive zero of the Bessel function $j_0(x)$, and $|D|$ is the area/volume of the region D. In a series of papers [159], Ashbaugh and Benguria

have proved the fundamental inequalities for the Laplace operator spectrum in two dimensions:

$$\frac{\lambda_2}{\lambda_1} \leq \left(\frac{\alpha_{1,1}}{\alpha_{0,1}}\right)^2 , \tag{4.39}$$

where $\alpha_{0,1}, \alpha_{1,1}$ are the first zeros of the Bessel functions $j_0(x)$ and $j_1(x)$, respectively.

The question is to what extent the change in the spectrum can be accounted for by the change in the boundary? In other words, can two differently shaped drums give different sounds? Unfortunately, in general, the answer is negative [160]. There are calculations showing that one can find totally different shapes generating practically the same spectrum. Knowing the sound of a drum, determined by the eigenvalues of the Laplace operator in a Dirichlet type problem, that is, the frequencies of the fixed membrane assuming the shape of this drum, one can find information about its shape only in some particular situations. The eigenvalues of the Laplace eigenvalue problem for Dirichlet type boundary conditions can indeed be used as a reliable feature descriptor for shapes, if the shape is a 'small' deformation of a shape with known spectrum. In other words, we can identify shape variations of a drum by listening to its Laplace spectrum if the drum is close to a disk. In more colorful language, referring to (4.39), Marc Kac states [161] that "one can hear a convex drum if its first eigenvalue, or ratio of the first and second eigenvalues, are close to those of a disk".

Along the same line of research, it has been shown that all eigenvalues of the Laplace operator are region-monotonic, i.e.,

$$\lambda_n(D_1) \geq \lambda_n(D_2) ,$$

if $D_1 \subset D_2$, and have the scaling property

$$\lambda_n(cD) = \frac{\lambda_n(D)}{c^2} , \quad c > 0 .$$

These relations show that the smaller the domain D, the larger the absolute values of the eigenvalues. In other words, it takes more geometric energy to confine a system in a narrower region of space, or the well known fact that a larger drum generates a lower pitch.

In the field of shape recognition, for example, scientists use the ratios of eigenvalues of the Laplace operator for a Dirichlet type of boundary condition, to identify certain features of shapes [160]. Shape analysis is a key component in object recognition, matching, registration, and analysis, and works by generating a feature vector that attempts to uniquely characterize the silhouette of an object. There are an increasing number of studies on eigenvalue-based shape recognition methods, especially related to translation-, rotation-, and size-invariant shape recognition, with robustness and tolerance to shape deformation and noise [162]. The general

procedure is to study the change in some 'feature descriptors' with the deformation of the shape of a region D. The feature descriptors used by shape analysis scientists are ratios of eigenvalues in the form

$$F_1(D) = \left\{ \frac{\lambda_1}{\lambda_n} \right\}_{n=2,\ldots}, \tag{4.40}$$

$$F_2(D) = \left\{ \frac{\lambda_{n-1}}{\lambda_n} \right\}_{n=2,\ldots}. \tag{4.41}$$

In [160], the authors study simple classes of shapes like ellipses, rectangles, or circular symmetric images with a certain number of lobes, versus separability and variations like geometric noise (change on the boundary of the shape), topological noise (re-connections or holes), or hand-drawn shapes. They show a class separability of 96% or greater, and a clear distinction between their F_1 or F_2 descriptor series. For example, the values of the first descriptors in F_1 change from 0.48 to 0.54 when they change from 4-lobe to 5-lobe shapes, regardless of rotations and translations. Other applications include face recognition [162], or the impact of offshore wave farms on the near-shore wave climate [163].

The qualitative analysis presented above raises two questions. Firstly, whether there is a possible analytic way to obtain a quantitative, measurable correspondence between the change in the spectrum and the change in the shape, and secondly, whether there is a mathematical intuition behind the fact that such a correspondence is not one-to-one, but it is rather a global average dependence. An answer can be obtained by using the implicit function theorem together with the chain rule and some geometric technicalities like the transversality theorem. A complete study can be found in Henry's book [164].

The problem of the effect of the perturbation of the boundary on the associate eigenproblem and spectrum has attracted well known mathematicians, starting at the end of the nineteenth century. Rayleigh (1894) and Hadamard (1908) mentioned the problem for the first time in independent articles, while Courant and Hilbert and Polya and Szëgo detailed it in their well known books [165], and research has continued up to recent work like [166].

A rigorous mathematical way to evaluate the effects on the solutions and spectrum of a boundary problem when the boundary is deformed, or at least perturbed, would need to consider a space whose independent variable is the domain of definition of the boundary value problem.

The study of deformations of surfaces occurs in a natural way in the hydrodynamics of liquid drops, shells, and bubbles. In [167–169], the authors consider a liquid drop with free liquid boundary determined by a closed hypersurface Σ, element of a set $\mathcal{S}$ of closed surfaces in $\mathbb{R}^3$ diffeomorphic to the boundary of a reference region D and enclosing the same volume as D. In the vast majority of situations, the reference region is a D^3 ball, but there are cases where the reference region does not have a regular boundary or is not simply connected, as for example in the case of boiling Leidenfrost drops (see Sect. 9.7). A variation $\delta\Sigma$ of the boundary Σ can be defined

as the infinitesimal variation of Σ in its normal direction $\boldsymbol{N}$, having zero value for its integral. This condition obviously leads to a serious loss of the generality in the variations because it involves only divergence-free deformations. On the other hand, from the physical point of view, a fluid which is compressible has no special free surface, and no surface tension properties, so it is not so interesting.

More rigorously, let us define the deformation of the region D by a diffeomorphism $\eta : D \to D_\Sigma \subset \mathbb{R}^3$ which preserves the volume (divergence-free) and has a smooth boundary Σ. If we denote by $\mathcal{C}$ the manifold of all these volume-preserving diffeomorphisms η, we can associate a tangent space $T_\eta\mathcal{C}$ and a cotangent space $T^*_\eta\mathcal{C}$ with any $\eta \in \mathcal{C}$. In this formalism the variation $\delta\eta$ of the domain boundary is an element of the tangent space. For any function $F : \Sigma \to \mathbb{R}$, we can define its functional derivative $\delta F/\delta\Sigma$ with respect to the variations of the boundary in the form

$$\int_\Sigma \frac{\delta F}{\delta \Sigma}\delta\Sigma \mathrm{d}A = D_\Sigma F \cdot \delta\Sigma \ , \tag{4.42}$$

where the symbol $D_\Sigma F(x)$ is the differential of $F(x)$ with respect to the variation of the boundary defined in the Fréchet sense, viz.,

$$D_\Sigma F(x) = \left.\frac{\mathrm{d}}{\mathrm{d}\epsilon}\right|_{\epsilon=0} F(x_\epsilon) \ ,$$

and x_ϵ is a curve based on $x \in \Sigma$ and tangent to the normal $\boldsymbol{N}(x)$ to Σ at x. This definition of the variation of the boundary is useful when we investigate the dynamics of a physical system with free boundary, like a liquid drop. Based on this variation, one can define functional derivatives with respect to the boundary and with respect to the other parameters of the system (like the fluid velocity or pressure fields), introduce a well defined Poisson bracket, and finally introduce a Hamiltonian formalism for the free boundary. The process is presented in more detail in Sect. 9.6. One physical advantage in using (4.42) for the differential of the boundary is that it leads to a Poisson bracket structure in $\mathcal{C}$ which organizes the equations for a liquid drop with surface tension into a Hamiltonian form. Another physics advantage is that the motion of the boundary in the Lagrangian representation can be reconstructed from that in the Eulerian representation, and conversely. The integral curves of the corresponding canonical Hamiltonian describe the motion in the space of shapes and in the space of fluid velocities.

However, the most important result obtained from this definition of the variation of the boundary is a geometric one, namely, the existence of an intimate connection between any arbitrary change of boundary and the additional field of velocities induced in the fluid by this change. All the geometric elements needed to understand this construction can be found in Sect. 4.5. This is the way it works. We consider $\mathcal{B}$ to be the manifold of all possible boundaries Σ of a region with constant volume (area), that is, boundaries diffeomorphic to D_Σ, which is the boundary of the reference region D. If we consider $\mathcal{B}$ to be a base space, we can build a fiber bundle upon it (called a principal bundle) which is exactly $\mathcal{C}$, the material

configuration space, i.e., the manifold of all volume-preserving diffeomorphisms η of the reference region.

Indeed, the projection $\pi(\eta)$ of a region diffeomorphism onto $\mathcal{B}$ is the boundary of the reference region from which this diffeomorphism was initiated. That is $\pi(\eta) = \partial(\eta(D))$. Furthermore, the Lie group of volume-preserving diffeomorphisms of D acts along the standard fiber, so $\mathcal{C}$ is a principal bundle. In order to endow this principal bundle with geometric structure, we have to introduce a connection, that is a horizontal subspace H_η at each $\eta \in \mathcal{C}$.

Any harmonic function f defined on $\eta(D)$ can be interpreted as a potential for the velocity of the fluid. The gradients of all such functions generate the horizontal subspace of $\mathcal{C}$, i.e.,

$$H_\eta = \left\{ \nabla f(\eta(D)) \,\middle|\, f \text{ harmonic in } \eta(D) \right\} .$$

Such a connection, which is actually called the Ehresmann connection, prescribes a manner for lifting curves from the base manifold $\mathcal{B}$ into the total space of the fiber bundle $\mathcal{C}$ so that the tangents to the curves are horizontal. These horizontal lifts represent the parallel transport for this formalism. For each $\eta \in \mathcal{C}$, this horizontal lift is given by

$$h_\eta : T_{\pi(\eta)}\mathcal{B} \longrightarrow T_\eta \mathcal{C} .$$

Physically, $h_\eta(\delta\Sigma)$ may be thought of as the velocity field of flow determined by the boundary variation $\delta\Sigma$. Of course, the geometric construction requires only volume-preserving flows and only irrotational velocity fields.

In the following, we provide a few practical examples and quantitative evaluations concerning the perturbation of a boundary. Let us assume we have a reference domain $D \subset \mathbb{R}^3$ which is connected and simply connected, with smooth boundary Σ, and suppose we deform it by a diffeomorphism $\eta(D)$. Any smooth vector-valued function $f : D \to \mathbb{R}^3$ defined on this region will suffer a pull-back $\eta^* f(x) = f(\eta(x))$ for any point $x \in D$. We assume in addition that f satisfies certain differential equations (in general nonlinear) with the form $L(x,f) = 0$, where L is a differential operator acting on some space of smooth enough functions defined on D. When the region D is deformed, so is the differential operator, and we have

$$L_{\eta(D)} : \eta(D) \times (\text{domain of } f) \longrightarrow \mathbb{R}^3 ,$$

whenever the differential operator has a Eulerian form, where the coordinates are functions of the deformation parameter (for example, time) in a fixed coordinate system. On the other hand,

$$\eta^* L_{\eta(D)} \eta^{*-1} : D \times (\text{domain of } \eta^* f) \longrightarrow \mathbb{R}^3 ,$$

is the Lagrangian form. The advantage of the Lagrangian form over the Eulerian form in terms of handling differential operators consists in the fact that the operators

act in spaces which do not depend on η, facilitating the use of the implicit function theorem, for example. In a simpler language, we can say that the *Eulerian formalism* is related to the *geography*, while the *Lagrangian formalism* is related to the *history* of the system.

More detail of the formalism and analysis of the Eulerian and Lagrangian points of view can be found in Sects. 4.6, 9.1, and 9.5, and in even more detail in [121, 155] and the references therein.

In order to make a quantitative evaluation of the change in the boundary and the variation of any function depending on the boundary, we need to use either the Fréchet derivative along a curve in the space $\mathcal{C}$ of deformations as above, or just parameterize the deformations with a one-parameter Lie group of deformations, e.g., denoting $y = \eta(t,x)$ and $D(t) = \eta(t,D)$, with $t \geq 0$ and $\eta(0,x) = x$. We can also calculate

$$\boldsymbol{V}(t,\eta(t,x)) = \frac{\partial \eta}{\partial t} \ ,$$

recalling that $\boldsymbol{N}_{\Sigma(t)}$ is the unit normal to the deformed boundary Σ. Considering a smooth function defined on the reference domain $f(t,x) : [0,t_{\max}] \times D \to \mathbb{R}$, let us calculate the variation of the integral quantity

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{D(t)} f(t,y)\mathrm{d}y \ . \tag{4.43}$$

We can consider the variation of the domain as a regular smooth change of variables, so we apply the substitution formula for the triple integral:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{D(t)} f(t,y)\mathrm{d}y = \frac{\mathrm{d}}{\mathrm{d}t} \int_{D} f(t,\eta(t,x)) \left| \frac{D(y)}{D(x)} \right| \mathrm{d}x \ , \tag{4.44}$$

where $J = |D(y)/D(x)|$ is the coordinate transformation Jacobian. Performing the calculations and using the Jacobi formula, the derivative of the Jacobian is

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \mathrm{Tr}\left[\mathrm{Adj}\left(\frac{D(y)}{D(x)} \right) \frac{\mathrm{d}}{\mathrm{d}t} \frac{D(y)}{D(x)} \right] = J \boldsymbol{\nabla} \cdot \boldsymbol{V} \ ,$$

where Adj is the adjugate of the coordinate transformation matrix, i.e., the inverse of the matrix times its determinant, and Tr is the trace of the matrix. Using the divergence theorem, we obtain the final formula for the rate of change of the integral quantity as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{D(t)} f(t,y)\mathrm{d}y = \int_{D(t)} \frac{\partial f}{\partial t} \mathrm{d}y + \int_{\Sigma(t)} f \boldsymbol{V} \cdot \boldsymbol{N}_{\Sigma(t)} \mathrm{d}A_y \ , \tag{4.45}$$

where $\mathrm{d}A_y$ is the area element for the deformed boundary.

More important though is to calculate the variation of an integral quantity over the boundary itself, viz.,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Sigma(t)} f(t,y)\mathrm{d}A_y \ . \tag{4.46}$$

We can use the divergence theorem once again, convert the surface integral into a volume integral, apply (4.45) to this volume integral, and use the formula $\boldsymbol{\nabla}\cdot\boldsymbol{N} = 2H$ from differential geometry for the divergence of the unit normal, which tells us that it is equal to twice the mean curvature (see Appendix 1 in Chap. 9). We thus obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Sigma(t)} f(t,y)\mathrm{d}A_y = \int_{\Sigma(t)} \left\{ \frac{\partial f}{\partial t} + \boldsymbol{V}\cdot\boldsymbol{N}\big[\boldsymbol{N}\cdot\boldsymbol{\nabla} f)\big] + 2H_{\Sigma(t)} f \boldsymbol{V}\cdot\boldsymbol{N} \right\} \mathrm{d}A_y \ , \tag{4.47}$$

where $\boldsymbol{N}$ means $\boldsymbol{N}_{\Sigma(t)}$, and the gradient and the area element are calculated in the deformed coordinates y. This equation shows that the variation of the mean boundary value of a function relative to the change in the boundary has a term dependent on the changes in f generated by the variation of the region, and one term proportional to the mean curvature, both also proportional to the normal velocity of the deformed boundary. Again, there is a physical interpretation of (4.47): the right-hand term actually represents the balance between the convective change in momentum of the boundary and the surface tension induced by the mean curvature.

In the following, we present two concrete examples to show how (4.45) and (4.47) work. The first example is chosen from the theory of elasticity. Let $u(x) : D \subset \mathbb{R}^2 \to \mathbb{R}$ be the strain function of a cross-section D of a uniform cylinder under constant and uniform torsion [164]. The dimensionless equation expressing the equilibrium of forces and torques implies

$$\triangle u|_D = -1 \ , \qquad u|_\Sigma = 0 \ , \tag{4.48}$$

where as always $\Sigma = \partial D$ is the boundary curve of the flat 2D region D. The torsional rigidity is defined as

$$R(D) = \int_D |\nabla u|^2 \mathrm{d}x \ , \tag{4.49}$$

and we want to calculate how the torsional rigidity depends on the variation of the boundary of the cross-section. We expand the deformation functions in a Taylor series

$$\eta(t,x) = x + tV(x) + \mathcal{O}(t^2) \ , \quad u\big(\eta(t,x)\big) = u_0(x) + tu_1(x) + \mathcal{O}(t^2) \ . \tag{4.50}$$

Using the Jacobi formula, we obtain

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \int_D (u_1 + u_0 \nabla \cdot V) \mathrm{d}x \,. \tag{4.51}$$

The divergence term allows us to transform the domain integral into a contour integral, and by repeatedly using the boundary conditions, we obtain the final form

$$\frac{\mathrm{d}}{\mathrm{d}t} R\big(\eta(t,x)\big)\Big|_{t=0} = \int_\Sigma VN \left(\frac{\partial u_0}{\partial N}\right)^2 \mathrm{d}A \,, \tag{4.52}$$

where differentiation with respect to N means differentiating along a straight line following the local normal to the boundary, in a neighborhood of the boundary. Equation (4.52) shows an important result: the rate of change of the torsional rigidity integral depends, for small deformations (i.e., in a neighborhood of $t \simeq 0$) only on a contour integral and on the normal speed $\boldsymbol{N} \cdot \boldsymbol{V} = VN$ of deformation of the boundary, and the rate of variation of the function along the normal to the square.

Our second example shows how the spectrum of a Dirichlet type of problem in a region changes with a change in the boundary. We introduce the Dirichlet eigenproblem

$$\triangle u + \lambda u = 0 \ \text{ in } D \,, \quad u|_\Sigma = 0 \,, \tag{4.53}$$

and we ask how the spectrum $\lambda(\eta(t, D))$ changes with the deformation of the boundary. By differentiating the eigenproblem equation with respect to the deformation parameter, and by consecutive applications of (4.44) and (4.47), one obtains the fundamental result, for simple eigenvalues and normalized eigenfunctions [164],

$$\frac{\mathrm{d}}{\mathrm{d}t} \lambda\big(D(t)\big) = \int_{\Sigma(t)} \boldsymbol{V} \cdot \boldsymbol{N}_{\Sigma(t)} \left[\frac{\partial u\big(t, \eta(t,x)\big)}{\partial N}\right]^2 \mathrm{d}A_y \,, \tag{4.54}$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \lambda\big(D(t)\big)\Big|_{t=0} = -\int_\Sigma 2\boldsymbol{N} \cdot \boldsymbol{V} \frac{\partial u_0}{\partial N} \left(\frac{\partial \dot{v}}{\partial N} - 2\boldsymbol{N} \cdot \boldsymbol{V} H \frac{\partial u_0}{\partial N}\right) \,, \tag{4.55}$$

where we use the same notation convention for u as in (4.50). In (4.54) and (4.55), the variation of the eigenvalues induced by the variation of the boundary depends on the rate of change of the eigenfunction in the normal direction. This does not imply that, if the normal derivative of the eigenfunction is zero, the eigenvalues are constant. These formulas apply for the Dirichlet conditions only, viz., (4.53). Imposing zero normal derivative on top of the homogeneous Dirichlet condition will reduce the eigenfunction to zero, the trivial solution, anyway.

The calculations presented in this section have theoretical value, and in addition they provide excellent modern tools in various applied fields. For example, morpho-

metric studies of brain structures were successfully measured using properties of the Laplace–Beltrami eigenvalue spectra. Such spectral geometry analyses were used to yield shape descriptors capable of localizing geometric properties and detecting shape differences between patients [170].

All the constructions described above were inspired by fluid dynamics. Of course, in order to bring more intuition to the concept of variation of a boundary, one needs a real world example. However, these constructions can be extended to greater mathematical abstraction without relying on hydrodynamic intuition and fluid flows. This leads to the notion of cobordism. In addition, the cobordism equivalence relation replaces the volume-preserving group of diffeomeorphisms by a simpler and more elegant structure. This, however, is the subject of the next section.

4.8 Differential Topology of Boundaries: Cobordism

Cobordism is essentially an equivalence relation between compact manifolds of the same dimension, set up using the concept of the boundary of a manifold.

Definition 3 Two manifolds of equal dimension are *cobordant* if their union is the boundary of a compact manifold one dimension higher.

Cobordisms are studied both for the equivalence relation that they generate and as objects in their own right. Cobordism is a much coarser equivalence relation than diffeomorphism of manifolds, and is easier to study and compute. An example is presented in Fig. 4.10. Cobordisms are the subject of study in geometric topology and algebraic topology, being intimately connected with singularity and critical point theory (Morse theory), and in surgery theory. This concept is also one of the fundamental tools of study in topological quantum field theory. An n-manifold M is

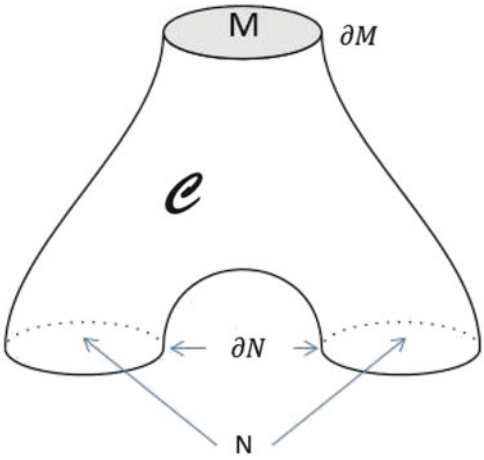


Fig. 4.10 An example of a cobordism $\mathcal{C}$ in $\mathbb{R}^3$ between a disk $M = D^1$ and a double disk manifold N

said to be null-cobordant if there is a cobordism between M and the empty manifold, that is, if M is a boundary of some $(n+1)$-dimensional manifold. A circle or an n-sphere are null-cobordant since they are boundaries of disks. Furthermore, every orientable surface is null-cobordant, because it is the boundary of a handle body (a sphere with two handles).

Definition 4 An $(n+1)$-dimensional cobordism W between n-dimensional manifolds M and N is an *h-cobordism* if the maps

$$M \hookrightarrow W \quad \text{and} \quad N \hookrightarrow W$$

are homotopy equivalences.

Two topological spaces are homotopy equivalent if there exist two continuous maps from one space to the other, and conversely, such that their compositions are the identity maps in both directions. In other words, two spaces are homotopy equivalent if they can be transformed into one another by continuous maps. For example, a solid disk or solid ball is homotopy equivalent to a point, and the plane without a point is homotopy equivalent to the unit circle S^1. As a comment, we mention that even if they are homotopic (one can deform the disk smoothly to a single point), a solid disk and one of its points are not homeomorphic since there is no bijection between them. Spaces that are homotopy equivalent to a point are said to be contractible.

Among other useful results, cobordism can provide a good classification tool for compact manifolds. The task of classifying compact manifolds of higher dimension by diffeomorphisms is a formidable one. It is possible to classify 2- and 3-manifolds up to diffeomorphism because these manifolds are geometrizable. Compact manifolds of dimension 4 are wild mathematical objects with highly exotic and unique properties. Higher-dimensional compact manifolds can be classified only using surgery, handle-body decomposition, and cobordism theories. We will elaborate more on this subject at the end of Sect. 5.8.

In order to simplify the presentation of this section, we consider all manifolds to be infinitely differentiable and embedded in some n-dimensional Euclidean space. We follow the introduction to cobordism in Milnor's book [152]. In order to introduce manifolds with boundary, we define the m-dimensional closed half-space

$$H^m = \left\{(x_1, \ldots, x_m) \in \mathbb{R}^n \mid x_m \geq 0\right\}.$$

The boundary of this half-space is the set ∂H^m defined by the hyperplane $\mathbb{R}^{m-1} \subset \mathbb{R}^m$. Then we can define:

Definition 5 A subset $X \subset \mathbb{R}^n$ is called a smooth m-manifold with boundary if for each of its points we can find a neighborhood U and a neighborhood V of H^m such that U and V are diffeomorphic. The boundary of X is the set ∂X of points corresponding to points on ∂H^m under such a diffeomorphism.

The boundary ∂X is itself a smooth $(m-1)$-manifold.

We can now introduce a more precise definition of cobordism. Consider two compact n-dimensional submanifolds N and N' of a larger manifold M of dimension m, and assume that all their boundaries are empty $\partial N = \partial N' = \partial M = \emptyset$. The difference of dimensions $m - n$ is called the codimension of the submanifolds.

Definition 6 We say that N is *cobordant* to N if the subset

$$N \times [0, t) \cup N' \times (1 - t, 1] \subset M \times [0, 1]$$

can be expanded to a compact manifold $X \subset M \times [0, 1]$ so that

$$\partial X = N \times \{0\} \cup N' \times \{1\} ,$$

in such a way that X does not intersect $M \times \{0\} \cup M \times \{1\}$ except at the points of ∂X. Here t is a real parameter in $[0, 1]$.

Cobordism is a fundamental equivalence relation on the class of compact manifolds of the same dimension, set up using the concept of the boundary of a manifold. Two manifolds of the same dimension are cobordant if their disjoint union is the boundary of a compact manifold one dimension higher. One of the most important facts of topology and geometry is that the boundary of an m-dimensional manifold M is an $(m-1)$-dimensional manifold ∂M that is closed, i.e., with empty boundary. In general, a closed manifold need not be a boundary: cobordism theory is the study of the difference between all closed manifolds and those that are boundaries.

A first very important result about boundaries is that there is no smooth map from a manifold with boundary onto its boundary that leaves its boundary fixed. In other words, there is no smooth map $f : X \to \partial X$ such that $f(\partial X) = \partial X$, that is, the boundary cannot be a fixed set (made of fixed points) for a contraction of the whole region inside. In Milnor's words, the identity map of a sphere S^n cannot be extended smoothly to a map $D^{n+1} \to S^n$. The proof is unexpectedly simple. Let us assume there is such a smooth map $f : D^2 \to S^1$ (see Fig. 4.11). Then for a

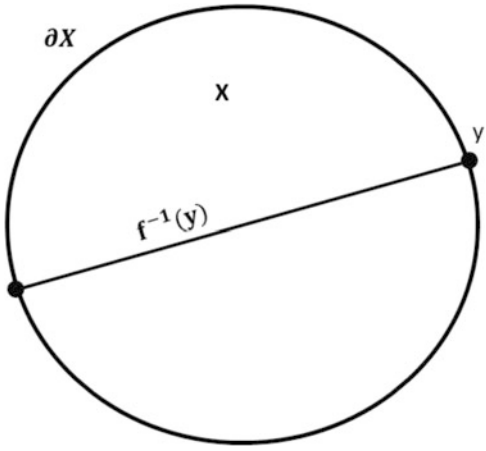


Fig. 4.11 Proof that the D^2 disk cannot be mapped smoothly onto its boundary, the circle S^1, in such a way that the circle is a collection of fixed points for this map

point $y \in \partial X$, its inverse image $f^{-1}(y)$ must be a compact manifold of dimension $2-1=1$, so it must be either a collection of segments with two boundary points, or a collection of circles without boundaries. So the number of boundary points of $f^{-1}(y)$ in X must be even. At the same time, if we want y in the boundary to be a fixed point of f, its inverse image must be the unique point y, so the inverse image must have only one boundary point, which contradicts the affirmation about an even number of boundary points.

Let us now present a few examples of how the theory of boundaries can help with the task of classifying manifolds. Consider a smooth manifold with non-empty boundary and consider the differential of the canonical inclusion $\iota : \partial M \to M$. Because the differential of this map [see (4.3) and (4.4)] is injective by construction, the image of $T_p(\partial M)$ is a well defined $(m-1)$-dimensional subspace of T_pM. This further implies that the vectors in $T_p(\partial M)$ viewed as vectors in T_pM have their last component equal to zero, independently of the chart. This has geometrical consequences. The tangent space of the boundary divides the tangent space of the manifold into two subspaces. In addition, any tangent vector to any curve touching the boundary has its last component non-negative, and positive inside M, which means that a vector from the inside tangent space points towards the inside if its last coordinate is positive.

A first result in the classification of manifolds according to their boundaries is the following theorem:

Theorem 3 *If the boundary of a compact smooth manifold M can be expressed as the disjoint union of two of its subsets, $\partial M = V_0 \cup V_1$, and if we can define a smooth real function f on the manifold which attains its maximum and minimum values only on these two subsets, respectively, and in addition it has nonzero differential everywhere, then the manifold is diffeomorphic to a hollow cylinder $V_0 \times [0,1]$ (see Fig. 4.12).*

The idea of the proof is to create a smooth vector field defined on M as the gradient of this function $f : M \to [0,1]$. It can be shown that there is always a smooth function g on M as the solution of the differential equation generated by this field, which maps one of the boundaries diffeomorphically, say V_0 as initial condition, throughout M.

Indeed, since the smooth function f has no extremal values in the interior of M, $\mathrm{d}f(M) \neq 0$, it attains its maximum and minimum values on M, and we can always renormalize it to have $f(V_0) = 0, f(V_1) = 1$, and $0 < f(\overset{\circ}{M}) < 1$. We consider the vector field

$$X = \frac{\nabla f}{\|\nabla f\|^2} ,$$

and let $g(p,t) : M \times \mathbb{R} \to M$ be the solution of the differential equation $\mathrm{d}g/\mathrm{d}t = \mathrm{d}g_{p,t}/\partial t = X_{f(p,t)}$ with initial condition $g(V_0, 0) = 0$. We have

$$\frac{\mathrm{d}}{\mathrm{d}t} fg = \nabla f \cdot \frac{\mathrm{d}g}{\mathrm{d}t} = \nabla f \cdot \frac{\mathrm{d}g}{\partial t} = \nabla f \cdot X = \nabla f \cdot \frac{\nabla f}{\|\nabla f\|^2} = 1 .$$

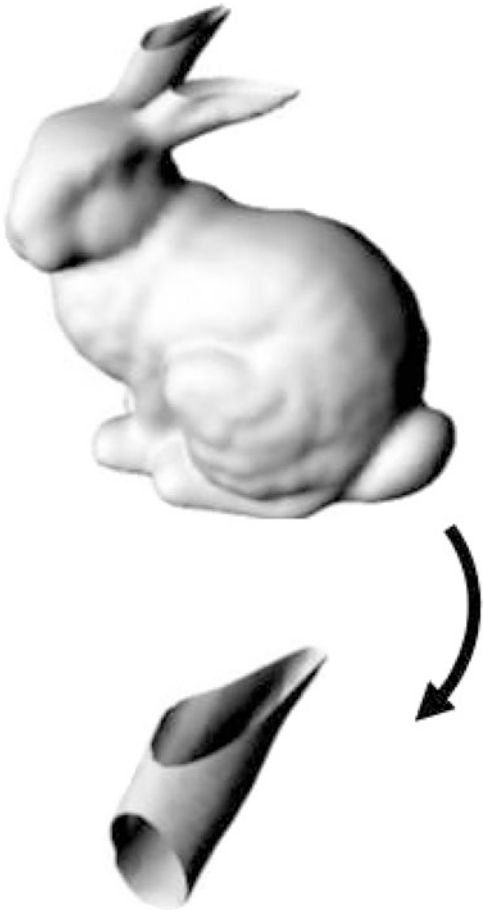


Fig. 4.12 A compact 2D manifold with boundary formed by two S^1 circles, the tips of the ears, is diffeomorphic to a finite cylinder

It follows by integration with respect to t that $fg(p,t) = t + f(p)$. Because M is compact and the resulting function fg is smooth, and because $g(p,t)$ cannot 'stop' inside M, it follows that t will assume all possible values when p runs over the whole of M, that is, $t = fg(p,t) - f(p) \in \big[-f(p), 1 - f(p)\big]$. In particular, the map $g(p,t) : V_0 \times [0,1] \to M$ is well defined, smooth, and invertible, so it is a diffeomorphism. Hence, when p covers one of the boundaries, say V_0, the parameter t maps this boundary all over M, and M is therefore a cylinder (see Fig. 4.12).

Another important result obtained directly from cobordism theory is the Hopf degree theorem:

Theorem 4 *A connected, oriented manifold M without boundary is smoothly homotopic to a sphere S^n if and only if they have the same degree.*

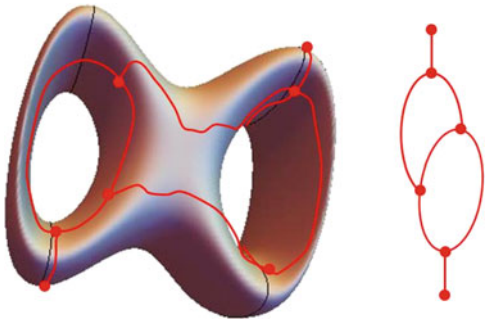


Fig. 4.13 A torus of genus 2 (*left*) and the graph of its critical points (*red and right*). Each contour determines an equivalence class which is represented in the critical point graph by a single point

One immediate consequence is:

Theorem 5 *Any smooth connected one-dimensional manifold is diffeomorphic either to the circle* S^1 *or to some interval on the real axis.*

The degree is calculated using (4.6). These results illustrate the scope of differential techniques before homotopy classification became primarily a project of algebraic topology. The degree of the map somehow measures the number of critical points of the boundary, because a change in the sign of df is in general associated with a zero, that is, a critical point, if the manifold is smooth and connected. The connection between the degree and the boundary shape revealed by the Hopf degree theorem has many applications. For example, in the case when the boundary is represented by the graphic of a real function $z = f(x, y)$, the type of topological changes on the surface are determined by the type of critical points, which is given by the number of negative eigenvalues of its Hessian. In this way, it is possible to extract a topological graph based on points, edges, and faces (what is known in algebraic topology as a CW complex). Figure 4.13 exemplifies this for a double torus surface that can be represented by a simple graph whose nodes are the critical points of the surface.

The graph representation of a surface through its critical points and the arcs between them is rather simple and can represent the main topological properties of a boundary. Because this property ensures efficient topological control in the compression and simplification of boundaries, it can be used to study the metamorphosis (morphing) of objects. This field of application needs a mathematical treatment that produces a rapid topological detection, classifies the common shape features, and preserves the low-level geometrical information. Such a representation has to allow for direct transformation and modification on the model and give an effective theoretical support. In fact, morphing is a technique used to analyse the evolution and metamorphosis from one image to another [171]. The idea is to get a sequence of intermediate configurations which, when put together with the original surface, would represent the change from one to the other. Applications include shape recognition, metamorphosis processes, ocean surface wave reconstruction, and space missions.

Chapter 5
Discrete Mathematics

There is no problem in the whole of mathematics which cannot be solved by direct counting

Ernst Mach

The revolutionary growth of experimental data in the sciences, and the availability of unprecedented computing power shape and challenge all the fields of mathematics, from traditional to contemporary, and from continuous to discrete. New fields like computational algebraic topology and computational geometry have appeared, combining efforts to develop mathematical tools for a broad new perspective of research. To enumerate only a few: the topological and statistical analysis of shapes, images, and high-dimensional data sets; algorithms for motion planning and the study of configuration spaces of mechanical systems; stochastic topology and the study of large growing systems; the theory of concurrent computation and computer networks, etc.

The fundamental relation between continuous and discrete mathematics is not the sole result of the computer revolution, and somehow the concept of boundary mediates the transition from continuous to discrete. There is an interesting relation between the topological boundary (a continuous concept) and the concept of dimension (a discrete concept). Indeed, based on the fact that in any n-dimensional Euclidean space the boundaries of n-dimensional disks have dimension $n - 1$, one can define the dimension of a space in an inductive manner in terms of the dimensions of the boundaries of suitable open sets. Put simply, we can define an *inductive dimension* counting up from -1, the dimension of the empty set, by induction. The inductive dimension of a set X is the smallest natural number $n = \text{ind}(X)$ such that any point in X has an open neighborhood included in X whose boundary has inductive dimension less than or equal to $n - 1$.

In this chapter we want to correlate the recent results of research with the importance of the boundary for graph and network theories. In order to take advantage of the numerous quantitative results relating parameters describing the boundaries of a graph or network (like volume, girth, etc.) with the type of graph, we shall introduce here some basic definitions and properties of abstract graphs.

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_5

5.1 Structured Finite Sets

A finite set of points can be structured by equivalence and order relations, and more importantly by directed and undirected links. Such objects usually fall into the broad category of graphs and networks. Graph and network theory is one of those areas of high interest and applicability where one can easily move from the geometric meaning to the algebraic and statistical one. In its most general sense, a graph is a category where objects are called *nodes* and morphisms between them are called *edges*. Functors between the graph categories would be modes of understanding graphs, and these come in various ways, depending on the reader's familiarity with a certain area of knowledge in computer science.

A possible interpretation of a graph can be the description of the states of a system, that is, a functor between a category of thermodynamical states and a drawing. We may gather all the states of the system as elements of an abstract space. Scientists like to introduce a state space injectively into some geometric space, for example some n-dimensional real affine space. This map associates with each state or configuration a point in the real affine space called a *node*. The system can perform transitions from one state to another, or even from one state to more than one state, the last possibility being understood as 'in principle'. As in the case of an ice crystal warmed up by an energy pulse: it can transform into a drop of water, it can vaporize, it can be ionized into a cloud of plasma, or it can be transformed into a shower of elementary particles. If we know all possible states, and we list the possible and impossible transitions, we build a graph with the states as nodes, and transitions as edges.

Another simple interpretation consists in selecting a natural number n and building an abstract set of cardinality n called the set of nodes. Next, we build a square matrix with zero for all its entries and size equal to the number of nodes. This matrix will be defined more exactly in Sect. 5.3. If we start to convert some of the entries of this matrix into 1s in some manner, we obtain a graph in which the coordinates of the 1 determine the labels of nodes connected by edges.

We divide this chapter into three sections as follows. In Sect. 5.2, we introduce the prerequisites for the formal theory of graphs, including definitions and classifications, with a few examples and some basic bounding theorems. In Sect. 5.3, we introduce algebraic approaches to graphs, the operators and matrices associated with a graph, their spectra and the relation between eigenvalues, and some topological, mainly boundedness, properties. Finally, in Sect. 5.4, we discuss the main topological properties of graphs relating to their boundaries.

5.2 Formal Theory of Graphs

Any set A described in this chapter is considered finite, and we denote the number of its elements $\text{card}(A) = |A|$. An abstract *graph* $G(N, E)$ is a finite set of points (nodes, vertices) submerged in a finite-dimensional Euclidean space and denoted

$N(G)$. We also introduce a set of continuous curves $E(G)$ joining the nodes and called edges or links. The set $E(G)$ is always a subset of $N(G)\times N(G)-\text{diag}\big(N(G)\times N(G)\big)$. The number of nodes of a graph is called the *order* of the graph, and it is denoted by $|G| = |N(G)|$, while the number of edges is denoted $\|G\| = |E(G)|$. An edge $e \in E(G)$ has two end nodes $i,j \in N(G)$, called adjacent or neighbor nodes. If two nodes $\{i,j\} \in N(G)$ represent the end points of an edge, we can write $i \sim j$ or $(ij) = e_{ij} \in E(G)$. We denote the edge by its end nodes $e = (ij) \in E(G)$. The set of all edges connected to a node $i \in N(G)$ is denoted $E(i) \subset E(G)$, and the *degree* of the node i (or its *valency*) is defined as $|E(i)| = d(i)$. All the nodes connected to this node by an edge are called its neighbors. If for any two nodes $(ij) = (ji)$, the graph is called *undirected*, and *directed* otherwise, and examples of graphs are shown in Figs. 5.1 and 5.2. A graph $G'(N',E')$ is subgraph of a graph $G(N,E)$ if $N'(G') \subset N(G)$ and $E'(G') \subset E(G)$, and if any two nodes are connected the graph is said to be *connected*.

If all the nodes have the same degree k, then we call G a *k-regular* graph (see the enumeration below). We define the average degree and the volume of a graph by [172]:

$$\langle d\rangle \equiv \frac{1}{|G|}\sum_{i\in N(G)} d(i) = \frac{\text{vol}(G)}{|G|} = 2\frac{\|G\|}{|G|}\ . \tag{5.1}$$

The sum of all degrees $\text{vol}(G) = \sum_{i\in N(G)} d(i)$ stands for the *volume of the graph*. The *density* of a graph is $2\|G\|/|G|(|G|-1)$. An *open path*, or simply a path between

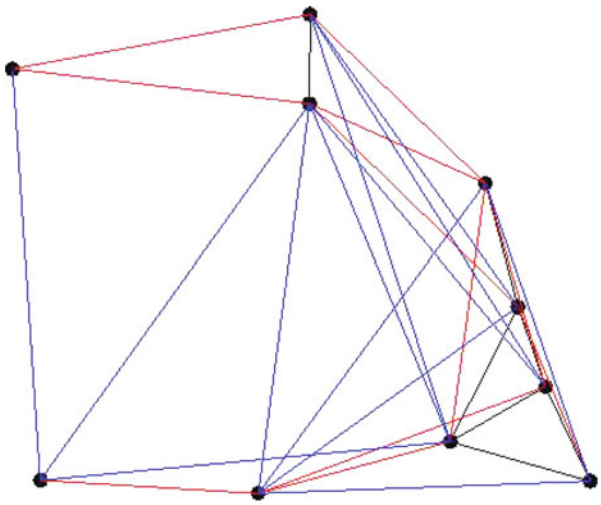


Fig. 5.1 Example of a graph with $|G| = 10$, $\|G\| = 62$. The graph's nodes are generated randomly and different edges are colored according to the length, in order *from black, red, blue, to green*. The nodes can represent geographic locations, while edges of different colors represent routes with different means of transportation

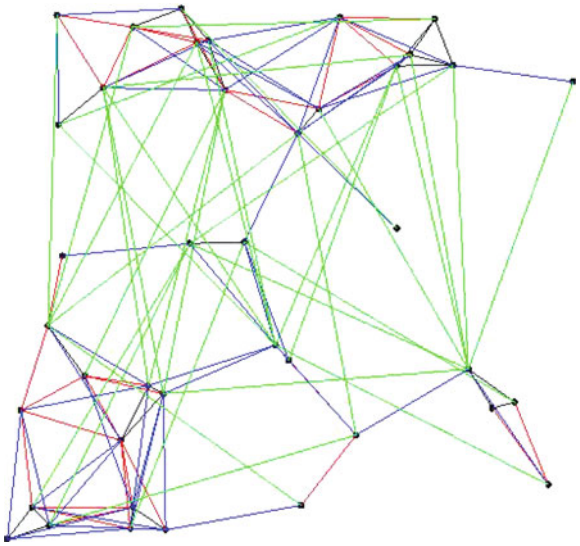


Fig. 5.2 A graph with $|G| = 40$, $\|G\| = 247$ which has the graph in Fig. 5.1 as a subgraph. In addition to that previous graph, we have added more distant nodes and more edges to connect at greater distances (*green*)

nodes i and j in a graph is a connected subgraph P containing i, j as nodes of order 1, and having all the other nodes of order 2. We denote by $\mathcal{P}(i,j)$ the set of all possible paths $P = P(i,j)$ joining the nodes i, j in G. In a path, the nodes do not repeat, unless the end nodes coincide, in which case the closed path P is a *cycle* Z with all its nodes of order 2. Furthermore, a cycle can have two orientations! The number of all cycles in a graph is given by the Euler formula $\|G\| - |G| + |CC|$, where we denote by $|CC|$ the number of its connected components.

Definition 7 We define the distance between two nodes as the minimum number of edges between the two nodes $i, j \in G$:

$$d(i,j) = \min \|\text{path } i \to j\| \ ,$$

and we define the diameter of G as

$$\text{dia}(G) = \max_{i,j \in N(G)} d(i,j) \ .$$

We can further introduce the eccentricity of G as

$$\text{ecc}(G) = \max_{i,j \in N(G)} d(i,j) \ .$$

The length of the shortest/longest cycle in a graph is the girth/circumference of the graph.

A *walk* is a connected union of paths with only 0 or 2 end nodes. In a walk, we can repeat the nodes, and the *length* of a path/cycle is the number of its edges,

$|P| = \text{card}\{(i,j)/(i,j) \in \mathcal{P}(i,j)\}$. We denote a k-path (path of length k) by P^k and a k-cycle (cycle of length k) by Z^k. Consequently, the sets of all paths, walks, and cycles are ordered sets. A graph G is *Hamiltonian* if there exists a path including all its nodes. Two paths in a graph are *disjoint* if they have in common at most their end points, and a *hop* is an operator allowing the motion from one node to a neighboring one.

Among all the many types of graphs, we would like to mention some special ones that provide interesting body–boundary relationships. A complete graph denoted $K_{|G|}$ has all its nodes adjacent to each other, and all nodes have the same maximum degree $d_{\max}|G| - 1$. A k-regular graph has all nodes of the same degree k. A 2-regular graph has its nodes aligned along a circle and has the maximum girth. A planar graph is an embedding in a compact and connected set in the real plane such that nodes correspond to points and edges correspond to arcs between points without multiple intersections. The bipartite graph has its nodes divided into two disjoint sets. A k-node connected graph has the property that, if we remove any k or fewer of its nodes, the graph remains connected. Random graphs are constructed by associating a certain probability criterion with a certain property of the graph, e.g., the probability of occurrence of an edge between any two arbitrary nodes is a random number between zero and one.

In the following, we present some quantifiers describing the topology of graphs. All topological definitions are based on the existence of a minimum (infimum) or maximum (supremum) of the set of distances in the graphs. We define the distance between two connected nodes $d(i,j)$ as the length of the shortest path between them in G:

$$d(i,j) = \inf_{P(i,j)\in\mathcal{P}(i,j)}|P(i,j)| \ .$$

For any node i in G, the longest path starting from it is called the *eccentricity* of the node:

$$\epsilon(i) = \sup_{j\in N(G)} d(i,j) \ .$$

The *radius* of a graph G is defined as the number $R(G) = \text{rad}(G)$, defined as the shortest of all its eccentricities, for all the nodes in the graph. In other words the radius of a graph is the shortest longest distance over the whole of G, i.e.,

$$R(G) = \text{rad}(G) = \inf_{i\in N(G)}\sup_{j\in N(G)} d(i,j) \ .$$

The *diameter* of G is the greatest distance between any two nodes in G (see an example of such a diameter path in Fig. 5.3):

$$D(G) = \text{dia}(G) = \sup_{i,j\in N(G)} d(i,j) \ .$$

The length of the shortest cycle in G is the *girth* $g(G)$ of G, while the length of the longest cycle is the *circumference* $\text{circ}(G)$ of G. The two numbers are always in

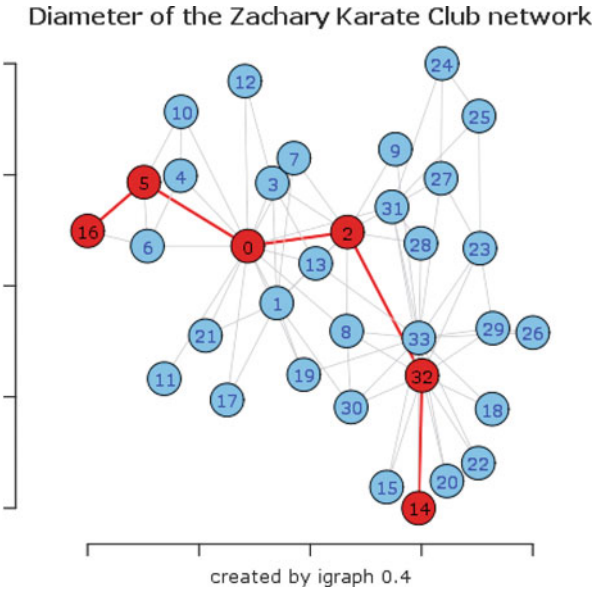


Fig. 5.3 Example of a diameter path. Courtesy of igraph 0.4

the relation $g(G) \leq \text{circ}(G)$. A node which is placed such that its longest distance to all other nodes is equal to the radius of the graph is called a *central* node. The radius and diameter of a graph are not in an exact algebraic relation, but they satisfy a sort of monotony: if one is larger (smaller) the other is also larger (smaller). This happens because

$$\text{rad}(G) \geq \text{dia}(G) \geq 2\text{rad}(G) \; . \tag{5.2}$$

We characterize different types of connected and undirected graphs by using two dimensionless parameters $1 \leq D/R \leq 2$ and $0 \leq g/\text{circ} \leq 1$ (see, for example, Fig. 5.4). The radius, diameter, and degree provide a first description of the topology of a graph. They are interconnected through some relations which we present here.

Theorem 6 *A graph G with radius and degree bounded from above, i.e., $R \leq R_{\max}$, $\sup_{i \in N(G)} d(i) \leq d_{\max} \geq 3$, has an upper bound for its number of nodes:*

$$|G| \leq \frac{d_{\max}(d_{\max}-1)^{R_{\max}}}{d_{\max}-2} \; .$$

The proof can be found in [172]. There is also a lower bound theorem for the number of nodes [172]:

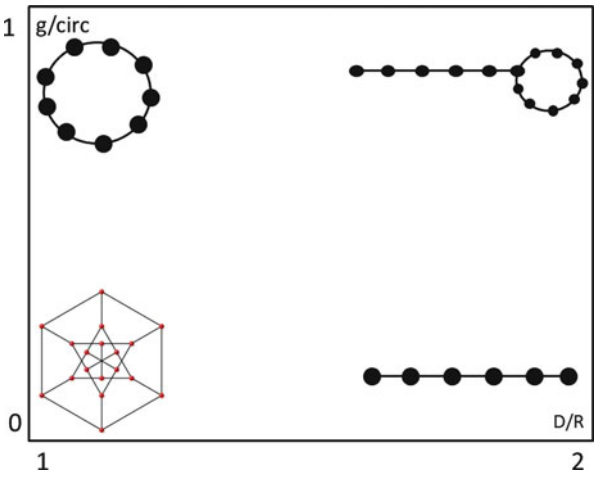


Fig. 5.4 Four types of graphs having extreme values for the ratios D/R and g/circ. *Clockwise from bottom right*: A linear graph $D \simeq 2R = |G|$ and $g = 0$; a Pappus graph $D \simeq R$ and $g = 3 <$ circ; a ring graph $R = D = n/2$ and $g =$circ; a combination of linear and ring $D = 2R$, $g =$circ

Theorem 7 *A graph G with average degree and girth bounded from below has a lower bound for G. That is, if* $d(G) \geq d_{\min}$ *and* $g(G) \geq g_{\min}$, *then*

$$|G| \geq \begin{cases} 1 + d_{\min} \sum_{i=0}^{r-1} (d_{\min} - 1)^i & \text{for } g(G) = 2r + 1 \ , \\ 2 \sum_{i=0}^{r-1} (d_{\min} - 1)^i & \text{for } g(G) = 2r \ . \end{cases}$$

In other words, scarcity of nodes in a graph involves either a low degree or narrow cycles. A regular graph with a minimum number of nodes for its given girth and degree is called a *cage* graph. We also mention the cornerstone of graph theory, known as Menger's theorem [172]:

Theorem 8 *For an undirected graph G and for any two arbitrary sets of nodes* $N_{1,2} \subset N(G)$, *the minimum number of nodes separating* N_1 *from* N_2 *in G is equal to the number of disjoint paths in G beginning at* N_1 *and ending at* N_2.

It is interesting to attempt to classify graphs in terms of the topology and geometry of their boundaries, when these concepts can be applied. In the case of planar graphs, we can mention the following important types of graphs [173–176]. Graphs with a regular polygon for boundary are called *wheel graphs* W_N, and are constructed as graphs with $|G| = N \geq 4$ nodes in such a way that one node of degree $n - 1$ is connected with all other nodes of degree 3 which form one $(n-1)$-cycle. The wheel graphs have diameter 2 and girth $g = 3$. The class of graphs with spherical shape includes regular graphs of higher degree than 3, small-world networks, and cage graphs. Cubic graphs have nodes of degree 3 and their boundaries take the shape of a polyhedron, or a crystal lattice.

5.3 Algebraic Theory and Spectra of Graphs

Graphs can be studied fairly comprehensively by using linear algebra. In this section, we introduce these algebraic properties, especially the graph spectra, while in Sect. 5.4, we discuss graph topology and boundaries. The main philosophy of the connection between graph spectra and graph topology, and in particular a graph's boundary, size, shape, etc., is this: the first two (the smallest) and the last two (the largest) of the eigenvalues of the spectrum of the normalized or non-normalized Laplacian control its size and topology, while the central part of the spectrum is of little relevance here. With (5.16), we will show how adding edges to a graph increases both the density and the length of these spectra. But adding edges may not change the graph topology at all. It is not so much how the spectrum looks, but rather its end eigenvalues that control the topology, a fact which will be abundantly exemplified in the following discussion.

For any undirected connected graph with $n = |G|$ nodes labeled from 1 to n, and $m = \|G\|$ edges, we can build six important matrices associated with the graph: the adjacency matrix $A_{n\times n}$, the incidence matrix $B_{n\times m}$, the diagonal matrix $\Delta_{n\times n}$, the Laplacian matrix $L_{n\times n}$, the signless Laplacian matrix $Q_{n\times n}$, and the normalized Laplacian matrix $L_{n\times n}$.

The graph matrix with the most properties is the adjacency matrix, which acts in the space of nodes and describes how the nodes are connected. In order to understand how this matrix works, we associate the n-dimensional real space $\boldsymbol{v} \in \{0,1\}^n$ with the graph, and also associate with each node a basis vector from the canonical basis in this vector space. Every time we apply the matrix A to the left of a basis vector i, the resulting vector has ones at the positions of the nodes connected to i. Recurrent application of A shows how connectivity propagates through the graph from the node i, making the graph work like an actual network. For example, in Figs. 5.5, 5.6, 5.7, we show the repeated action of A upon the vector associated with node 1.

The diagonal matrix Δ has the form $(\delta_{ij}d(i))$ and it measures the degree of each node. The incidence matrix B is no longer square, and it describes what edges are connected to what nodes and how. A way to understand the action of B is to label and order the edges by their end nodes in lexicographic order, i.e.,

$$(e_1, e_2, \dots, e_\alpha, \dots) = (e_{i,j}, e_{i,i+j}, \dots, e_{i+l,i+l+n}, e_{i+l,i+l+n+m}, \dots) ,$$

with $i, j, l, n, m > 0$. The incidence matrix is the most interesting for our approach in this book, because it is nothing but the *boundary operator* from homology theory in its discrete version (see also Sect. 5.4). The matrix B also plays the role of a conservation law because it acts on the space of edges and has values in the space of nodes. It is practically identical to Kirchhoff's second law for electrical networks [172, 173]. Sometimes, especially when undirected graphs are used, one uses instead the matrix $R = (|B_{i,\alpha}|)$.

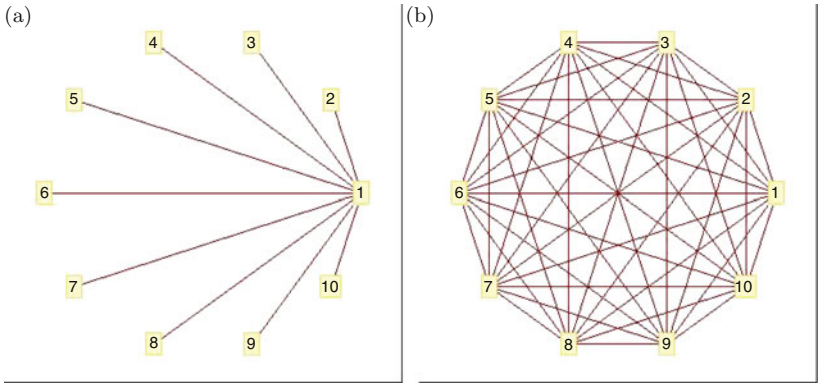


Fig. 5.5 Action of operator A on a complete graph with 10 nodes starting from node 1. It took just two iterations to complete the graph. (**a**) $A(1,0,0,0,0,0,0,0,0,0)$. (**b**) $A^2(1,0,\ldots,0) = K_{10}$

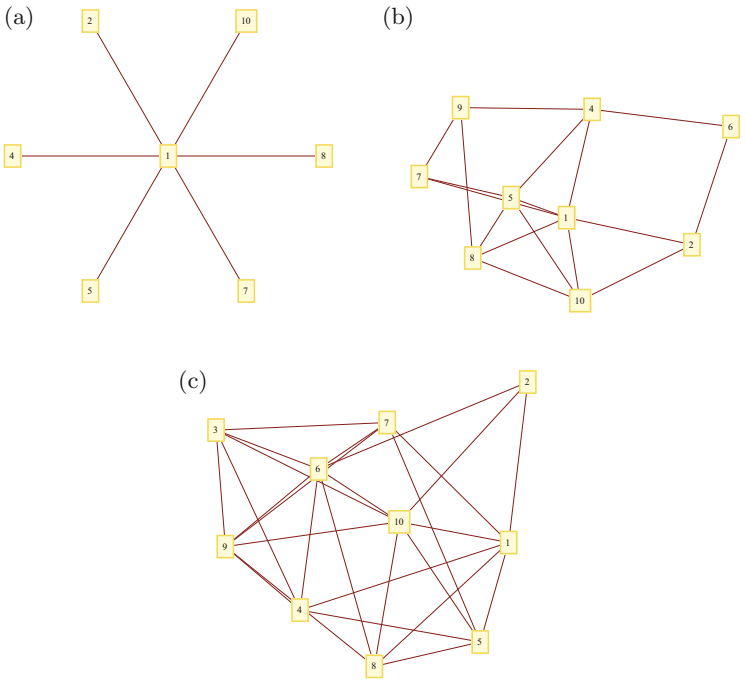


Fig. 5.6 Another example of propagation through the graph by the action of the matrix A. The graph is strongly connected, with size 10 nodes, and the initial vector is $(1,0,0,0,0,0,0,0,0,0)$. It took three steps to complete the graph with the operator A, as compared with eight steps to complete the graph in the previous figure. (**a**) $A(1,0,\ldots,0)$. (**b**) $A^2(1,0,\ldots,0)$. (**c**) $A^3(1,0,\ldots,0) = N(G)$

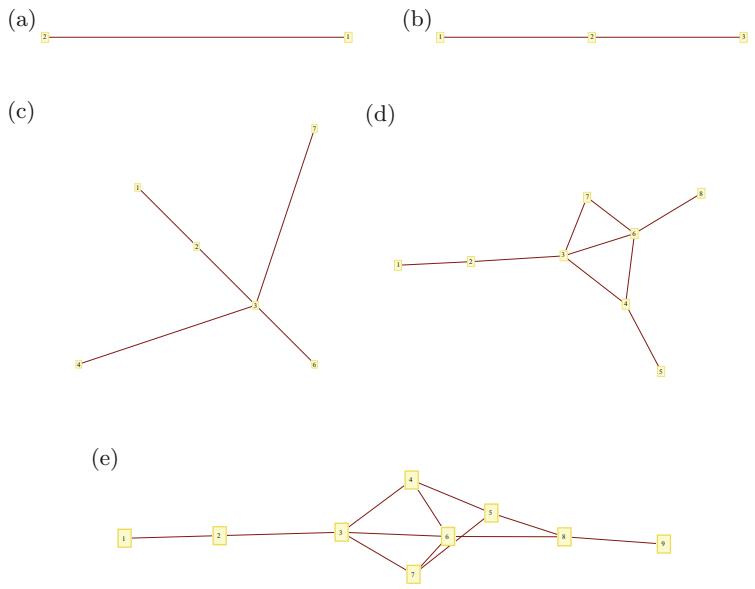


Fig. 5.7 Another example of propagation through the graph by the action of the matrix A. The graph is weakly connected, almost a single path, with size 9 nodes, and the initial vector is still $(1,0,0,0,0,0,0,0,0)$. It took 6 steps to complete the graph. (**a**) $A(1,0,\ldots,0)$. (**b**) $A^2(1,0,\ldots,0)$. (**c**) $A^3(1,0,\ldots,0)$. (**d**) $A^4(1,0,\ldots,0)$. (**e**) $A^5(1,0,\ldots,0)$

The next three matrices are built from A, B, and Δ :

$$L = \Delta - A\,, \quad \text{the Laplacian}\,, \tag{5.3}$$

$$Q = \Delta + A\,, \quad \text{the signless Laplacian}\,, \tag{5.4}$$

$$\mathcal{L} = \begin{cases} 1 & \text{if } i = j\,, \\ -\dfrac{1}{\sqrt{d(i)d(j)}} & \text{if } (ij) \in E(G)\,, \\ 0 & \text{otherwise}\,, \end{cases} \quad \text{the normalized Laplacian}\,. \tag{5.5}$$

The matrix L is also called the Kirchhoff matrix (it is useful in electrical network studies), or the admittance matrix, while the matrix Q is also called the co-Laplacian, and the matrix $\mathcal{L}$ is also called the correlation or transition matrix because it plays an important role in the study of random walks.

Except for the matrix B, all the operators defined above are symmetric and positive-definite, so they have real, discrete, and bounded spectra. In the literature,

one labels the eigenvalues of the three fundamental graph matrices in the form:

$$\begin{aligned} A\boldsymbol{u} &= \theta_i \boldsymbol{u}\ , \quad & \theta_1 \geq \theta_2 \geq \cdots \geq \theta_n\ , \\ L\boldsymbol{u} &= \lambda_i \boldsymbol{u}\ , \quad & 0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n\ , \\ \mathcal{L}\boldsymbol{u} &= \bar{\lambda}_i \boldsymbol{u}\ , \quad & 0 < \bar{\lambda}_1 \leq \bar{\lambda}_2 \leq \cdots \leq \bar{\lambda}_n\ , \text{ and always } 0 \leq \bar{\lambda}_i \leq 2\ , \\ Q\boldsymbol{u} &= \mu_i \boldsymbol{u}\ , \quad & \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n\ . \end{aligned} \tag{5.6}$$

The largest eigenvalue of A is called the degree of G. The normalized Laplacian eigenvalues are all bounded between 0 and 2, its spectrum is symmetric with respect to 1, and a high multiplicity for the eigenvalue 1 may show duplications in the graph. Furthermore, the high density of its eigenvalues near 2 is a sign that the graph is bipartite or close. There are relations between these matrices, viz., $Q = BB^{\mathrm{t}}$, $\mathrm{Tr}Q = \mathrm{Tr}L$, and $\mathcal{L} = \Delta^{-1/2} L \Delta^{1/2} = BB^{\mathrm{t}}$.

All these matrices act like operators on functions $g : N(G) \to \mathbb{R}$ defined on nodes. For example, the normalized Laplacian acts in the form [175, 176]

$$\mathcal{L}g(i) = \frac{1}{\sqrt{d(i)}} \sum_{ij \in E(G)} \left[\frac{g(i)}{\sqrt{d(i)}} - \frac{g(j)}{\sqrt{d(j)}} \right] . \tag{5.7}$$

The incidence matrix has a profound topological meaning, because it acts like a *boundary operator* in homology, whence it is itself called the *graph boundary operator*. The boundary operator B transforms $\|G\|$-dimensional vectors $\boldsymbol{u} \in \{0,1\}^{\|G\|}$, i.e., vectors describing subsets of edges, into $|G|$-dimensional vectors $\boldsymbol{v} \in \{0,1\}^{|G|}$, i.e., vectors identifying subsets of nodes. If applied to a chain of edges (we can use the C_1 notation of a one-dimensional cycle from the graph considered as a simplicial complex), it generates its boundaries (the corresponding C_0 set of points), that is, the nodes describing the beginning and end of the cycle $B(C_1) \to C_0$, $C_1 \leftarrow B^{\mathrm{t}} C_0$. Figure 5.8 shows an example of a simple directed graph with incidence matrix in the form

$$B = \begin{pmatrix} -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} . \tag{5.8}$$

If we apply the boundary matrix B on the cycle $(1,0,0,-1,1,0) \subset E(G)$, that is the cycle (125), we obtain no boundary, or an empty set of nodes $(0,0,0,0,0) \subset N(G)$. If we apply B on the chain $(1,1,1,0,0,1) \subset E(G)$, we obtain the boundary 1 and 5 of this chain, namely the nodes $(-1,0,0,0,1)$.

The Laplacian matrix acting on functions of nodes acts in the same way as the Laplace–Beltrami operator acts on functions defined on oriented Riemannian manifolds (see Sect. 4.3) [175]. The theorem for the Rayleigh quotient and the minimum eigenvalue can be used rather as an approximation in the Riemannian

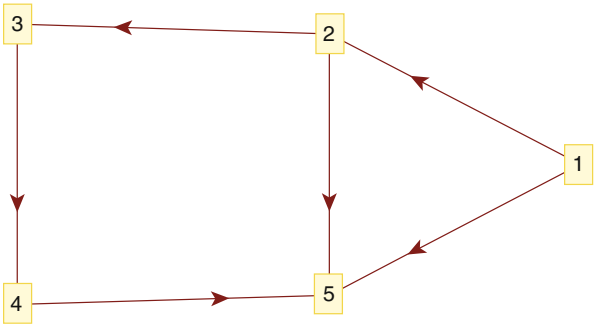


Fig. 5.8 A directed graph with edges labeled in the order 12, 23, 34, 15, 25, 45. If we apply its B matrix to the cycle (125), we obtain the null boundary set. If we apply B to the chain of nodes $\{1, 2, 3, 4, 5\}$, we obtain the boundary 'from' 1 'to' 5

manifold case, while in graph theory where all norms are finite sums, this theorem becomes a rigorous and important tool.

For a Dirichlet boundary problem, it is known that the solutions form a discrete bounded spectrum $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{\max}$, and the following *principle of the minimum eigenvalue* holds:

$$\min RQ = \lambda_1 \,, \tag{5.9}$$

where RQ is the Rayleigh quotient. The equivalent of the principle of the minimum eigenvalue can be translated in terms of graph theory and enables us to write

$$\begin{aligned} \bar{\lambda}_1 &= \mathrm{vol}(G)\, \inf_f \frac{\sum_{i\sim j}(f(i)-f(j))^2}{\sum_{i\sim j}(f(i)-f(j))^2 d(i)d(j)} \,, \\ \bar{\lambda}_{n-1} &= \sup_f RQ_G[f] \,, \end{aligned} \tag{5.10}$$

where f runs through the orthogonal complement of the vector with components $\big(d(1), d(2), \ldots, d(|G|)\big)$, and where $0 < \bar{\lambda}_1 \leq \bar{\lambda}_2, \ldots, \leq \bar{\lambda}_n$ are the eigenvalues of the normalized Laplacian $\mathcal{L}$.

We mention that such calculations become useful for example in conformal field theory [176], where one is interested in calculating the partition function as an integral of a negative exponential of the Hamiltonian over a set of state functions f defined on the graph with prescribed values σ on the node boundary of a subgraph S of a larger lattice. We have

$$Z(\sigma) = \int_f \mathrm{e}^{-cH[f,S]} \,,$$

where c is a positive constant, $H[f, S] = \langle f, \mathcal{L}f \rangle$, $\mathcal{L}$ is the normalized Laplacian of the lattice graph, the scalar product is taken on the nodes of S, and the integration is taken over all the functions f having the value σ on δS.

5.3.1 Relations Between Eigenvalues and the Diameter

In order to find the exact relations between the graph spectrum and its diameter, we need an exact definition for the diameter. Definition 7 is the traditional one: the maximum over all ordered pairs (i, j) of the shortest path from i to j [172–176]. For example, in [177], the authors proposed a definition based on the average distance in a graph, namely the length of the shortest path between two nodes averaged over all ordered pairs (i, j). The difficulty here is that, according to this second definition, a single disconnected pair has an infinite average distance which completely changes the results. Moreover, as emphasized in [178], a large network like the web is rife with such pairs, so discarding a few outliers before taking this average would not help the situation. In order to avoid such problems, recent literature uses a revised definition where either the max or sup operators, or the average, are taken only over pairs of connected nodes. Obviously, in the case of a connected graph there will be no need for such a special amendment.

For a finite and connected graph where the diameter can be calculated exactly, it has been shown that the value of the diameter is strongly related to its Laplacian spectrum, and there are a number of interesting bounding relationships between the diameter of a graph and the smallest and largest Laplacian eigenvalues [174, 175]. The general findings prove that the diameter cannot be smaller than a limit imposed by the smallest eigenvalue of its normalized Laplacian and by the graph volume [175]:

$$\mathrm{dia}(G) \geq \frac{1}{\bar{\lambda}_1 \mathrm{vol}(G)} \, .$$

The lower bound shown above can also be re-expressed in terms of the second smallest eigenvalue of the Laplacian:

$$\mathrm{dia}(G) \geq \left[\frac{4}{|G| \lambda_2} \right] \, , \tag{5.11}$$

where the square bracket in this equation and the next one represents the integer part. Among the relations providing upper bounds for the diameter of a graph [176], we have the Alon–Milman inequality which relates the diameter to the smallest Laplacian eigenvalue:

$$\mathrm{dia}(G) \leq \left[\sqrt{\frac{2 d_{\max}}{\lambda_2} \log_2 |G|} \right] + 1 \, . \tag{5.12}$$

If G is not a complete graph, then

$$\mathrm{dia}(G) \leq 1 + \frac{\ln(|G|-1)}{\ln \dfrac{\lambda_n + \lambda_2}{\lambda_n - \lambda_2}} \,. \tag{5.13}$$

Let G be an undirected k-regular graph having all its A eigenvalues smaller in magnitude that a certain eigenvalue of the Laplacian, i.e., $|\theta_i| < \lambda$, for all $i = 1, 2, \ldots, n$. Then the diameter is bounded above by the relation

$$\mathrm{dia}(G) \leq \frac{\ln(|G|-1)}{\ln(k/\lambda)} \,. \tag{5.14}$$

Finally, for a connected graph G with $\{\lambda_i, m_i\}_{i=1,\ldots,\sharp}$ distinct Laplacian eigenvalues together with their multiplicities, we have [173]

$$\mathrm{dia}(G) \leq |G| - 1 - \sum_{i=1}^{\sharp} (m_i - 1) \,.$$

This relation shows that graphs with a large diameter have Laplacian eigenvalues with small multiplicities, i.e., the majority of the Laplacian eigenvalues are distinct. In the following paragraph, we introduce diameter bounds for special types of graphs. A Moore graph is the regular graph with the maximal number of nodes for a given diameter and given degree. For Moore graphs we have $\mathrm{girth}(G) = 2\,\mathrm{dia}(G) + 1$ [174]. For a generalized polygon graph, we have $\mathrm{girth}(G) = 2\mathrm{dia}(G)$.

A very particular graph, having the smallest diameter for a given number of nodes and degrees, is shown in Fig. 5.9. This is a Kautz graph, K_d^{n+1}, of degree m and dimension $|G| = n$ [176]. It has $(n+1)m^n$ nodes and $(d+1)d^{n+1}$ edges. It is a directed graph, constructed from a finite alphabet with $m+1$ symbols, whose nodes

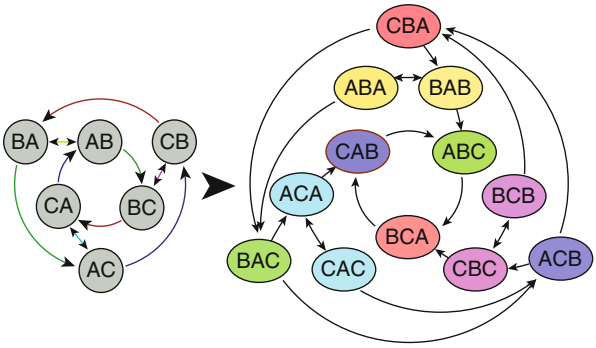


Fig. 5.9 Example of a Kautz graph

are all possible words of length $n + 1$ symbols written with this alphabet, without adjacent repetitions of symbols in the same word.

5.3.2 Relations Between Eigenvalues and Connectivity

The connectivity of a graph is related to its Laplacian spectrum, too. The second smallest eigenvalue λ_2 of the Laplacian matrix for a graph is called its *algebraic connectivity* and denoted by $a(G)$. The number of connected components of a graph is equal to the multiplicity of its $\lambda = 0$ null Laplacian eigenvalue. For a regular graph of degree k, the multiplicity of its largest Laplacian eigenvalue gives the number of its connected components [173–176].

5.3.3 Relations Between Eigenvalues and the Topology of a Graph

A graph is almost always finite. The counterexample would be provided by infinite networks, which are in fact no longer fictional theoretical examples, especially in light of the last counting of Facebook social Internet documents and registered users: 1.1–1.3 billion in 2013! Nevertheless, it is a little strange to study the topology of a graph. Such a study can be done in terms of connectivity, as we presented above, or in terms of excision or removal theorems. A graph cannot have holes of the kind we have in traditional homotopy and homology theory, but we can remove parts of it, and check how this surgery changes the graph spectra. Through such procedures, we can identify what limitations are introduced by the graph spectra with regard to how large, or how separated, etc., such subsets can be.

A bipartite graph definitely has a specific type of topology. The criterion for having a bipartite graph is to arrange for its Laplacian and its signless Laplacian to have identical spectra, and conversely [175]. Furthermore, for a bipartite graph we have $\theta_1 + \theta_n = 0$.

If a graph is k-regular then the second largest eigenvalue of the incidence matrix A satisfies the Alon–Boppana formula [173]

$$\theta_2 \geq 2\sqrt{k-1}\left[1 - \mathcal{O}\left(\frac{\ln(k-1)}{\ln k}\right)\right].$$

There is a strong result concerning the maximal size of disconnected subsets of a given graph [173]. Let us choose two disjoint subsets of nodes of a graph, $X, Y \subset N(G)$, $X \bigcap Y = \emptyset$, such that they are completely separated (there is no edge

between their respective elements). Then we have

$$\frac{|X|\,|Y|}{(|G|-|X|)(|G|-|Y|)} \leq \left(\frac{\lambda_n - \lambda_2}{\lambda_n + \lambda_2}\right)^2 , \tag{5.15}$$

where $n = |G|$, and of course λ_2, λ_n are the smallest nonzero and largest eigenvalues of the Laplacian spectrum, respectively. This theorem explains that the sizes of two disjoint sets of nodes (disjoint and not being neighbors of one another) is limited by the spacing of the Laplacian spectrum. The left-hand side of the in equation (5.15) is a strictly increasing function of the size of each of the subsets, so the larger the right-hand bound, the larger the separated sets that can be selected inside the graph. It is known from the Zhang–Luo–Anderson–Morley theorem that [173, 176, 179]

$$\frac{|G|}{|G|-1} d_{\max} \leq \lambda_n \leq \min(|G|, 2d_{\max}) ,$$

and

$$\frac{4}{\mathrm{dia}(G) \cdot n} \leq \lambda_2 \leq \frac{|G|}{|G|-1} d_{\min} .$$

A contour plot of the right-hand term of the inequality (5.15) reveals that, for large enough n (e.g., $n > 50$), the range of the maximal degree and the diameter of G are not much relevant, while the minimum degree of the graph drastically changes the ranges of the right-hand term. A lower value for the minimum degree increases the value of the right-hand term towards its maximum value of 1, while higher values of the minimum degree in a graph allow much lower values for the right-hand term. Consequently, graphs with higher values for their minimum (hence average) degree are less likely to be partitioned into disjoint and separated subsets than those with lower values for the minimum degree. Figure 5.10 exemplifies a separation analysis for a graph with 100 nodes.

5.3.4 Relations Between Eigenvalues and Paths

A *path* is a connected non-repeating subset of edges of a graph such that the intersection of any two of these edges is either empty or a common node, and the intersection of three edges is the empty set. A *walk* is a parameterized connected subset of edges of a graph. Both paths and walks can be open or closed, and in the latter case, the path is called a cycle.

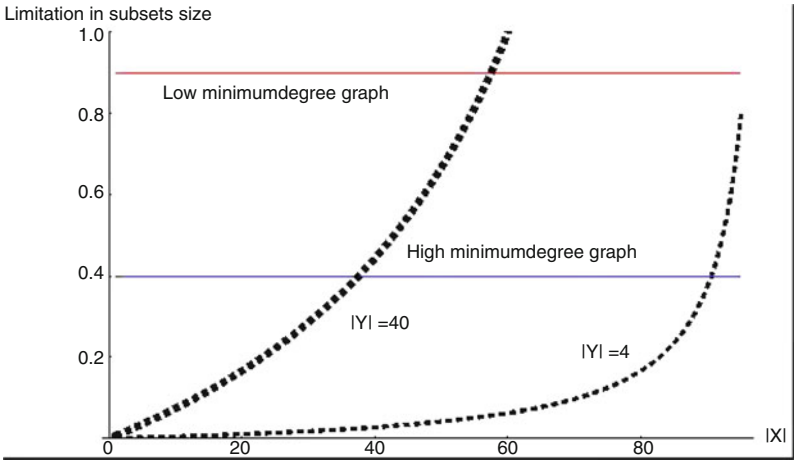


Fig. 5.10 Plot of the left-hand term in the inequality (5.15) versus the size of a subset X disjoint and separated from another fixed size subset Y of the same 100 node graph. For the *thin dashed curve*, the subset Y has 4 nodes, while in the *thicker dashed curve*, Y has 40 nodes. For a given X, Y, the *abscissa* of the plotted curves represent the admissible $\lambda_{2,n}$ eigenvalue limitation. For example, a highly connected graph with minimum degree 40 has, for the right-hand side of the inequality, the limit indicated by the *blue horizontal line*, hence showing that the maximum size of the set X is around 38 for $|Y| = 40$. A graph with minimum degree 4 involves the limit indicated by the *red line*, hence showing that the maximum size of the set X is around 76 for the same $|Y| = 40$. The conclusion is that it is harder to build large separated sets in a highly connected graph

The cardinality of the set P_k of independent closed walks of length k in G is given by

$$\text{card}\{P_k\} = \sum_{i=1}^{|G|} (\theta_i)^k \ ,$$

also called the k th spectral moment of G. For example, there are $|G| = n$ closed walks of length zero, $\|G\|/2$ closed paths of order 1, $|G|\langle d\rangle$ closed walks of length 2, etc.

5.3.5 *Other Relations Between Eigenvalues*

This is a summary of useful inequalities and relationships between eigenvalues of a graph. For example, it is known that the second smallest eigenvalue of the Laplacian, λ_2, is (the algebraic connectivity) superadditive with respect to unions of graphs. For a k-regular graph, its degree k is equal to the largest eigenvalue of the matrix A.

Otherwise, we have [173]

$$k_{\min} \leq \langle d \rangle \leq \max(\lambda_A) < k_{\max} \, .$$

Furthermore, $\lambda_n \leq \mu_1$, and equality holds for bipartite graphs. If a graph is connected, the largest A eigenvalue θ_1 has multiplicity 1. This eigenvalue itself measures the average degree $\langle d \rangle$ of the graph:

$$\max\left(\langle d \rangle, \sqrt{d_{\max}}\right) \leq \theta_1 = \theta_{\max} \leq d_{\max} \, ,$$

and therefore we have $\max|\theta_i| \leq \max_{i \in N(G)} d(i)$.

The number $\theta_1 = \theta_{\max}$ is called the *degree* of G, and we have

$$\langle d \rangle \leq \theta_1 = \theta_{\max} \leq \max d \, .$$

For a bipartite graph, we have $\theta_1 = -\theta_n$, and for a regular graph, we have $\theta_n = n$.

There is a very important theorem on the Laplacian spectrum related to reduction or growing of a graph. Consider G with L-spectrum λ_i. Then if $e \in E(G)$, consider the smaller subgraph $G' = G - \{e\}$ with its L'-spectrum λ'_i. We have

$$0 = \lambda'_1 = \lambda_1 \leq \lambda'_2 \leq \lambda_2 \leq \ldots \lambda'_n \leq \lambda_n \, , \tag{5.16}$$

which means that the roots of the two corresponding characteristic polynomials for matrices L and L' alternate, in exactly the same way as the zeros of orthogonal polynomials of different orders, or zeros of derivatives of functions.

The spectrum $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_n$ of the adjacency matrix A provides a number of important inequalities between the number of edges and the degrees of the nodes [174]:

$$\theta_1 \geq \frac{1}{\|G\|} \sum_{i \sim j} \sqrt{d(i)d(j)} \, , \tag{5.17}$$

$$\theta_1 \geq \sqrt{\frac{\|G\|}{\sum_{ij \in E(G)} 1/d(i)d(j)}} \, , \tag{5.18}$$

$$\theta_1 \geq \sqrt{\frac{\sum_{i=1}^{n} \left[d(i)\right]^2}{|G|}} \, , \tag{5.19}$$

$$\theta_1 \leq \frac{1}{2}\left(\sqrt{8\|G\| + 1} - 1\right) \, , \tag{5.20}$$

$$\theta_1 \leq \max_{ij \in E(G)} \sqrt{d(i)d(j)} \, . \tag{5.21}$$

We also have the inequalities [175]

$$\sum_{i=1}^{n} \bar{\lambda}_i \leq |G| \,, \quad \bar{\lambda}_1 \leq \frac{|G|}{|G|-1} \,, \quad \bar{\lambda}_{n-1} > \frac{|G|}{|G|-1} \,,$$

where $0 < \bar{\lambda}_i$ are the eigenvalues of the normalized Laplacian $\mathcal{L}$, labeled in increasing order.

5.4 Graph Topology and Boundaries

The past thirty years have seen increasingly rapid advances in the field of topological methods in the study of graphs and networks. During these years a considerable amount of literature has been published on topological quantifiers for graphs, especially emphasizing the relationships between topology and spectra. Traditionally, the topology of a graph has been assessed by evaluating a certain number of topological quantifiers based on edges (connectedness) or on the distance between nodes (radius, eccentricity, diameter), or the path structure (average path length), not to mention cycles or other closed substructures (girth, circumference, clustering coefficients), separability of subgraphs (isoperimetric number, expanders, Colin de Verdière number, Cheeger constant), boundaries, embeddability, etc.

In spite of there being so many topological tools, there are still difficulties with graph and network topology when attempting to implement results from traditional continuous topology, homotopy, and homology. Such limitations arise from the fact that one can apply topological concepts on different levels when modeling or attempting to understand discrete structures. For example, there is a 'data level' representing the way data initiated at a node is actually forwarded to other nodes. Topologies based on the data level reflect physical nodes and connections, but these topology graphs are hard to create and even harder to validate for correctness and completion. There is also a 'control level' representing policies for sending data, and this topology is based rather on the structure of edges than nodes. Different types of graph topologies have become a well defined area of research.

5.4.1 The Graph Topology and the Diameter

The difference between this section and Sect. 5.3.1 consists in the criteria used to evaluate the diameter. Here we wish to relate the diameter to the possibility of separating disjoint subsets of the graph, which is rather a topological than an algebraic property. All these inequalities are upper bounds for the diameter and all depend on the logarithm of the size $|G|$ of the graph. This phenomenon has been known for a long time [180–182] from statistical studies on large-scale networks or

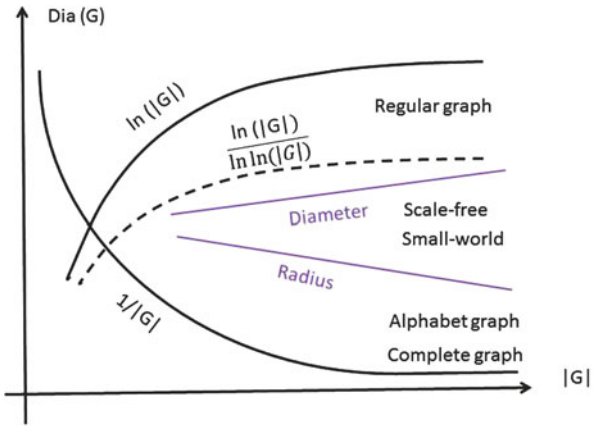


Fig. 5.11 The bounds of the radius and diameter of a graph versus its number of nodes $|G|$

from theoretical studies on random graph models. For example, in [183], the same behavior was noticed on a very large regular graph. In [184], the authors proved that, even for regular networks with just a little bit of randomness, the logarithmic upper bound of the diameter is a stable phenomenon.

Figure 5.11 presents a synthetic view based on various theorems concerning the graph diameter with upper and lower bounds. From the large number of available studies for small connected graphs, one can conclude that the diameter is bounded below by a constant divided by $|G|$, and bounded above by another constant times the natural logarithm of $|G|$. There is, however, another result [180, 185] arising form large computer simulations performed on some special graphs with large sizes (random networks), in which the upper bound for the diameter is given by $\log|G|/\log(\log|G|)$.

When the size of a graph increases, by increasing its number of nodes and edges, the diameter value is bounded between the lower hyperbolic limit and the upper logarithmic (loglog, normalized logarithmic) limit. Complete graphs and alphabetic combinatorial graphs (like the Kautz graph) lie around the lowest limit for the diameter, while scale-free networks approach the upper logarithmic limit. If we could interpret the number of nodes to be proportional to the 'area' or even the volume of the graph, the upper bound logarithmic dependence suggests an exponential area dependence on the radius: $\mathrm{dia}(G) \leq \ln|G| \to |G| \sim \mathrm{e}^{\mathrm{dia}(G)}$. Even in a very high number of dimensions, there is no such geometrical dependence. This exponential increase in the 'area' of the graph with its diameter rather suggests a combinatorial or even multi-fractal type of geometry.

We begin our analysis from the Van Dam–Haemers relation relating the diameter to the two largest Laplacian eigenvalues [174, 175]:

$$\mathrm{dia}(G) \leq \left\lceil \frac{\log_2(|G|-1)}{\log_2(\sqrt{\lambda_n}+\sqrt{\lambda_{n-1}}) - \log_2(\sqrt{\lambda_n}-\sqrt{\lambda_{n-1}})} \right\rceil . \tag{5.22}$$

A similar inequality is provided by the Mohar inequality, which involves another topological invariant of a graph:

$$\mathrm{dia}(G) \leq 2\frac{\log_2 \frac{|G|}{2}}{\log_2 \frac{d_{\max}+h_G}{d_{\max}-h_G}} . \tag{5.23}$$

Here h_G is the Cheeger constant, also known as the *isoperimetric constant*. We will introduce this constant later on [see (5.25) in Sect. 5.4]. In [175], the authors obtained another type of upper bound for the diameter. In terms of the Laplacian eigenvalue, it reads

$$\mathrm{dia}(G) \leq \left\lceil \frac{\ln(|G|-1)}{\ln \frac{1}{\bar{\lambda}_1 - 1}} \right\rceil ,$$

while in terms of the normalized Laplacian, it reads

$$\mathrm{dia}(G) \leq \left\lceil \frac{\cosh^{-1}(|G|-1)}{\cosh^{-1}\left[(\bar{\lambda}_{|G|-1}+\bar{\lambda}_1)(\bar{\lambda}_{|G|-1}-\bar{\lambda}_1)^{-1}\right]} \right\rceil ,$$

which is somehow understandable since inverse hyperbolic functions are actually logarithms. This last inequality has important implications for developing more general topological quantifiers of separation. If we choose two disjoint subgraphs $X, Y \subset N(G)$, we can generalize the concept of distance from two nodes to two sets:

$$d(X,Y) = \min\left\{d(x,y)|x \in X, y \in Y\right\} .$$

In the study in [175], it was shown that, with the notation $\bar{X} = N(G) - X$ for the complementary set, one can write an upper bound for the separation distance between the sets:

$$d(X,Y) \leq \left\lceil \frac{\cosh^{-1}\sqrt{\frac{\mathrm{vol}(\bar{X})\,\mathrm{vol}(\bar{Y})}{\mathrm{vol}(X)\,\mathrm{vol}(Y)}}}{\cosh^{-1}\left[(\bar{\lambda}_{|G|-1}+\bar{\lambda}_1)(\bar{\lambda}_{|G|-1}-\bar{\lambda}_1)^{-1}\right]} \right\rceil .$$

One can generate various formulas measuring the separation between two disjoint subsets X, Y of a graph in terms of the spectrum of its (normalized) Laplacian, variations being induced by choosing different measures on the subsets [175]. All such definitions obey a certain pattern, generally in the form $\bar{\lambda}_1 \leq F\big(\mu(X), \mu(Y), d(X,Y)\big)$, where F is a continuous decreasing function in all three

arguments. More interestingly, if the distance between the sets can be explicitly solved from this inequality, one obtains an upper bound for the distance between any two disjoint subsets of a graph in terms of their measures and the first (smallest) eigenvalue of the normalized Laplacian matrix. In general, this upper bound has a logarithmic decreasing dependence on the measure of X and Y if they are 'small' compared to G, and it is rather constant and proportional to $1/\sqrt{\bar{\lambda}_1}$ if the sets are the 'largest' possible, that is, if they grow up to 'half of G'.

If the smallest eigenvalue of the normalized Laplacian is arbitrarily close to zero, there is no strict limitation on how large and disjointed the sets can be that we cut out of the graph. However, if this eigenvalue is bound from below and away from zero, such sets are limited in size. This is the case for small diameter graphs that also have large edge or node expansions. Explicit examples are random graphs, or k-regular random graphs where the diameter can be approximated by $\mathrm{dia}(G) \sim \log_{k-1}|G| + \log_{k-1}\ln|G|$.

5.4.2 Embeddings

For a loopless graph G (i.e., $i \nsim j$ for any pair of nodes), we define a *graph embedding* by mapping the nodes one-to-one into points in $\mathbb{R}^n$, and by having an injective map of the edges into continuous curves in $\mathbb{R}^n$, intersecting only at the images of the graph nodes. Obviously, any finite graph can be embedded in $\mathbb{R}^3$ by mapping its nodes in arbitrary order along a line segment called the node axis, and connecting the corresponding nodes with edges such that each edge is embedded into another Euclidean plane from a family of planes mutually intersecting in the node axis, as in Fig. 5.12. Of course, some non-intersecting edges can be mapped into the same plane.

The simplest type of graphs from the embedding classification point of view is the *outerplanar* graph. A graph is outerplanar if it can be drawn in the plane without crossings, in such a way that all of the nodes belong to the boundary of the drawing, that is, no vertex is totally surrounded by edges. Put more simply, it can be drawn without intersecting edges in a plane in such a way that all nodes can be placed on the circle, and all edges are inside this circle. The outerplanar graphs are related to electrical circuit design, and especially to series–parallel graphs.

An even more complex graph which is very important for applications is the *planar graph*, i.e., one that can be embedded in the Euclidean plane. The planar graph and the so-called planarization problem are studied in relation to the design of printed circuit boards and the routing of very large scale integration circuits. There are sophisticated algorithms for finding the near-maximal planar subgraph from a given in general non-planar graph.

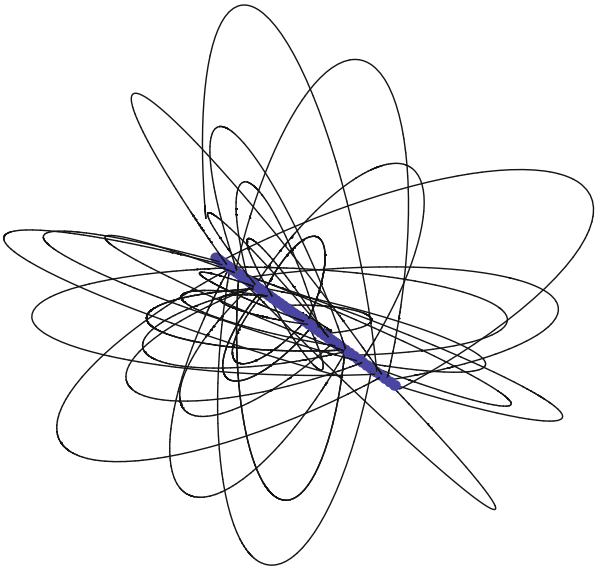


Fig. 5.12 Embedding a graph in $\mathbb{R}^3$ as a 'book'. The nodes are mapped along a real segment like a book spine, and each edge is embedded into a different page of the book. Sometimes non-intersecting edges that form locally planar subgraphs are mapped onto the same page

A more complex and particular type of graph is the *linklessly embeddable* graph. This has the property that no pair of its cycles (that is, their images in $\mathbb{R}^n$) are linked.[1]

The embedding properties (like being planar, outerplanar, linklessly embeddable, or embeddable in some type of surface) are closed under taking *graph minors* (Robertson theorem [173]). A minor version of a given graph is the smaller graph obtained by removing isolated nodes, or removing edges and merging the two remaining nodes. The planarity of a graph can be easily checked by using graph minors and the Kuratowski–Wagner theorem [173]: a finite graph is planar if and only if all its minors include neither the complete graph on five vertices, nor a complete bipartite graph on six vertices (see Fig. 5.13).

One of the most important exact results on the question of classification of graphs according to the simplest manifold in which they can be embedded concerns the Colin de Verdière parameter $\mu(G)$ [173, 186]. This parameter can classify graphs according to their embedding properties. More surprisingly, it has deep implications in Riemannian geometry problems like the spectrum of the Laplace–Beltrami operator for Dirichlet problems, Cheng's eigenvalue comparison theorem, or even in physics problems related to mesoscopic type II superconductivity, the spectrum of the magnetic Schrödinger operator, and solutions of Diophantine equations.

[1] Two disjoint Jordan curves are linked in $\mathbb{R}^3$ if there is no sphere S^2 separating them.

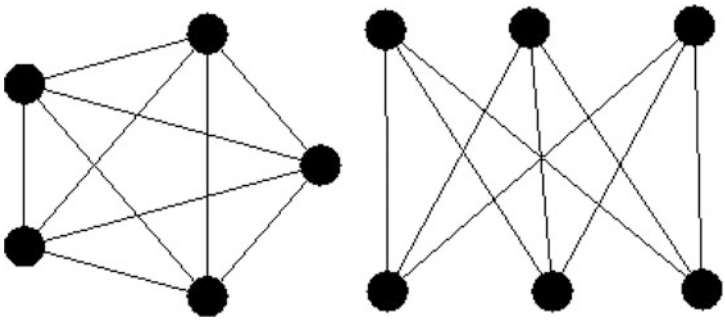


Fig. 5.13 The complete graph with 5 nodes (K_5) on the left, and the 'utility' graph with 6 nodes on the right are the only graphs preventing a larger graph to be planar if it contains one of them as minors

As mentioned above, the introduction of this parameter was motivated by the study of the maximum multiplicity of the second eigenvalue of certain magnetic Schrödinger operators. A major problem in applications involving these operators is that they need to be defined on Riemann surfaces because the magnetic Schrödinger equation very well describes the Landau–Ginzburg model for superconductivity, and the superconducting materials come in compact objects like disks, pellets, spheres, cylinders, etc. It turned out that in the study of these operators one can approximate the surface by a sufficiently densely embedded graph G, in such a way that $\mu(G)$ is the maximum multiplicity of the second eigenvalue of the operator, or a lower bound to it. Yet again, we see the importance of the second (largest) eigenvalue of a discrete bounded spectrum.

The parameter $\mu(G)$ can be described fully in terms of properties of matrices related to G, mainly the adjacency matrix A, its spectrum, and in particular its second largest eigenvalue θ_2. Even if the $\mu(G)$ parameter is defined on the basis of purely linear algebra, explicit calculation for a given graph involves a tedious and as yet unsatisfactory amount of work.

For a given integer N, consider the set of symmetric $N \times N$ matrices with real entries, viz., $M = (M_{ij}) \in \mathbb{R}^{N\times N}$. We can always regard such a matrix as a linear operator acting on vectors from $\mathbb{R}^N$ and it is easy to identify the null space $\ker(M)$ of any such matrix, i.e., the vector space of vectors mapped to the null element by this operator. The linear dimension of this null space $\dim\big(\ker(M)\big)$ is called the *co-rank*. Given a graph $G(N, E)$, we can construct a set of special matrices M by a sort of generalization of the adjacency matrix obeying three rules.

The first rule requires M to have the same structure as the negative adjacency matrix $-A$ for the graph G, except that it can have any elements (even nonzero) on the diagonal, and the 1 entries are generalized to any arbitrary negative numbers. In other words the off-diagonal elements of these matrices (M_{ij}) are simply negative numbers if $i \sim j$ in G, and zero otherwise. There is no specification regarding its diagonal elements. The second rule requires M to have only one negative eigenvalue of multiplicity 1. Here is an example showing how easily this works: choose $N = 3$

and G as the graph with $1 \sim 2$, $2 \sim 3$. We can easily find a matrix M satisfying the first two rules above:

$$M = \begin{pmatrix} 0 & -1 & -2 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}.$$

This matrix also has one eigenvalue 1, and the other two $(1 \pm \sqrt{21})/2$, so only one eigenvalue is negative and has multiplicity 1. However, there is a third rule which minimizes the possibilities for M, namely, that it should satisfy the *strong Arnold property* [187]: any square matrix X of the same dimension $N \times N$, with zero diagonal, having zero entries in every place where M has nonzero entries, and in addition such that $MX = 0$ is zero, must itself be zero, viz., $X = 0$. In other words, the kernel (the null space) of M has zero intersection with any matrix X orthogonal to M in the direct product sense. Continuing with our example, we can find a matrix

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & x \\ 0 & y & 0 \end{pmatrix}$$

which is constructed according to the above prescriptions, except that, in order to satisfy $MX = 0$, we need to set $x = y = 0$.

Finally, let us pretend we have found for a given graph several examples of such matrices M satisfying all three rules. The matrix with the largest co-rank among all these matrices M provides the constant $\mu(G)$, and this constant is its co-rank. In our 3×3 example presented above, the co-rank is 1.

The Colin de Verdière number of a graph has several surprising graph-theoretical properties. The most basic is that $\mu(G)$ is minor-monotone, i.e., if H is a minor of G, then $\mu(H) \leq \mu(G)$. Furthermore, the number is subadditive, that is, if G, H are two graphs, then $\mu(G + H) = \max\{\mu(G), \mu(H)\}$. For a complete graph of order k this number is $k - 1$. If its value is less than 1, then the graph must be a disjoint union of paths. If it is less than 2, the graph is outerplanar, and if it is less than 3, the graph must be planar. Finally, if the number is less than 4, the graph is linklessly embeddable [186].

5.4.3 *Isoperimetric Problems*

The typical isoperimetric problem in planar geometry, especially when we need to find the maximum area for a given perimeter, can be translated into graph theory as the problem of finding the maximum $|G|$ in a set with a boundary of prescribed cardinality.

For any two disjoint subsets of nodes of a graph, $S, T \subset N(G)$, $S \cap T = \varnothing$, we define the set of the edges connecting them by

$$E(S,T) = \left\{(ij) \in E(G) | i \in S, j \in T\right\} .$$

We define the *edge boundary* of a subset of nodes S by $\partial S = E(S, N(G) \setminus S)$, that is, the set of edges emerging from S towards nodes that are not in S.

We define the *isoperimetric number* (or *conductance*) of a graph

$$i(G) = \min_{0<|S|\leq|G|/2} \frac{|\partial S|}{|S|} . \tag{5.24}$$

The isoperimetric number has the following properties:

$$i(G) \leq \frac{2\|G\| \left[|G|/2\right]}{|G|(|G|-1)} ,$$

$$i(G) \leq \frac{a(G)}{2} ,$$

$$i(G) = \min \frac{\boldsymbol{x}^{\mathrm{t}} L \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{t}} \boldsymbol{x}} ,$$

where the minimum is taken in the set of all vectors $\boldsymbol{x} \in \{0,1\}^{|G|}$ such that $1 \leq \|\boldsymbol{x}\|^2 \leq |G|/2$.

Here we present the point of view developed in [175] on isoperimetric problems. The authors of this work define a *cut* in a graph as a subset of the set of edges which disconnects G, i.e., which make $N(G)$ into a disconnected set. There are two types of cut: *edge cuts* and *node cuts*, and consequently, there are two types of isoperimetric problem for the two types of boundary measure. Both cuts deal with the question of finding the subgraph of given volume (chosen to be less than the volume of its complement, in order to avoid redundancy) which guarantees the minimum measure for its boundary.

The *edge cut* around a subgraph $S \in N(G)$ is made through the *edge boundary* for S, namely $\partial S = \{(ij) | i \sim j, i \in S, j \neq S\}$. If we denote the complement of S with respect to the graph by $\bar{S} = N(G) - S$, we have $\partial S = \partial \bar{S} = E(S, \bar{S})$. The measure of this type of boundary can be expressed in two different ways:

$$h_G(S) = \frac{|E(S,\bar{S})|}{\min\left\{\mathrm{vol}(S), \mathrm{vol}(\bar{S})\right\}} , \qquad h'_G(S) = \frac{|E(S,\bar{S})|}{\min\left\{|S|, |\bar{S}|\right\}} . \tag{5.25}$$

The advantage of these definitions is that they allow the introduction of an overall characterization of any graph (and hence its spectrum) in the context of isoperimetric problems. We have

$$h_G = \min_S h_G(S) \ , \qquad h'_G = \min_S h'_G(S) \ , \tag{5.26}$$

where both minima are taken over all subsets S of node numbers containing no more than half of the nodes of the original graph, and the numbers h_G, h'_G are called the Cheeger constant(s) of G [175]. The difference between prime (also called modified Cheeger constants) and non-prime constants consists in the way one decides to normalize the boundary: versus the volume which implies the degree, or only versus the number of nodes. This also involves the use of either the Laplacian or the normalized Laplacian, respectively. As mentioned above, this constant bears some relationships with the graph eigenvalues [175]:

$$2h_G \geq \bar{\lambda}_1 \ , \qquad \bar{\lambda}_1 \geq \frac{h_G^2}{2} \ , \qquad \bar{\lambda}_1 > 1 - \sqrt{1 - h_G^2} \ ,$$

but

$$2h'_G \geq \lambda_2 \ ,$$

because the case of the modified Cheeger constant involves Laplacian eigenvalues, and $\lambda_1 = 0$.

The *node cut* is built with the help of the *node boundary* for S, that is, $\delta S = \{i \neq N(S) | ji \in E(G), j \in N(S)\}$, which represents the set of all nodes which are not in S, but are neighbors of S. Similarly with the Cheeger constant, one can define two equivalent parameters for the node boundary:

$$g_G(S) = \frac{\mathrm{vol}(\delta S)}{\min\left\{\mathrm{vol}(S), \mathrm{vol}(\bar{S})\right\}} \ , \qquad g'_G(S) = \frac{|\delta S|}{\min\left\{|S|, |(\bar{S})|\right\}} \ , \tag{5.27}$$

with the corresponding parameters defined independently of the test cut S:

$$g_G = \min_S g_G(S) \ , \qquad g'_G = \min_S g'_G(S) \ . \tag{5.28}$$

Again, the only difference between prime and non-prime definitions consists, as above, in the way one decides to normalize the measure of the boundary: versus degree and nodes, or just versus nodes, respectively. These two parameters bear their own relationships with the graph spectrum [175]:

$$g_G \geq h_G \ , \qquad 2g_G \geq \bar{\lambda}_1 \ , \qquad \bar{\lambda}_1 > \frac{g_G^2}{4d_{\max} + 2d_{\max} g_G^2} \ .$$

5.4.4 Separations

The problem of *separation* is to find the most efficient cut in a graph, i.e., to identify the minimum set S which can generate a cut in the graph through its maximal boundary. The concept and subsequent formulas are very useful for evaluating graphs and networks. Another possible question: what is the maximum set of cuts that completely separate the network into two disjoint sub-networks? The edge boundary has the property [174]

$$\lambda_2 \frac{|S|\,|N(G) \setminus S|}{|G|} \leq |\partial S| \leq \lambda_n \frac{|S|\,|N(G) \setminus S|}{|G|} \,, \tag{5.29}$$

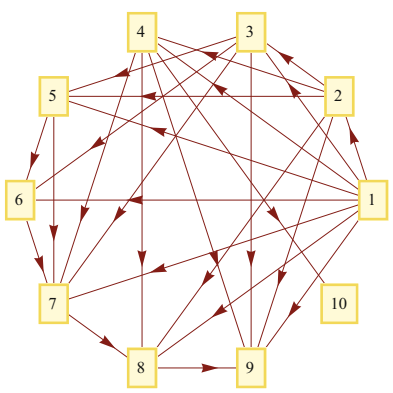
where λ are the eigenvalues of the Laplace operator. Equation (5.29) expresses the fact that the span $\lambda_n - \lambda_2$ of the Laplacian spectrum provides the admissible relative range for the size of the edge boundary with respect to the size of its inside S. For a given graph and a subgraph of given size $|S|$ (the size of $|N(G) - S|$ is prescribed, too), the boundary of the subgraph cannot be smaller or larger than a certain limit. We can rewrite (5.29) in the form

$$\lambda_2 \leq \frac{|\partial S|}{|S|} \frac{|G|}{|N(G) \setminus S|} \leq \lambda_n \,.$$

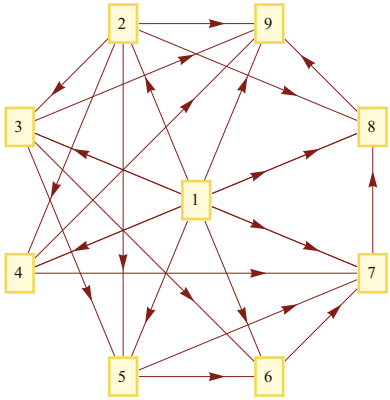
For graphs whose Laplacian spectrum has a narrow range, the edge boundary of any subgraph has almost constant size, which explains why algorithms based on cuts have good results on random graphs. Furthermore, there are upper and lower limits for the edge boundary size [174]:

$$\begin{aligned} &\max |\partial S| \leq \frac{|G|\lambda_n}{4} \,, \\ &\min_{|S|=[|G|/2]} |\partial S| \geq \begin{cases} \dfrac{|G|}{4}\lambda_2 & \text{if } |G| = \text{even} \,, \\ \dfrac{|G|^2 - 1}{4|G|}\lambda_2 & \text{if } |G| = \text{odd} \,, \end{cases} \end{aligned} \tag{5.30}$$

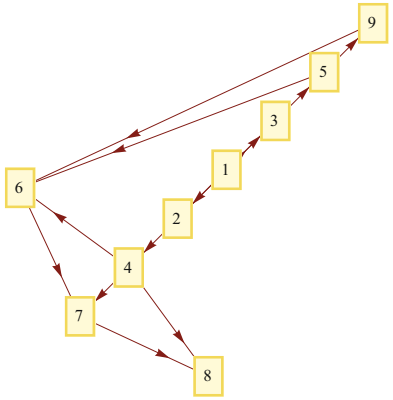
also referred to as the *bipartite width* of the graph. We present example applications of (5.29) and (5.30) for ∂S in Fig. 5.14. In the limiting cases of strongly or weakly connected graphs, the range provided by the formula for a given $|S| = 4$ subgraph does not change if we add an extra isolated node or change the shape of the graph. The upper limit for the size of the edge boundary is always dictated by the number of edges $\|G\|$. However, for a regular lattice type of graph, where the standard deviation of the average degree is small, the size of the edge boundary is mostly controlled by the node number $|G|$. It almost seems that the formulas in (5.29) and (5.30) are not really useful for practical purposes, because they give a very wide range. We



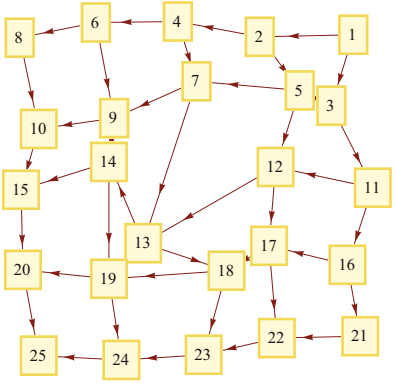
|G|=9, ||G||=25, <d>=5.7, |∂S|=7−26



|G|=10, ||G||=26, <d>=5.2, |∂S|=7−29



|G|=9, ||G||=12, <d>=2.7, |∂S|=2−14



|G|=25, ||G||=40, <d>=3.2, |∂S|=1−24

Fig. 5.14 Examples to evaluate the size of the edge boundary of a subgraph with $|S| = 4$, cut out from four different types of graphs: strongly connected (*upper left*), strongly connected with one extra weakly connected node (*upper right*), a weakly connected almost path graph (*bottom left*), and a square lattice graph (*bottom right*). Under each graph we indicate the nodes, edges, average degree, and range of boundary size

find that a strongly connected graph is more likely to accept a sophisticated (even fractal) shape for the boundary of a given cut (removal of a subgraph of fixed size) than other types of graphs. Put simply, regular graphs as opposed to random graphs, offer better optimized solutions for isoperimetric problems.

We define the *node cut* or *node boundary*: for a given subgraph S of G, δS is defined as the set of all nodes outside of S which are adjacent to S. We have the following relations between the two measures of the boundary of a subgraph [174]:

$$|\delta S| \leq |\partial S| \leq d_{\max}|\delta S| \,, \tag{5.31}$$

and

$$j(G) = \min_{1 \leq |S| \leq \lfloor |G|/2 \rfloor} \frac{|\delta S|}{|S|} \,. \tag{5.32}$$

The expression $j(G)$ is called the *node expansion* of a graph, and it bears the following relationship with the algebraic multiplicity: if $a(G) \geq \epsilon > 0$, then

$$j(G) \geq \frac{2\epsilon}{d_{\max} + 2\epsilon} \,,$$

and conversely, if $j(G) \geq c > 0$, then

$$a(G) \geq \frac{c^2}{4 + 2c^2} \,.$$

5.4.5 Expanders

A connected and undirected graph G is called an *expander* if, for any subset $S \subset N(G)$, the number of nodes in $G \setminus S$ placed at distances less than or equal to one hop from S (we denote this number of nodes by P) is greater than a constant times $|S|$ [188]. In other words G is an expander if there exists a positive constant ϵ such that $P \geq \epsilon |S|$: the larger the set, the larger the node boundary, and in a proportional way. Expanders are useful due to their role in procedures for sorting networks [173]. Figure 5.15 shows such an expander graph. The set $P \setminus S$ is called the *adjacent* set to S. We have the following (Tanner [173]) theorem: if G is a k-regular undirected graph so that there is an eigenvalue of the Laplacian which is an upper bound for the moduli of all adjacency matrix eigenvalues, i.e., $k \neq |\theta_i| \leq \lambda$ for any $i = 1, 2 \ldots, n$, then for any subset of nodes $R \subset N(G)$ having its adjacent set $\tilde{R} \subset N(G)$, we have

$$\frac{|\tilde{R}|}{|G|} \geq \frac{|R|/|G|}{\frac{|R|}{|G|} + \frac{\lambda^2}{k^2}\left(1 - \frac{|R|}{|G|}\right)} \,. \tag{5.33}$$

This inequality tells us that the size of any adjacency set is bounded from below, or that the boundary for any subset in such a graph is always large. Figure 5.16 plots the ratio of the size of the boundary $|\tilde{R}|$ to the size of the inside $|R|$ against the size of the inside, for different orders k of regular graphs and different eigenvalues of the Laplacian. When the size of the graph increases indefinitely, the ratio approaches unity, while for more than $n/2$ neighbors, the ratio is practically 1 for any size of graph.

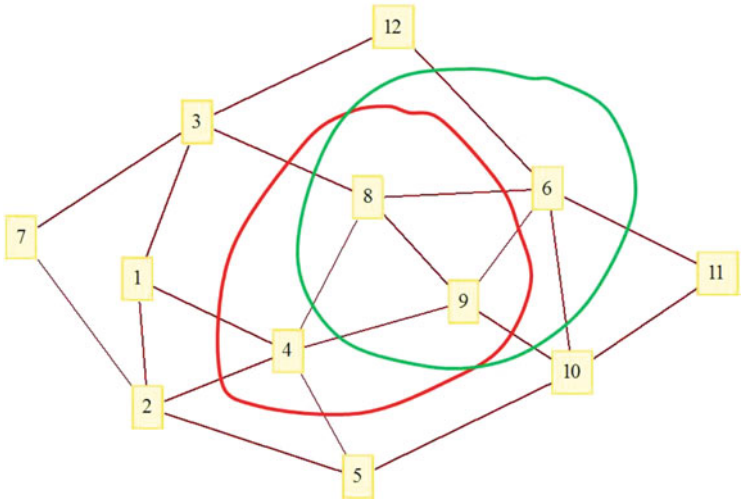


Fig. 5.15 An expander graph. If we consider the set of nodes $\{4, 8, 9\}$ encircled by the *red line*, the set P of nodes placed at distance one hop or less is $\{1, 2, 3, 5, 6, 10\} \cup S$, so $|P|/|S| = 3$. If we consider for S the set encircled with *green*, we have for the same ratio $8/3 = 2.6(6)$, and so on. It is straightforward to check that this ratio has a minimum value of $\epsilon = 1.(09)$

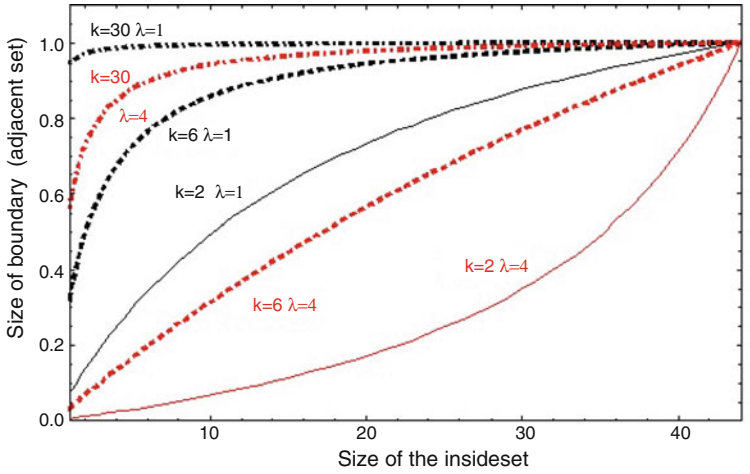


Fig. 5.16 Ratio of the size of the boundary to the size of the inside, plotted against the size of the inside, for three different orders k of regular graphs and two different eigenvalues of the Laplacian [see (5.33)]

5.5 Algebraic Topology

Algebraic topology is the branch of mathematics which serves as a bridge between discrete and continuous mathematics. A major strength of algebraic topology has always been its wide degree of applicability to other fields: physics, differential geometry, algebraic geometry, and number theory. Modern algebraic topology is the study of the global properties of spaces by means of algebra. Poincaré was the first to link the study of spaces to the study of algebra by means of the fundamental group.

In order to emphasize the way continuous mathematics, for example, differential geometry, interacts with algebraic topology, we can think of the example provided by a sphere. One way of telling that we live on a sphere is to measure the sum of the three angles of a triangle. For a small triangle, it is slightly more than 180°, but for a large triangle, it is much more. This non-constancy property of the sum of the angles of a triangle, and the proportionality relationship between the size of the triangle and the sum of its angles tells us that we live on a surface with (nonzero) positive curvature. But since we can use small triangles, this is a differential geometry property, because it is a local one. Algebraic topology is concerned with the whole surface and points to the obvious fact that the surface of a sphere is a finite area with no boundary, while the flat plane does not have this property. So far all these concepts are related to continuous mathematics, but there are also discrete mathematics ideas that characterize the difference between the plane and the sphere: the plane has no holes in it, while a sphere surrounds a three-dimensional hole. And since the number of holes in a geometric object is a discrete quantity, a theory like algebraic topology describing the number and properties of these holes must have a discrete counterpart.

Algebraic topology deals with the differences between the plane and the sphere by assigning groups. Indeed, it assigns special groups, with invariance properties against deformations, to any space. These groups are called homotopy and homology groups. The groups are invariant under homeomorphisms, that is, against the continuous deformation of space. The sphere is assigned an infinite group which is a measure of the fact that the sphere has a hole in it, while the plane is assigned the zero group because it does not. The fact that these groups are different tells us that the spaces are fundamentally different from a global vantage point. And of course, algebraic topology is not confined to the study spaces of dimension three, but includes the study of higher-dimensional spaces as well.

In algebraic geometry, the connection between discrete and continuous mathematics is obtained by combining finite numbers of smooth geometrical objects (simplices, cells, spheres, disks, planes, etc.), each having established topological and differential properties, in a finite-dimensional algebraic structure like a complex, or CW complex, and inducing an algebraic structure on the set of homeomorphisms between these geometric elements. For example, a torus T^1 embedded in $\mathbf{R}^3$ can be cut around its two circles (major and minor) and then mapped into a rectangle, providing an equivalence relation between its edges and corners. This rectangle is

further divided into a finite number of triangles (simplices), whose simple properties are very well known (this process is known as triangulation).

The fundamental construction in algebraic topology is the *simplex*. Consider a set of $p+1$ distinct points in an n-dimensional Euclidean space $\mathbb{E}^n$, denoted as affinely independent vertices $\{u_0, u_1, \ldots, u_p\}$, that is, the vectors $u_j - u_0, j = 1, 2, \ldots, p$ are linearly independent. The p-simplex generated by this set of points is

$$\sigma_p = [u_0, u_1, \ldots, u_p] = \left\{ \sum_{i=0}^{p} \lambda_i u_i \;\middle|\; u_i \geq 0, \sum_{i=0}^{p} \lambda_i = 1 \right\} \subset \mathbb{E}^n \,. \tag{5.34}$$

The convex hull of the vertices of a p-simplex is called a face of the simplex. Faces are $(p-1)$-simplices themselves, the 0-faces are the vertices, the 1-faces are called the edges, and the $(p-1)$-faces are called the facets. A line segment is a 1-simplex, a triangle in the plane is a 2-simplex, a tetrahedron is a 3-simplex, and so on.

Given a collection of simplices of different orders, we call it a *simplicial complex* $\mathcal{K}$ if the collection satisfies the conditions:

Definition 8 The union of a collection of simplices forms a simplicial complex $\mathcal{K}$ if and only if:

1. any face of a simplex from $\mathcal{K}$ is also in $\mathcal{K}$,
2. the intersection of any two simplices $\sigma_1, \sigma_2 \in \mathcal{K}$ is a face of both σ_1 and σ_2.

The set-theoretic union of all simplices from $\mathcal{K}$, when viewed as a subset of $\mathbb{R}^d$, is called a polytope. If a subset X of $\mathbb{R}^d$ is homeomorphic to a polytope, we say X is triangulated by $\mathcal{K}$. The procedure to construct a simplicial complex associated with (homeomorphic to) a given topological space is triangulation. We need to find the set of simplices which combine in a complex to describe the given space. The general procedure is to puncture the topological space and remove the point. Then, starting from that hole, apply a series of cuts in the original topological space and in the end flatten out the resulting surface over a plane. In order to respect the positions of all points before the cuts were performed, we have to label points which used to coincide before the cut with the same label. Then, we start to divide the resulting geometrical figure with lines and planes until all the edges and nodes of all the simplices satisfy the axioms of the simplicial complex (see Definition 8).

An example of triangulation of a 2-torus ($S^1 \times S^1$) is presented in Fig. 5.17. The four numbers 7 show that these corners first have to join together to form a cylinder, which is then bent and glued into a toroidal surface. We draw a diagonal (the one containing 5, 6), and in this way we divide the square into two 2-simplices which have a common edge, and also two common vertices, which is forbidden by Definition 8. The next step is to divide the square with lines, for example the edges 3–6 and 6–1, and continue this procedure until all the triangles we draw satisfy the axioms of the simplicial complex in Definition 8.

Note that the empty set is a face of every simplex. See also the definition of an abstract simplicial complex, which loosely speaking is a simplicial complex without an associated geometry. A simplicial complex together with its topology and

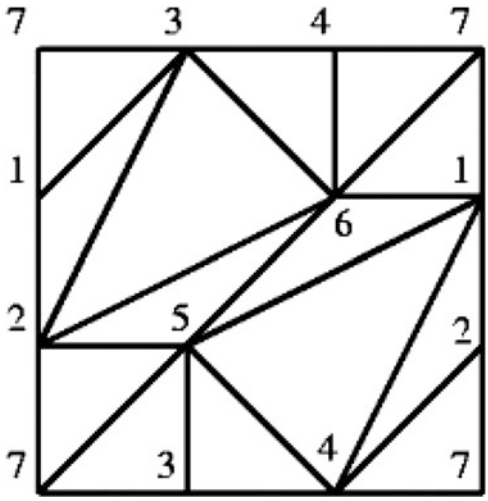


Fig. 5.17 Example of a triangulation of a torus, which needs 14 triangles to satisfy the simplicial complex axioms of Definition 8. The two initial cuts on the torus surface map it to a square whose corners were joined, and which are now all labeled 7. In order to satisfy all the requirements of the complex, we have to divide the square into more 2-simplices

closure properties can be considered a CW complex, although the latter is a more general and more abstract structure. More precisely, by a result due to Milnor, the geometric realization of a locally finite simplicial complex is a CW complex. The main difference, besides level of generality, is that a CW complex is not necessarily ordered, as the simplices in a simplicial complex are.

In the following, we present an inductive constructive definition of a CW complex:

1. Begin with a discrete space X^0 whose points are called zero-dimensional cells.
2. Suppose we have already constructed X^{n-1}. For every element j of an index set J_n, take a map $f_j : \partial D_j^n \to X^{n-1}$ and define

$$X^n = \bigcup_j (X^{n-1} \cup_{f_j} D_j^n) \ .$$

   The interiors of the disks D_j^n are the n-dimensional cells denoted by e_j^n.
3. We can stop the construction for some n and put $X = x^n$, or proceed to infinity and put

$$X = \bigcup_{n=0}^{\infty} X^n \ ,$$

   where in the latter case X is equipped with the inductive topology, which means any topological affirmation is true if it is valid for any of the topological subspaces X^n.

For example, if X^0 is a set of $2k$ points, $k \in \mathbb{Z}$, we introduce a set of 1D disks D_j^1, which are just a collection of segments, and map their end points, i.e., the ∂D_j^1, to the set X^0. In this way, any pair of points in X^0 has a line segment associated with it, which together with X^0 form X^1. The construction continues. The sphere S^n is a CW complex with one cell e^0 of dimension 0, one cell e^n of dimension n, and the constant map $f : S^{n-1} \to e^0$.

In any simplicial complex $\mathcal{K}$, or in any CW complex, we can introduce a group structure by defining the addition of p-simplices to form the chain group C_p. In the following, in order to present the contents efficiently and briefly, we demonstrate the concepts on simplicial complexes only. Its elements consist of p-chains, the linear combination of a finite number of oriented p-simplices $c_p = a_i \sigma_i^p$ with integer, rational, or real number coefficients a_i, a negative coefficient representing reversal of the orientation of the simplex.

Consider a simplicial complex $\mathcal{K}$ containing simplices of all orders up to n. We introduce the *boundary operator*

$$\partial_p : C_p \to C_{p-1} \ ,$$

acting on $\mathcal{K}$ with values in $\mathcal{K}$ as a linear operator that maps a p-simplex onto the oriented sum of all $(p-1)$-simplices in its boundary:

$$\begin{aligned} \partial_p [x_0, x_1, \ldots, x_p] = [x_1, x_2, \ldots, x_p] - [x_0, x_2, \ldots, x_p] \\ + \cdots + (-1)^p [x_0, x_1, \ldots, x_{p-1}] \ . \end{aligned}$$

The action of the boundary operator on the chain groups leads to the definition of three more groups. Firstly, the image of ∂_p is a subgroup of C_{p-1} called the boundary group B_{p-1}. Secondly, the set of all p-chains that have empty boundary forms the group of k-cycles Z_p. These two groups are related by the fact that the boundary of a boundary is empty. This is the fundamental property of the boundary operator, viz., $\partial_p \partial_{p-1} = 0$. It implies that B_p is a subgroup of Z_p. Figure 5.18 presents an example of the action of the boundary operator on a 3D cube $c_3 \in \mathbb{R}^3$.

The *homology groups* are defined as the quotient groups $H_p = Z_p / B_p$, also denoted $H_p()$. This means that two p-cycles w_p and z_p belong to the same homology class if their difference is the boundary of some chain, i.e., $z_p - w_p = \partial_{p+1} v_{p+1}$ (see an illustration of these exact chains in Fig. 5.19). Finally, the number of distinct equivalence classes of H_p is the pth Betti number β_p, which effectively counts the number of p-dimensional holes in X. When $p = 0$, the Betti number counts the number of path-connected components of X. With the Betti numbers, we can calculate the *Euler characteristics* χ_{E} of the simplicial complex $\mathcal{K}$ of dimension n (maximal rank of its simplices) in the form

$$\chi_{\mathrm{E}}(\mathcal{K}) = \sum_{p=0}^{n} (-1)^p \ \mathrm{rank}\ H_p(\mathcal{K}) \ .$$

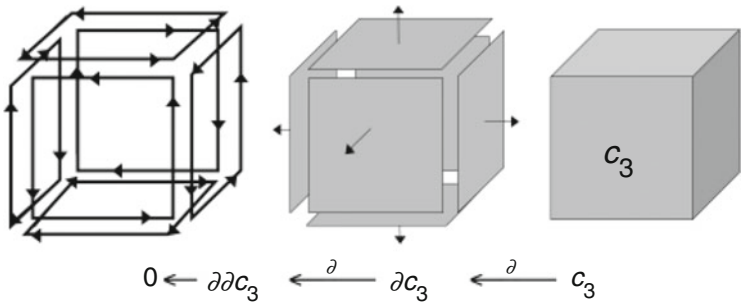


Fig. 5.18 Example of the action of the boundary operator ∂ on a cube c_3 submerged in $\mathbb{R}^3$. *From right to left*: The first action of the boundary operator generates a 2D cube surface, whose boundary is the null set. However, if we decompose this surface into square faces, i.e., 2D cells, each such face has its boundaries, which are the square frames. Their boundaries are the null set, too. In the picture, we notice the essence of the way algebraic topology works: divide any complicated geometric object (in this example c_3) into simpler geometric objects with known differential properties, and structure these simpler cells algebraically

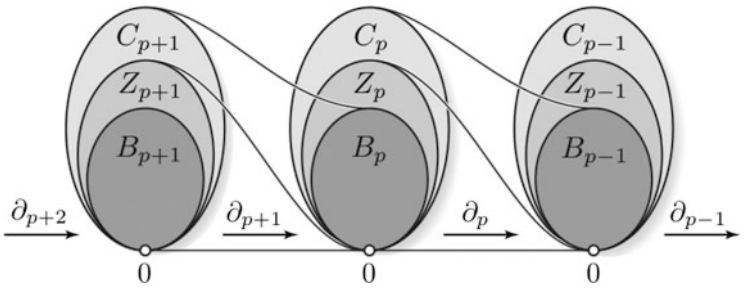


Fig. 5.19 The key picture of homology. The chain group C_p is a set of p-dimensional simplices, the cycles Z_p are chains without boundary, and boundary cycles B_p are cycles, but they are also boundaries of higher order chains. The boundary operator ∂_p homeomorphism connects all these homology structures in an *exact sequence*, that is, the image of one homomorphism equals the kernel of the next

For subsets of $\mathbb{R}^3$, we can interpret β_1 as the number of independent tunnels, and β_2 as the number of enclosed voids. For example, the solid torus $D^1 \times D^1$ has $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 0$, whereas the surface of a torus $T^2 = S^1 \times S^1$ has $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 1$. The latter calculation shows that a T^2 torus (a 2-torus) has 2 holes, out of which one is 2D (the large circle) and one is 3D (the inside). In other words $\chi_{\mathrm{E}}(T^2) = 1 - 2 + 1 = 0$. This affirmation can be written

$$\operatorname{rank} H_p(T^2) = \binom{2}{p},$$

and in general,

$$\text{rank } H_p(T^k) = \binom{k}{p} .$$

If we have a sub-complex $\mathcal{A} \subset \mathcal{X}$, we can define the *relative homology* group as $H_p(\mathcal{X}, \mathcal{A})$, which is constructed in the same way as $H_p(\mathcal{X})$, except that we use simplices from $\mathcal{X} - \mathcal{A}$ only. The special thing about the relative homology is that we can write an exact sequence[2]:

$$\cdots \longrightarrow H_p(\mathcal{A}) \longrightarrow H_p(\mathcal{X}) \longrightarrow H_p(\mathcal{X}, \mathcal{A}) \longrightarrow H_{p-1}(\mathcal{A}) \longrightarrow \cdots . \tag{5.35}$$

The homology groups are the basis for a system of theorems which allow us to calculate the algebraic topological properties of any complex (simplicial or CW). In the following we will denote the simplicial complex $\mathcal{X}$ associated with a topological subspace X simply by X. One important theorem is the *excision theorem*, which states that, if we have a sequence of topological subspaces $C \subseteq A \subseteq X$ and the closure of C is included in the open part of A, then we have a group of isomorphisms

$$H_p(X - C, A - C) \cong H_p(X, A) . \tag{5.36}$$

Let X be a disjoint union of simplicial complexes. Then we have

$$X = \bigsqcup_{i \in I} X_i \longrightarrow H_p(X) \cong \bigoplus_{i \in I} H_p(X_i) . \tag{5.37}$$

From these equations, it is possible to calculate all the relative homology groups for a wide variety of topological spaces. For example, for the spheres and disks, we have

$$\text{rank } H_0(S^n) = \begin{cases} 2 & \text{if } n = 0 , \\ 1 & \text{if } n > 0 , \end{cases}$$

$$\text{rank } H_p(S^n) = \begin{cases} 0 & \text{if } 0 < p < n, \text{ or } p > n , \\ 1 & \text{if } p = n , \end{cases}$$

$$\text{rank } H_p(D^n, S^{n-1}) = \delta_{pn} ,$$

where δ_{pn} is the Kronecker delta symbol. In addition to the description and classification of topological spaces, other applications of homology include fixed

[2] A sequence of group morphisms is exact if the image of each map is the kernel for the following morphism, as in Fig. 5.19.

point theorems, homeomorphism theorems, calculations of the degree of maps, and the Künneth formula for direct sums of groups.

The most valuable mixture of algebraic and differential geometry (topology) methods with the most prolific results is *frame bordism* theory. This theory developed mainly because the algebraic problems are more tractable and their solutions lead to geometric consequences [152]. We introduce here a few elements of frame bordism inspired by Davis and Kirk's course in algebraic topology [189]. Let M be a compact and smooth manifold with or without boundary, and $V \subset M$ a compact submanifold whose boundary is contained in the boundary of M in such a way that V meets the boundary of M *transversely*, that is, the tangent spaces to V and to ∂M at the same point generate together the tangent space of the ambient manifold. In the following, we need the definition of the normal bundle of V from Sect. 4.5, denoted by $\nu(V \hookrightarrow M)$ [see (4.19)]. We define the *tubular neighborhood* of the submanifold V as the embedding $f : \nu(i) \to M$ which restricts to the identity on V. In other words, we associate with every point of the submanifold V the orthogonal complement of its corresponding tangent space, and then we canonically project this orthogonal complement in M. We introduce two definitions:

Definition 9 We define a *framing* of a submanifold $V^{k-n} \subset M^k$ to be an embedding $\Phi : V \times \mathbb{R}^n \to M$ such that $\Phi(p, 0) = p$ for all p in V. We call the pair (V, Φ) a *framed submanifold*.

Definition 10 If (W^{k+1-n}, Ψ) is a framed submanifold of $M \times I$, then the two framed submanifolds of M given by intersecting W with $M \times \{0\}$ and $M \times \{1\}$ are *frame bordant*.

We denote by $\Omega_{k-n,M}$ the set of frame bordism classes of $(k-n)$-dimensional framed submanifolds of M.

For any framed submanifold, we can construct the *collapse map*, defined on M with values in $\mathbb{R}^n \times \{\infty\} \cong S^n$, sending any point in M which is not in the range of Φ to the point at ∞. Further, for any point of M which is an image of some $(p, v) \in V \times \mathbb{R}^n$, the collapse map sends it to v. This map, also known as the Pontrjagin–Thom construction, introduces the translation from bordism, as a concept in differential topology, to algebraic topology through the following theorem [151, 152]:

Theorem 9 *The collapse map induces a bijection between $\Omega_{k-n,M}$ and the set of classes of equivalence modulo homotopy of maps from M to S^n.*

We understand by the homotopy class between two submanifolds the set of all maps between the two submanifolds which can be smoothly deformed one into the other. Theorem 9 states that, if we can map the total manifold M^k into a sphere S^n in several ways, and we structure these ways modulo homotopy (two ways in which we can map are indiscernible if they can be smoothly mapped one into the other), then the same algebraic structure is obtained if we classify all possible $(k-n)$-dimensional submanifolds V^{k-n} of M modulo frame bordism. Put simply, Theorem 9 shows that the study and classification of the framing submanifolds of M is an essential property

of M, because it tells us the structure of all possible maps of M into spheres of different orders.

Here is an example (without proof) to provide some geometric insight: any framed circle in S^2 is frame bordant to the empty set (so called null-bordant). Take for V the equator with the framing along its intersecting meridians. The equator is an S^1 submanifold which is the boundary of the D^2 disk in the ball D^3, and hence contractible to a point. It can be mapped to just a point of the sphere. In this way, the structure induced by homotopy classes of the equator in the S^2 sphere is in one-to-one relation with the structure induced by frame bordant classes of the 2D disk in the 3D ball.

As we can see, algebraic topology is not only a traditional chapter in mathematics, but it forms the background for several new areas of application-oriented mathematical research. The last few decades have shown a growing number of connections between continuous and discrete mathematics, including emerging fields like computational algebraic topology, topological robotics, stochastic topology, and combinatorial algebraic topology and concurrency.

5.6 Classification of Continuous Structure by Discrete Criteria

Mathematics, like any other scientific creation, be it generated in our brain or acknowledged and understood through our brain, has an inner duality between its discrete (digital, additive, musical) and continuous (analog, geometric, visual) counterparts, spiced with flavors from statistical methods and stochastic systems.

The continuous part of mathematics, initially constructed as geometry, has developed in modern times into the more abstract topology and calculus. Its fundamental building blocks are topological spaces, maps, continuity and differentiability (smoothness), and the discipline where they all combine into the most continuous of all mathematics is manifold theory (see Fig. 5.20).

Manifolds are spaces with continuous properties, and depending on continuity/smoothness criteria, they can be studied in four main classes:

1. **TOP.** Topological manifolds are among the weakest continuous structures. They do not necessarily admit a linear or differential structure, and if such a structure is defined on them, it is not necessarily unique (*Hauptvermutung*, i.e., the main conjecture).
2. **Handles.** These are topological manifolds admitting a topological decomposition into handles.
3. **PL.** Piecewise linear manifolds (formerly called combinatorial) are topological manifolds together with a piecewise linear structure defined by means of an atlas, such that one can pass from chart to chart in the manifold by piecewise linear functions. This category is slightly stronger than the topological procedure of

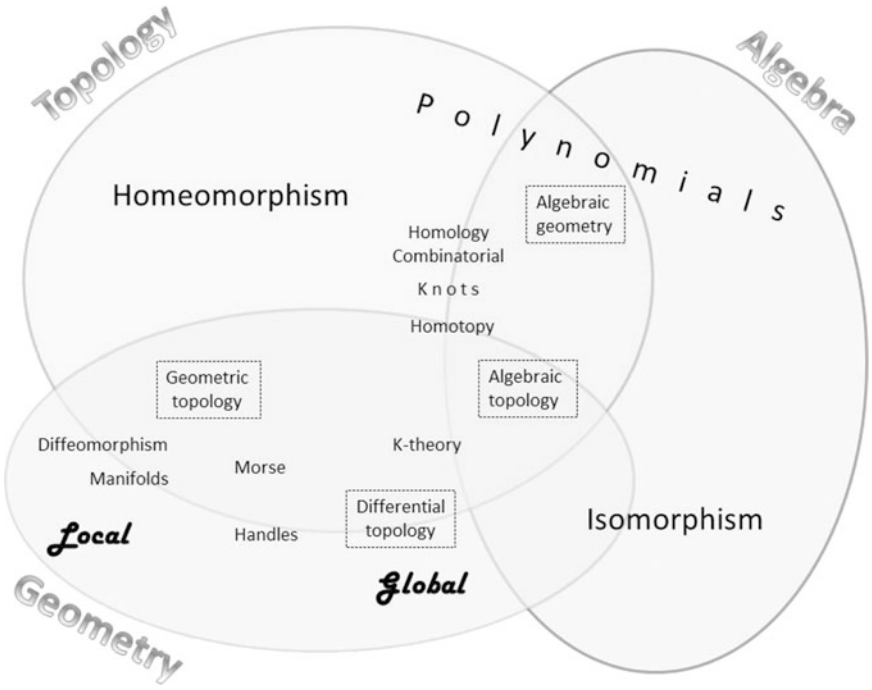


Fig. 5.20 Different areas of topology, algebra, and geometry and their interactions

triangulation. However, PL do not always have smooth structures, and are not always smoothable.

4. **DIFF.** Differentiable (smooth) manifolds. These have canonical PL structures and they are uniquely triangulable (Whitehead's theorem on triangulation, 1940).

Smoothness properties are strongly dependent on the number of dimensions of the manifold. For dimensions less than or equal to 3, the TOP, PL, and DIFF properties, if satisfied, are simultaneously satisfied.

In any dimension other than 4, a compact topological manifold has only a finite number of essentially distinct PL or smooth structures. In dimension 4, compact manifolds can have a countably infinite number of non-diffeomorphic smooth structures. Four is the only dimension for which a manifold can have an exotic smooth structure. A 4-manifold has an uncountable number of exotic smooth structures [190].

One of the methods used to study topological manifolds is based on the representation of a manifold as a union of topological balls with non-intersecting interiors and with boundaries intersecting in a special way.

A handle decomposition is to a manifold what a triangulation is to a topological space. In many regards, the purpose of a handle decomposition is to have a language analogous to triangulations (or CW complexes, we will see about these shortly), but

adapted to the world of smooth manifolds. Thus a handle in the world of differential manifolds is a smooth analogue of a cell in a triangulated (CW complex) world.

The problem of decomposition of topological manifolds into handles is relatively complicated. It is known that any closed topological manifold of dimension higher than 4 can be decomposed into topological handles. Manifolds of dimension less than 4 are combinatorially triangulable, and so can be decomposed into handles. It has been proved that there exists a manifold of dimension 4 which does not admit a handle decomposition, because 4-manifolds have a handle-body decomposition if and only if they are smoothable [191].

Moreover, there is a close relationship between smooth handle decompositions of manifolds and smooth functions defined on these manifolds having non-degenerate points, the so-called Morse functions [192].

5.7 Triangulations and CW Complexes

At the manifold level of complexity, discrete mathematics interferes with continuity properties mainly through the attempt to linearize the manifold in order to simplify it. Making a manifold look like a vector space, at least locally, makes it much simpler to study. One traditional procedure for intertwining discrete and continuous approaches by linearization is the *triangulation of manifolds*, that is, approximating them with linear algebraic spaces, and mapping them into linear geometric quantities, like simplices, polygons, or polyhedra. A triangulation of a topological space is the construction of a simplicial complex homeomorphic to the manifold, together with that homeomorphism [193]. Triangulation is useful for determining the algebraic properties of a topological space, and consequently for determining the algebraic structure of its topological invariants. For example, one can compute homology and cohomology groups of a triangulated space using simplicial homology and cohomology theories rather than more complicated theories [194].

For topological manifolds, there is also a stronger notion of triangulation: a piecewise-linear triangulation. This is one with the extra property that the *link* of any simplex is a piecewise-linear sphere [194, 195].

Definition 11 The link of a simplex σ in a simplicial complex K is a sub-simplicial complex $U \subset K$ consisting of the simplices τ that are:

1. disjoint from σ,
2. such that both σ and τ are faces of some higher-dimensional simplex in K.

For instance, in a 2D piecewise-linear manifold formed by a set of vertices, edges, and triangles, the link of a vertex v consists of the cycle of vertices and edges surrounding v: if τ is a vertex in this cycle, it and v are both endpoints of an edge of K, and if τ is an edge in this cycle, it and v are both faces of a triangle of K. This cycle is homeomorphic to a circle, which is a 1D sphere. In any simplicial complex

homeomorphic to a manifold, the link of any simplex can only be homeomorphic to a sphere. As we can see, there is a strong correlation between algebra (graphs) and topology (spheres) through the concept of link.

Differentiable manifolds admit a piecewise-linear triangulation, technically by passing via the PL category. Topological manifolds of dimensions 2 and 3 are always triangulable by an essentially unique triangulation. Any of their triangulations is a piecewise linear triangulation. In dimension 4, there are examples of compact manifolds having an infinite number of triangulations, all piecewise-linear inequivalent. In dimensions greater than 4, there are manifolds that do not have piecewise-linear triangulations.

When combining algebraic and topological methods (see Fig. 5.20), one very productive method is to use graph theory. For example, one procedure of triangulation of a surface is the embedding of a graph onto the surface in such a way that the faces of the embedding are exactly the cliques of the graph (the Whitney triangulation). In this way every face is a triangle, every triangle is a face, and the graph is not itself a clique. The clique complex of the graph is then homeomorphic to the surface [193–196].

The simplest way to analyze a manifold is to triangulate it towards a simplicial complex, that is, to build a simplicial complex homeomorphic to the manifold. A simplicial complex is a topological space of a certain kind, constructed by 'gluing together' points, line segments, triangles, and their n-dimensional counterparts (see Sect. 5.5). For example, any compact topological manifold of dimension greater than 4 which has a piecewise linear (PL) structure is triangulable to a simplicial complex. There is a well-developed technique based on the Kirby–Siebenmann invariant which tells us whether or not a topological manifold admits a PL structure [193, 197]. However, there are topological manifolds which do not admit any PL structure, but are still homeomorphic to some simplicial complex.

Simplicial complexes are rich in algebraic properties. Basically, a simplicial complex is a hypergraph with a closure property. For example, they have order-preserving morphisms between ordered finite sets. Simplicial complexes should not be confused with the more abstract notion of a simplicial set appearing in modern simplicial homotopy theory. The purely combinatorial counterpart to a simplicial complex is an abstract simplicial complex. Simplicial complexes are very often used in combinatorics.

A more sophisticated algebraic-topological structure is the *CW complex*, which is a type of topological space introduced to meet the needs of homotopy theory (see Sect. 5.5). This class of spaces is broader, and it has some better categorical properties than simplicial complexes, yet retains a combinatorial nature that allows for computation. The notion of CW complex has a natural adaptation to smooth manifolds through the handle decomposition, which is closely related to surgery theory. A CW complex is made of basic building blocks called cells. The precise definition given in Sect. 5.5 prescribes the way the cells may be topologically glued together [198]. CW complexes have good topological properties: they are Hausdorff, they are locally contractible, the product of two CW complexes can be made into a CW complex, and they are paracompact. Graphs, polyhedra, differentiable

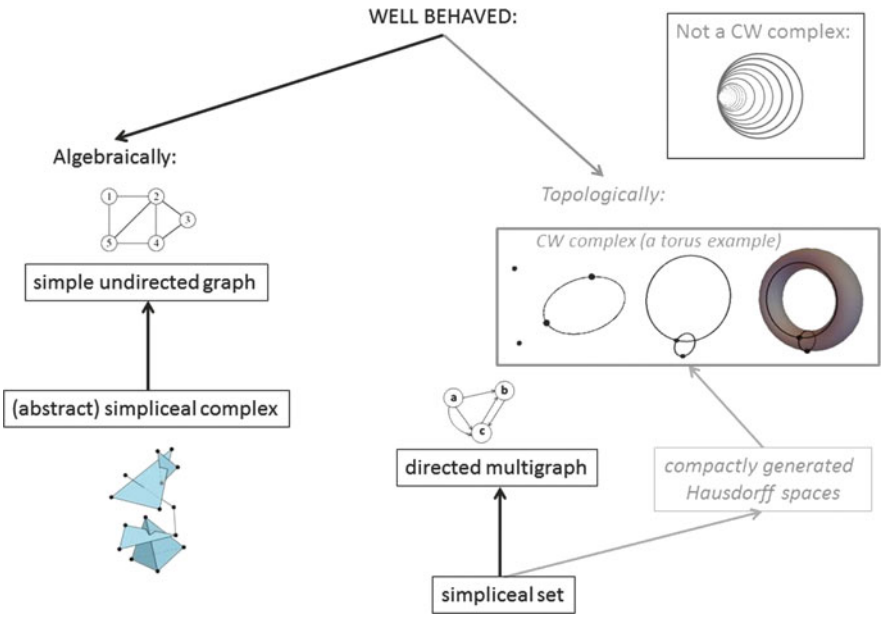


Fig. 5.21 Examples of structures from algebra and topology with good behavior

manifolds, and algebraic and projective varieties all have the homotopy-type of CW complexes. The 'Hawaiian earring' is an example of a topological space that does not have the homotopy-type of a CW complex. An infinite-dimensional Hilbert space is also not a CW complex (see a helpful diagram in Fig. 5.21).

In addition to simplices and CW complexes, mathematicians have invented a useful structure called a *simplicial set*, which is a (purely algebraic) model capturing those topological spaces that can be built up from simplices and their incidence relations. A simplicial set is a combination of elements from algebraic topology (because simplicial sets with ordered vertices are useful in algebraic topology) and geometric topology (by using simplicial sets one can generalize triangulations).

Simplicial sets work in a similar way to CW complexes in modeling topological spaces, with the crucial difference that simplicial sets are purely algebraic and do not carry any actual topology (see Fig. 5.21). A simplicial set is a collection of simplices of different orders and a collection of morphisms between simplices of different orders, like shrinking down to faces, or expanding into a higher order degeneracy. In a simpler language, a simplicial set is a CW complex made of simplices. These simplices can be glued to themselves or multiply glued to each other. The corners of each simplex are consistently locally ordered. Moreover, every generalized triangulation with consistently locally ordered vertices is represented by a unique simplicial set.

Simplicial sets should not be confused with abstract simplicial complexes, which generalize simple undirected graphs rather than directed multigraphs. While

algebraic topologists largely continue to prefer CW complexes, there is a growing contingent of researchers interested in using simplicial sets for applications in algebraic geometry, where CW complexes do not naturally exist.

5.8 Connecting Discrete and Continuous

Algebraic geometry, algebraic topology, differential topology, and geometric topology form a frontier between topology, differential geometry, and algebra. Their main tools include studies of homotopy, homology, and cohomology defined on CW complexes. The ordinary homology theories satisfy the 'dimension axiom', viz., the homology of a point vanishes in dimensions other than 0. They are determined by an abelian group G. In addition to these theories, we have the K-theories, which are related to vector bundles over topological spaces. Different sorts of K-theory correspond to different structures that can be put on a vector bundle. The next tier up is represented by bordism and cobordism theories which study manifolds. Here is where the boundary reaches its greatest theoretical importance because in these theories a manifold is regarded as trivial if it is the boundary of another compact manifold. The cobordism classes of manifolds form a ring algebraic structure that is usually related to some generalized cohomology theory, and to the Thom spaces of certain groups. Finally, on top of all these theories, we have the theory of elliptic curves.

Quantum field theories, which have important applications in string theory, statistical mechanics, and condensed matter physics, are mainly studied using tools from algebraic topology. In 2 dimensions, for example, where there is an infinite-dimensional group of local conformal transformations, physicists like to use conformal field theories. Each conformal or superconformal field theory is labeled by two integers which are the genus and the number of punctures of some Riemann surface, what topologists call a 'pair of pants'. Each distinct way to sew such Riemann surfaces together from pairs of pants corresponds to a different Lagrangian description of the theory, and a Lagrangian description is weakly coupled in the region of the parameter space where the pairs of pants are sewn together by long tubes. Each tube represents an $SU(2)$ gauge group, and each pair of pants represents a block of matter hypermultiplets. The sewing of the Riemann surface encodes the detailed structure of the matter representations. Basic topological operations have a direct translation in the language of these field theories. Sewing two Riemann surfaces together, or adding a handle to a Riemann surface, corresponds to gauging two flavor symmetries at the two sewn-together punctures.

Part III
Applications

In this third part, we match the theoretical fields presented in Part II with selected applications in interesting areas of science.

Much effort has been devoted in the literature [199] to understanding synchronization phenomena and critical phenomena in complex networks, in biological systems, or in computer science. This includes both extensive numerical work as well as analytic approaches. Studies of the statistical properties of large real-world networks have revealed their highly inhomogeneous and hierarchical structure. In addition, comparative analysis has been carried out on large networks from different fields, in which real-world networks were compared with large scale-free random graphs. It has been found that in, contrast to random scale-free graphs, real-world networks can be characterized by correlations in the node degrees, a small-world property, by the specific motifs, short cycles, and communities of nodes that are linked together in densely connected groups [199].

The local structure of a graph influences the size and shape of its boundary. For example, in undirected networks, the high density of short loops (high clustering coefficient) together with small graph diameter gives rise to the small-world effect. In directed networks, the correlation between the numbers of incoming and outgoing edges modulates the expected number of short loops. If the two opposite directions in an edge are not correlated, the number of short loops is significantly reduced as compared to the case where the degrees are positively correlated. Such differences between directed and undirected networks can explain the shape of very large networks, like the structure of the tissue in nervous systems and bone marrow, the Internet, or large social networks.

The geometry and topology of graphs have a deep influence on the physical properties of dynamical processes defined on complex networks. It is interesting to identify the geometrical and topological properties of graphs which might affect the dynamical behavior of a model. For example, increases in urban populations can lead to problems of urban decay, such as widespread poverty, high unemployment, and rapid changes in the racial composition of neighborhoods, inducing social movements in response. Theoretical network methods can be used to identify isolated neighborhoods in big cities with a complex web of roads, walkways, and

public transport systems [199]. Estimates of the size of the boundaries of isolated neighborhoods can provide good solutions for urban planning.

This part is divided into five chapters. Chapter 6 introduces some ideas regarding the importance of boundaries in the philosophy of science. Then Chap. 7 reviews the literature on networks and their relations to boundaries. Section 7.1 deals with complex networks, and this is pursued in a natural way with Sect. 7.2, dedicated to world networks. Section 7.3 approaches an interesting new problem, namely the shape of the Internet, discussing findings which have emerged from the analysis of graphs and networks using topological and geometrical tools. Finally, Sect. 7.4 approaches the topic of the previous section, the shape and boundary of very large networks, from the standpoint of spaces with high dimensions.

Chapter 8 describes and discusses the very modern topic of big data and their handling from different mathematical perspectives. Section 8.1 describes how to frame big data sets in a geometrical structure which allows convenient and efficient study. The methods of algebraic topology presented in Part II and the results obtained from them are applied in Sect. 8.2, which is devoted to special homology methods for handling big data sets. The topological analysis of big data sets continues in Sect. 8.3, where we describe ways of handling situations in which data sets contain gaps and holes.

Chapter 9 presents some physical applications to free liquid boundaries. In the introduction, we review the main implications of the existence of a boundary for a mass of fluid, then list and comment upon the dimensionless numbers and parameters important for fluid dynamics. In Sect. 9.1, we apply the concepts introduced in Sects. 4.5 and 4.6 on fibre bundles and discuss the formal approach to hydrodynamics from this mathematical perspective. Since this section focuses mainly on fluid kinematics, we devote Sect. 9.2 to a differential geometry approach to the Navier–Stokes equations and fluid dynamics.

Section 9.3 contains an almost complete mathematical description of 2D soap films with boundary, presenting the main uniqueness theorems for shapes. In Sect. 9.4, we return to full 3D liquid systems with boundary, namely drops, and discuss their Hamiltonian properties and stability. In Sect. 9.5, we specifically study the rotation of 3D liquid drops, and the various equilibrium shapes one can obtain from different regimes. Then in Sect. 9.6, we return to 2D drops and apply the same Hamiltonian principles in order to study their stability. Moving even further into the subject, Sect. 9.7 presents very recent results concerning patterns generated by 2D drops, such as Leidenfrost systems. Section 9.8 is devoted to the rotation and shapes of such 2D drops, and presents both the experimental aspects and the theoretical approaches. A summary of the main findings and of the principal issues and ideas which have arisen in the previous chapters and sections is provided in Sect. 9.9, which is probably one of the key sections of the book. We draw attention to the similarity between 2D rotating fluid systems at different scales and in different physical systems, emphasizing the universality of their behavior through the dynamics of their boundaries.

Chapter 9 contains three appendices (Appendices 1–3) treating in more detail the second fundamental form for surfaces embedded in $\mathbb{R}^3$, the calculus of variations, and n-dimensional spheres in rotation.

The last chapter of Part III (Chap. 10) aims to unify the ideas presented throughout the book, from such different angles of human knowledge. Towards the end of the chapter, we present a pioneering idea, namely the hypothetical importance of other senses and perceptions (for example, taste and olfaction) that are not supported by theoretical models in the artistic or mathematical approaches.

Chapter 6
The Boundary in the Philosophy of Science

Anything worthy of the name boundary will effect set-theoretically describable divisions, even if more complex ones than the simple twofold division envisaged by traditional philosophy

Mark Sainsbury

In [200] Nancy Cartwright elaborates on the idea that the picture of the world offered us by physicists and mathematicians is governed by a few simple general laws, a vision that would have been familiar to David Hume in the eighteenth century. Observable regularities in nature allow us to infer causal connections. However, the fact that up to now a certain event has always been observed to go in a certain cause-and-effect way does not necessarily mean that it will always do the same thing in the future. For what guarantees that nature will not change tomorrow? Physical laws assert what are supposedly eternal regularities, but there is nothing necessary about them.

If one accepts this fear, the world in which we live, unlike the one inside the laboratory, is an unpredictable place, marked by discontinuities and catastrophes. Cartwright does not deny the essence of the laws, but makes a point about limitations to the confidence we must have in laws when they apply outside lab boundaries. The confidence scientists have in their laws comes from the fact that, unlike for example climatologists, medical doctors, or economists, physicists are able, in the closed world of the laboratory, to ensure that the outcomes they predict are in fact attained. Physicists create very restrictive conditions under which their predictions will come true. As the econometrician Tyrgve Haavelmo commented: "physicists confine their predictions to the outcomes of their experiments." When it comes to predicting things in the real world, the results are trickier. So science has an inherent boundary of applicability: the more precise and deterministic the law, the less applicable it becomes outside of laboratory boundaries.

The laws of physics are constructed in such a way that they can operate only at a high level of abstraction. It is because of this that they are remote from reality. Take, for example, the first law of dynamics, the law of inertia. In order to apply it, one has to be in a place far away from any type of interaction, completely isolated, which contradicts the possibility of verifying the law, since to observe the behavior, one needs a measurement device which will induce an interaction in the system.

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_6

For example, a paradigm of scientific knowledge for Cartwright is exemplified by propositions of the form: 'Aspirins cure headaches'. By this we do not mean that aspirins always cure headaches: sometimes they do, sometimes they do not. What must be said is only that the property of being an aspirin carries with it the capacity to cure headaches. Another boundary for applied science observed by Cartwright is the concern that science has no potentiality to change the world. For example, quantum mechanics, instead of representing a step forward in our knowledge of the atomic world, actually involves a decrease in our knowledge, being severely limited in its scope of operation. Quantum physics works only in specific situations when classical physics fails, and vice versa, in a way without reason and cause.

6.1 Boundaries in Epistemology

In epistemology, the concept of boundary is used abundantly in various contexts, yet none of them can be seen as purely epistemic [201]. There are disciplinary boundaries, temporal boundaries (real or assumed), linguistic boundaries, generic boundaries, medium boundaries (oral versus written), etc. Lehoux believes that all these boundaries can acquire a place in the spotlight, when we look closely at the creation myth of science, or philosophy, but "the epistemic moral should perhaps be held in that same suspension as the causes invoked by any other good fiction" [201].

The authors in [202] argue that, in drawing epistemic boundaries between science and non-science, we have to look at how boundaries between science and pseudo-science have been constructed historically and the ways in which these boundaries have been put to use by scientists trying to secure resources and credibility while denying them to others. The question of how to draw boundaries between science and non-science is sociological rather than philosophical. Philosophers like Karl Popper believe it is possible to draw a line separating science and non-science, yet other scholars studying science now talk about a multitude of boundaries. Once firm boundaries, like those between science and the public, and between religion and science, have become increasingly difficult to pin down, and new boundaries have emerged within science itself.

One way to investigate the boundaries in science is to focus on their particular 'epistemic communities', that is, from where to where a certain domain of science appears to be true for a certain community. Disciplinary boundaries in science have always been ontological: what a biologist can explain differs from what a physicist can explain. Hence, the objects of science become epistemically unstable when they move across boundaries. In order to understand the historical meaning of an object, Delbourgo believes [202] that we must pay attention not just to the boundaries that are being crossed, but also to the particular modes of transfer. The question is how the particular dynamics of the epistemic conversion can be induced and by whom. What cultural practices did some successful scientists bring to the established community of scientists? Does it concern the completeness of these new observations? Or could elements of this process be understood using the idea

of a trading zone? This includes the idea of 'creolization', the process by which new languages and cultural practices develop as formerly distinct cultures intermix, and the idea of the hybrid, the notion that every individual's identity is in fact fractured. According to Baus [202], there are four complementary dimensions in the formation of new scientific communities: the construction of a communicative space, the formation of new institutions, the flow of symbols in the form of texts and instruments, and the migration of individual scientists between communities.

An exotic and new type of boundary in the philosophy and history of science relates to 'conspiracy theories' [203]. Even dismissed in social sciences as irrational, bad science, religious belief, scientific communities practise a sort of persistent disqualification which is a form of boundary work, e.g., the exclusion of lay knowledge by scientific experts forming a global power elite. Given their critique, which resonates with social scientific understandings of science, it is concluded that conspiracy theorists compete with (social) scientists in complex epistemological battles [203].

6.2 Triadic Classifications, Complexity, and Boundaries

Boundaries can be created by or between different groups of scientists. According to Spee and Jarzabkowski [204], there are three such types of knowledge boundaries: syntactic, semantic, and pragmatic. Syntactic boundaries are the simplest, providing that there is a common syntax and assuming that knowledge can be transferred between scientists. A semantic boundary is more complex because common meanings need to be developed in order to translate knowledge. Pragmatic boundaries are the most socially and politically complex, as common interests need to be developed to transform knowledge at a pragmatic boundary. For example, during periods of strategic uncertainty, groups of scientists within different labs might have different political interests about what constitutes the appropriate course of strategic action. Boundary objects assist in the transfer, translation, and transformation of knowledge across the different syntactic, semantic, and pragmatic strategies.

At present the market success of a new research project is influenced, if not determined, by its content of neoteric concepts like sustainable development, quantum computers, complex systems, globalization of information, the information society, big data, computational science and engineering, the knowledge society, the knowledge economy, and nanoneurotechnology, etc. But what made these terms win out over other phrases? It may well be because of a diversification of disagreements, or an increase in cooperation between different fields of knowledge. This is an explanation borrowed from the science of complex systems. In particular, when analyzing such interactions between traditional fields, we need to explain the existence (and dynamics) of their boundaries.

According to Eugene Gendlin, and in the spirit of Heidegger and Husserl, there are three main types of interactions involving living systems: (1) the battle, or absolute adversity when one disappears, as in the case of the immune response;

(2) the contention or the relative adversity when only one can win, as in niche segregation in natural selection; and (3) the debate, or complex adversity when nobody wins, but a new state (truth) is born. In the same spirit, in his book *The Structure of Scientific Revolutions* from 1962, Thomas S. Kuhn distinguishes three types of incommensurabilities in the theory of evolution and revolution: (1) methodological, when there is no common measure because the methods of comparison change; (2) perceptual/observational, when observational evidence cannot provide a common basis for theory comparison, since perceptual experience is theory-dependent; (3) semantic, when the languages of theories from different periods of normal science may not be inter-translatable from old to new theories.

By induction, there may be at least three different approaches for understanding and partially breaking down a complex system into its minutest parts and their boundaries. One way is to begin from a particular complex system and to address a variety of questions coming from that particular domain and its points of view. This approach uses causality constraints, analytic tools, similarities, and induction, which basically means a search for functions.

The second approach almost eliminates the deterministic point of view and uses statistical theories like random matrices, Bayesian theory, computational information, or statistical complexity. This way, patterns and rules can be detected, and the method searches for structure. These two traditional procedures have been successful, but are still tributary to some reminiscences of linear thinking.

The third approach to complex systems and their boundaries cuts across particular domains, dissolves the formal and traditional boundaries, and most importantly, thanks to modern technology, shrinks things down to the nanoscale where the complexity clearly arises from nonlinear interactions, preventing us from obtaining a realistic description of a system by dissecting it into its components [205].

While the first two approaches lead to domain-specific cross-disciplinary fields, the third is responsible for the interdisciplinary type of inquiry. This approach starts from fundamental questions relevant to all domains, and searches for rigorous methods to solve particular open problems.

This third approach is handy when we deal with nonlocal mathematizable models with multiplicity properties, but which are nevertheless robust and flexible. It is widely used in classic semantic theories where a predicate has meaning through its extension only. Concepts are boundary drawers in semantic theories. A good example to mention here are the endless attempts to predict laminar flow and turbulence through smooth solutions to the Navier–Stokes equations (one of the Clay Millennium Problems). New interdisciplinary and transdisciplinary solutions then seem necessary, in particular to treat challenging problems like correlations between phenomena at different levels, self-organization, robustness and flexibility of the system, evolutionary memory systems, and growth of supplementary structures. In order to be able to construct such solutions, we need to understand the boundaries of the disciplines, the boundaries of different phenomena, of systems, and even the boundaries of the methods. The dynamics of these boundaries can provide shortcuts and simpler solutions, by choosing cross-examples from a wide variety of topics and disciplines. Boundaries are what count, because a solution to a complex problem

must use its boundary to segregate the consequences which fall under it from the consequences which do not.

6.3 Boundarylessness as the Philosophy of Vagueness

Finally, the question is: how can we handle diversity and interaction by avoiding such an intense study of the boundaries? The vast majority of concepts classify by setting boundaries, but there are some that do not. We can make the difference between two types of boundaries.

1. **Separating boundaries.** These are the boundaries between different subsystems or distinct phases of one system. In this case we have a total complex system which, in some of its regions, usually hyper-subsystems having one dimension less than the dimension of the system, behaves completely different. The properties of the system tend to change so abruptly across these regions, and in that violate the systems laws, that we have to give these regions special attention and laws. The boundary may separate same type of sub-systems, or totally different ones. Such boundaries usually involve transfers and interaction between the regions they separate. These boundaries occur in classificatory processes. Examples are benthic regions, cell membranes, phase separation surfaces, scenes or acts in performing arts, national boundaries, concepts.
2. **Defining Boundaries.** They define the system, contain or surround the system and isolate it from the void outside. These boundaries occur in constructive processes, in boundary-drawing descriptions. As examples we enumerate a painting on wall, a liquid drop in vacuum, a network, the boundary of a geometric object, moral or legal concepts.

According to Sainsbury [206], there are concepts that can be classified without setting ('sharp') boundaries. Examples of boundaryless concepts are traditionally considered to be vague, e.g., red, heap, child, bald. Because there are objects for which the classification of being red is well posed, impossible, or relative. Hence there is no set of red things, and red does not draw boundaries. Sainsbury's theory of *unsharp boundaries* does not use previous attempts to describe boundarylessness as vagueness, as happens in the theory of categories, fuzzy logic, or supervaluation theory.

Philosophers are interested in vagueness especially because of the fascination, and threat, posed by the so-called sorites paradoxes: does a non-bald person become bald by gradually losing a single hair at a time? And if so, when? A situation where there is no longer any hair loss marks the transition, so there can be no transition, hence no sharp boundary. How can this problem be avoided? In [207], the solution to this question is to consider vague concepts, that is, ones without boundary.

According to Sainsbury's ideas, a vague boundary divides things into an infinite number of sets with the cardinality of the continuum. This can be proved by induction. Let us assume that we can define a vague boundary by three sets: one for which an action, property, or predicate can be proved positively true (the positive

extension), one for which it can be proved positively false (the negative extension), and one for which the proof has a borderline conclusion (the penumbral extension). But an action or predicate which effects such a threefold partition is no longer vague, so we have a contradiction. The same contradiction would occur if we tried to use fuzzy logic or supervaluation, because these theories also divide the actions into three sets: the 1s, the 0s, and the ones in-between 0 and 1. In conclusion, if we understand boundarylessness as the absence of borderlines, we end up in vagueness. Boundarylessness cannot be described sharply.

Sainsbury gives an example with the 'set' of strawberries, which apparently draws boundaries because there are no borderline cases of strawberries. But this is just an accident, because there could be plants having common features between strawberries and raspberries, and if there are not yet, this fact may well soon be amended by progress in genetic engineering. Concepts like strawberries do not impose boundaries, but constitute a system of contrary boundaryless concepts (Locke used to call these boundary-defying monsters). It would be interesting to find possible future applications of such boundaryless concepts, e.g., in the cognitive sciences, where one can try to discover ways of making a machine understand the purely human concept of vagueness. Other examples can be found in the fields of law, morality, or history.

It seems that some concepts classify by setting boundaries, but some do not. According to the classical philosophical or linguistic picture, the job of classificatory concepts is to sort or segregate things into classes by providing a system of pigeon-holes, by placing a grid over reality, by demarcating areas of logical space. When trying to define boundaries, may be a helpful question to ask is 'what is vagueness?'

Such classification of boundaries may compel us to try to classify all concepts and systems through their boundaries. This sounds a lot like trying to explain the world through the set theory, which direction was investigated and was proven to be a dead end.

It is a well posed question to look for the boundaries of a well-defined system or concept. This happens because well-defined concepts must use a boundary to segregate. The reciprocal, however, is not always true. According to Sainsbury some concepts classify by setting boundaries, but some concepts do not classify by setting boundaries [208]. Sainsbury asks if instead of trying to find exact definitions for boundaries, we should rather solve the preliminary question of 'what is vagueness?' Philosophers have been interested in vagueness for centuries. One reason is the fascination, and threat, posed by paradoxes. Nevertheless, vagueness may be of interest independently of the paradoxes, and it can be a tool for the definition of unclassifiable entities. Sainsbury says 'concepts can classify without setting boundaries.' For example, the concepts 'childhood' or 'red' are vague because they do not classify in a sharp way other concepts. There are things for which 'red' is neither true nor false, for which being 'red' or not doesn't make sense. In Sainsbury system, such vague concepts like 'children' are concepts without boundaries.

From a mathematical-epistemological angle it seems more natural to solve Sainsbury's problem on vagueness and on the non-existence on boundaries for some concepts by using a fundamental theorem in geometry: the boundary of a boundary

is empty. By using pure math in this epistemological context there is always the risk of shifting the initial problem of 'definitions through boundaries' into classical semantics. However, instead of simply using the set theory we can follow the geometrical fact that a boundary has no boundary. From here we can infer that *vague concepts* are themselves the boundary of the sets of *exactly defined concepts*. We illustrate this idea in the case of the entities 'children' and 'children now in this room,' Fig. 6.1. In the left side of this figure we show the concept of 'children present now in this room' by points inside as a solid disk each point representing such a gathering of children. The rest of all possible gathering of children relative to this moment and this room are presented as exterior points from the disk. For the concept of 'children' in general we can use the boundary of the disk, separating the two clearly defined situations. The entity 'children' in this case is represented by a boundary, and hence itself has no boundary as in Sainsbury's entities with vagueness. On the contrary, the way Sainsbury sees the structure of vagueness is presented in the right side of Fig. 6.1. There are situations when the children can be clearly defined and classified with respect to a moment of time and a room. However, according to Sainsbury the 'children' entity floats over the classified entities without structure, limits or bounds.

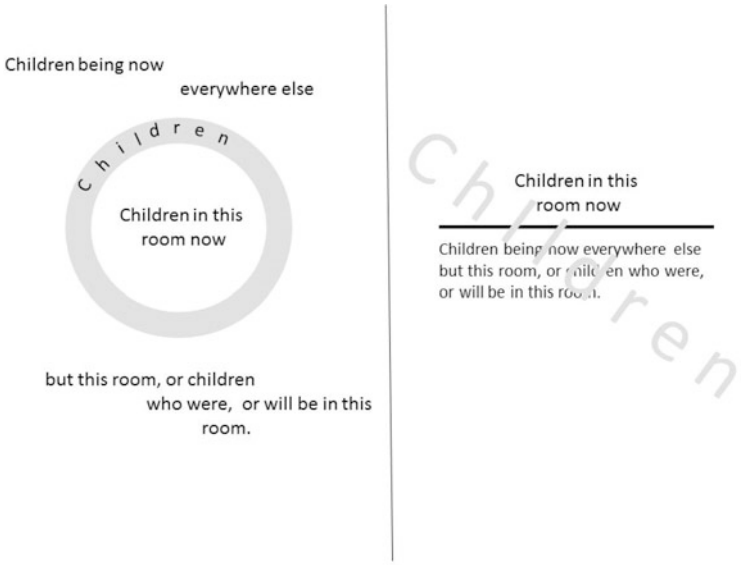


Fig. 6.1 Geometric and Sainsbury representation of the entity 'children.' *Left:* The concept of 'children present now in this room' represented as points inside as a solid disk, each point representing such a specific gathering of children. The rest of all classifiable gathering of children relative to this moment and this room are presented as exterior points from the disk. The entity 'children' in general is represented by the disk boundary, and hence itself has no boundary. *Right:* The Sainsbury vagueness of the entity 'children.' There are classifiable situations, and vague situations like 'children' which float uncertain over the classified entities.

Chapter 7
Networks and Their Boundaries

The idea for the topic in this chapter came from the following question: what is the shape of the Internet? Or, more generally, does a very large network have a boundary, and if so, how can we define it? There is definite interest in the literature for these questions. Callon, for example, discusses boundaries (of socioeconomic networks) in the context of actor network theory (ANT), defining them through the *level of convergence* of the network [209]. The network boundaries are built by classification/elimination of their elements. In his approach, an element lies outside of a network if it weakens the alignment and coordination (that is the convergence) of the latter when moved into the network. Alignment, as opposed to disalignment, is a measure of the network's commensurability, while coordination refers to the process of imposing local rules that stabilize connections. The stronger the coordination, the more predictable the network.

In the Internet case, convergence can be evaluated by crawling procedures, among many others. Figure 7.1 presents the result of a simple experiment of network crawling: searching one word and measuring the number of resulting links and the duration of the search, all on the same standard machine. The horizontal axis represents the number of links on a logarithmic scale, and the vertical axis is the search time in milliseconds. In addition to the mosaic of significations obtained, we notice that the center of mass of the positions of the words tends to remain on a horizontal line, implying that the search time is independent of the number of links browsed and found. This result is a naive finding in favor of the argument that the Internet, or at least the way the Google search engine works in 2014, is indeed a 'small world' network: all links are placed at about the same distance.

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_7

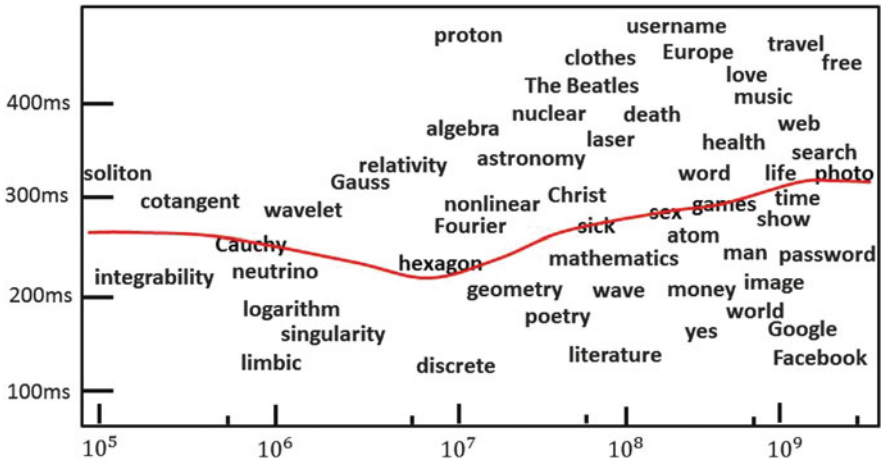


Fig. 7.1 Result of a Google search for different one-part words versus the number of viewable links and the duration of the search. *Red curve*: Average center of mass of the positions of the words in this plane

7.1 Complex Networks

Networks are present in many aspects of everyday life, from river basins to veins, the lymph and nervous systems, ground communication cables, electric power grids, and Internet. In all these cases, the network properties work together to optimize some 'cost function', such as the number of points connected with respect to the length of the web [210].

Let us introduce some quantifiers used in the description of nodes and edges in networks. Among the parameters describing the *centrality* of a node, we have the *betweenness*, which represents the number of shortest paths between all other possible nodes that pass through that node. Betweenness is a global parameter, and its local version is the *betweenness centrality* of a node. The betweenness of a node i is defined as the sum over pairs of nodes of the ratios between the number of shortest paths crossing i between any two nodes, divided by the number of shortest paths between those two nodes all in all.

The *assortativity* is the degree variance. The *coreness* k of a node is the integer showing that this node belongs to a subgraph where all the nodes have at least degree k, and it does not belong to any subgraph of higher coreness.

In connected networks, we define the *farness of a node* as the sum of its distances to all other nodes, and its *closeness* as the reciprocal of the farness. Closeness is a measure of how many edges it will take to spread information from this node to all other nodes in the network.

The *closeness centrality* of a node is defined as the reciprocal of the sum of the shortest paths from that node to any other node. The *global clustering coefficient* is the ratio between the number of closed triplets in the network and the total number

of triplets in the network. The *local clustering coefficient* γ_i for a node i is (2 times according to some authors) the number of edges between its neighbors divided by the number of all possible edges that may exist between its neighbors. We also recall the concept of a *clique*, which represents any complete subgraph; a maximal clique cannot be extended to a larger clique by adding any of its neighbors. An important clique is the triangle which plays a role in the definition of the clustering properties of networks. An example of a highly clustered network is the so-called 'caveman' network, which is a disjoint union of cliques.

Another important parameter of a network is the *hub number* of G, which is the minimum size of a *hub set*. A hub set is a subset in a graph containing a part of a path between any two nodes of the graph outside the hub set. A subset of nodes in a graph which is adjacent to all nodes of the graph is dominating, and in a connected graph, any dominating set is also a hub.

Large graphs are generally classified according to their statistical properties. A few important prototypes are described in the following section.

7.2 World Networks

A very large graph (the number of nodes exceeding hundreds) with a high degree of connectivity may qualify for the definition of a network. However, a network in its most general sense is a large graph together with some dynamics, including real-time graph modifications and transfer of information from node to node. A network can be highly organized and ordered, if it has as substrate a regular or random graph, and is stochastic if it is constructed (and keeps growing or changing) according to some random and probabilistic principles. In addition to this crude classification, a network may also be classified or analyzed through many different approaches, from psychiatric, neurological, social, linguistic, industrial, astrophysical, biochemical, to computing. I remember asking my math professor in college what an operator is? "Anything can be an operator," he replied, "this train passing by can be an operator." So, similarly to an operator, I mention that anything can become, or be understood as, a network.

However, the most interesting dynamical and unpredictable networks are generated by complex systems, so they are complex networks whose behavior is hard to understand and describe in traditional rigorous mathematical terms. The complex networks scale from millimeters in the case of the central nervous system, up to hundreds of thousands of kilometers in the case of the satellite communication web. Probably the oldest complex network ever identified is provided by a fossil, estimated to be 520 million years old, which has revealed the earliest known central nervous system of an animal [211]. The three-centimeter-long fossil (a megacheiran) is a distant relative of spiders, scorpions, and horseshoe crabs. It contains a pair of long, forceps-like extensions from the head. This may have been one of the first successful complex networks built by nature, unless we count the first

metabolic process on Earth, believed by a group of biogeochemical astrobiologists from UCLA to have been formed 3.85 billions years ago [212].

Among the complex networks, probably the most useful to understand in our daily lives are the social networks made 'of the people, by the people, for the people'. Real world social networks are difficult to model because they have highly non-local dynamics, and their complexity grows with size. Apparently, there is still no reliable mathematical theory for such social networks, in spite of dozens of existing theoretical models. All the social network models try hard to match the increasing complexity with network size by creating scaling laws and size-invariant network laws, e.g., random networks, small-world networks, and scale-free networks. As of this moment, none of these models can generate the vast complexity of features and high levels of unpredictability in a real world network.

One interesting and intriguing result is the famous 'six degrees of separation' phenomenon occurring in some special social networks, believed to exist even in some natural networks. The theory is that a person can be connected to any randomly selected individual in the whole world through just five or six intermediary individuals. The concept was first mentioned in the short story *Chains*, by Frigyes Karinthy in 1929, probably following the impact of Guglielmo Marconi's Nobel Prize speech from 1909. Later on, Stanley Milgram conducted experiments to prove this theory, which he termed the *small world* problem. In the 1960s, he conducted landmark studies in order to quantify the typical distance between actors (i.e., nodes) in a social network and to show that one should expect it to be small. His series of experiments attempted to test the idea that the world had become increasingly interconnected thanks to growing globalization.

Interest in this topic has led to research and it has recently become one of the key fields of study in social and biological networks. From many experiments, we know that almost all random networks (i.e., built by adding edges to existing nodes, or by bringing new nodes in a stochastic way) share two global properties: they have a similar value for their diameters (for sufficiently high values of the degree) and are likely to spread. That is, when we count the neighborhood of each node in the neighborhood of a fixed initial node, and keep doing this recursively further and further away, we are likely to find new nodes that are not included in the initial neighborhood. In addition, random networks have low values for the clustering coefficient and the average distance between nodes is small. This can be summarised as follows:

$$\begin{aligned} \text{dia}\, G_{\text{random}} &= \text{dia}_{\max}(|G|)\,, \\ |A^{j+1}\boldsymbol{v}_i| &> |A^j \boldsymbol{v}_i|\,, \quad \forall i,j\,, \\ \langle d_{\text{random}} \rangle &\sim d_{\min}(|G|)\,, \\ \langle \gamma_{\text{random}} \rangle &\sim \gamma_{\min}(|G|)\,, \end{aligned}$$

where $\boldsymbol{v}_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$. But how are these global properties reflected in local properties, and which are more interesting and useful for the individual? It

is legitimate to ask what people really want from a web query, or how individuals can find short paths in a social network using only local information [213].

An interesting construction, although it does not directly answer the questions above, emerges from the small-world model for networks. A small world network can be built according to many recipes [185, 214, 215]. It is definitely known that one can construct a sort of hybrid, or superposition, of a network with high diameter and high clustering coefficient and a random network. The construction is possible by adding new edges between hubs in a random manner. By randomly connecting highly disconnected clusters, the average distance between nodes and the clustering coefficient start to decrease. One can find an intermediate situation where the diameter starts to drop quickly to small values (like the famous 5–6 number), while the clustering coefficient is still large. If such a mixed situation can be obtained before both of these parameters drop to their smallest values, then the corresponding network is called a small world [215].

Among other consequences, in a small-world network, the average value of the distance between any two nodes is polynomially dependent on the logarithm of the number of nodes, i.e., $\langle d(i,j)\rangle \sim (\log |G|)^j$, so the number of edges to be crossed in going from one node to another is exponentially smaller than the number of nodes. Further, we have $|G_{SWk,m}| = n$ and (see Sect. 5.3 for the relevant definitions)

$$\bar{\lambda}_1 = \frac{\sin \dfrac{\pi(m-1)(2k+1)}{n}}{\sin \dfrac{\pi(m-1)}{n}} \tag{7.1}$$

and

$$\mu_{n+1-m} = 2k - \bar{\lambda}_m \,, \tag{7.2}$$

where $\bar{\lambda}$ and μ are the normalized Laplacian and adjacency matrices, respectively. Small worlds are hybrids between highly clustered networks and random networks.

The small-world network model has a problem finding a decentralized algorithm (whose decisions are based solely on local information) and can produce paths of small expected length relative to the diameter of the network. Moreover, as efficient as it appears to be, the small-world model of networks, especially social and biological networks, fails to answer an important question: how is meaningful information extracted from document streams? In a highly connected world, the problem is to defend oneself from wrong information, and to somehow select only good information. Kleinberg addressed these types of problem in his studies on algorithmic issues at the interface between networks and information theory, especially in social and information networks [213]. So the problem is to identify the 'most relevant' nodes (web pages, documents, immune responses, etc.) following a given broad query, and what is needed is an automated way to filter out the most 'authoritative' nodes. Kleinberg's solution is to understand and model the networks

as systems with an infinite-dimensional phase space, where bursts of activity are phase transitions.

White and Houseman demonstrated the existence of cohesive islands or communities with boundaries [216] by removing the bridges between clusters and making the cohesive clusters more apparent. They found that multi-level networks, including hierarchies of cohesive subgroups in networks like islands within islands, are critically related to issues of self-organization.

An interesting network dynamics occurs in some models trying to generate small worlds by an optimization process. Mathias and Gopal [217] introduced an optimization model for networks based on the minimization of a Lagrangian as a linear combination between the average network distance and the total length of all edges (like a normalized wiring cost). In contrast to previous work focusing on small-world behavior, they succeeded in showing that, by considering what arises as a result of the random rewiring of a few edges with no constraint on the length of the edges [185, 213–215] and subsequently introducing this constraint, they could obtain an alternate route towards a small-world network through the formation of hubs. While the node at each hub center serves to contract the distance between each pair of vertices within the hub, the introduction of a hub center into each neighborhood serves to maintain the clustering coefficient at its initially high value. The procedure used by Mathias et al. consists in starting with a regular network of some degree, using a certain high value of the wiring cost as phase transition parameter. As the wiring price is slowly decreased, some hubs emerge and grow in size and number. An increase in the range and number of inter-hub links is subsequently observed, together with a reduction in the number of hubs, and a trend toward the formation of a universal hub. The process is illustrated in Fig. 7.2. This study provides an interesting example of the shape and boundary variability of networks. Just by following certain optimization steps, the network changes dramatically from a typical ring shape for a regular network.

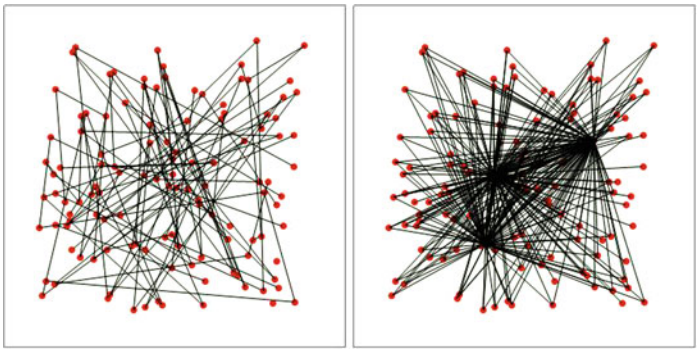


Fig. 7.2 Random graph with $n = 100$ and no hubs. Evolution of hubs for the same $n = 100$, but $k = 4$ optimized network. Note the emergence of hubs and their variation in size and number

Sporns and Honey, and Bassett et al. [218] have found strong evidence for the existence of functional networks exhibiting small-world attributes in the human brain. Using wavelet analysis on magnetoencephalography recordings acquired from human subjects, they obtained patterns of functional connectivity across a large number of recording sites. The correlations between signals in wavelet space express a statistical association between recording sites. The matrix of wavelet correlations obtained for different frequency bands was mapped to a binary matrix which was ultimately interpreted as an undirected graph and analyzed using network analysis tools that measure clustering, path length, centrality, and synchronizability. They found that the global topology of the functional networks at different frequency bands was both highly clustered and highly integrated, forming a small-world type of network.

7.3 The Shape of the Internet

Network models provide the conceptual tools for systematically and clearly representing social relations [219, 220]. Social networks can be studied, like any network, through graph theory methods. The graph model of the social network represents individuals or members by vertices and relationships between society members by edges. A fairly comprehensive yet concise review of quantitative methods for graph-based models of Internet topology can be found in [221].

As an example of a social network, see Fig. 7.3, where we represent the association between location and time of burglaries in a city. The upper graph represents areas where the time of day of the event was not relevant for the statistics. The lower four connected graphs represent burglaries happening during the night, and a few disconnected events during the mornings.

One question we would like to raise in this section concerns the boundary of the Internet as a (social) network. Is it as in Fig. 7.4, Fig. 7.5, or Fig. 7.6? And does it actually have a shape, or is it blurred?

In a recent monograph [220], the author explains that social network analysis shows that the associated graph model is usually clustered, as in the case of small worlds where everybody knows almost everybody through a few connections, but where the boundaries of the cluster are as yet unclear.

The main problem for a sociologist, psychologist, or biologist is to identify individuals and classify them according to their group identity on the basis of relational information alone. If this ‘kinematic’ task can be accomplished, the next problem is to find and explain the cohesion that binds a certain group together. There are procedures, and among them the topological approach, and especially boundary types of approaches, are the most efficient. In this section, we will discuss such procedures and results concerning the explanation of social cohesion and clusterization in social networks, based on their boundaries and separations.

First, we introduce elements of the Internet social network. As a graph or network, the web is not only a fascinating subject of academic study. It also yields

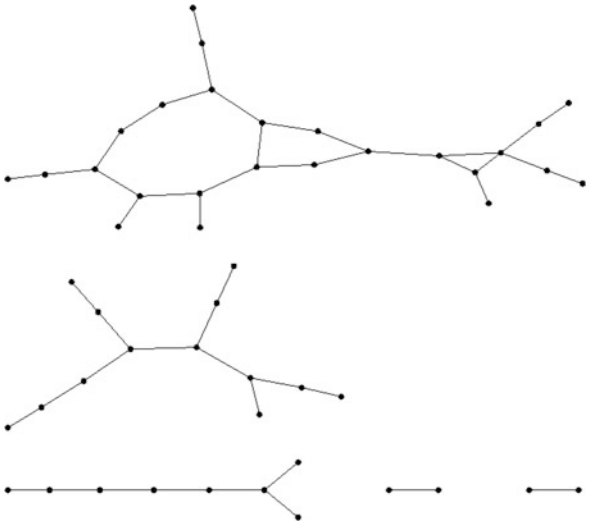


Fig. 7.3 Burglary crime analysis in Daytona Beach, FL 2010 [222], courtesy of DBPD and Master Thesis, Dan Antolosh 2012. The nodes represent burglary locations and the edges represent correlations by the same time of day for the crime. The *upper connected graph* represents burglaries during daytime, while the *lower four connected graph components* represent nighttime events in the same city

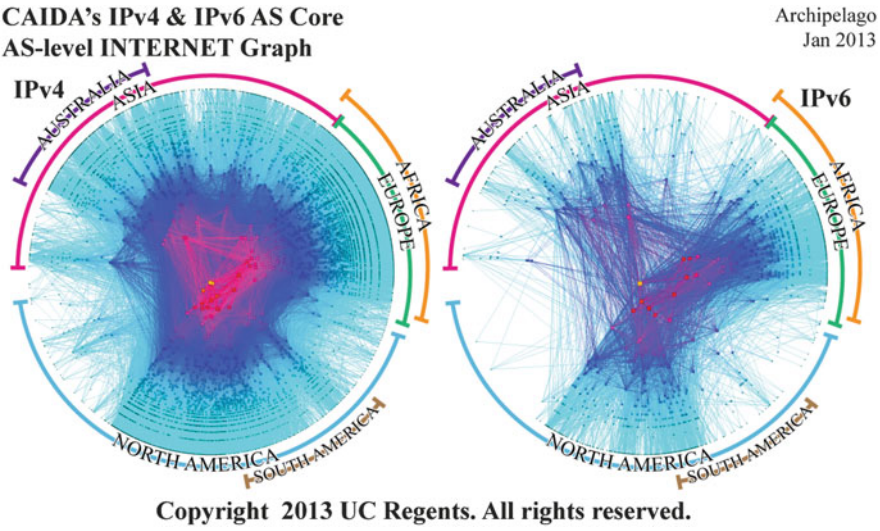


Fig. 7.4 Visual representation of the Internet topology at a macroscopic scale in 2013 for IPv4 AS Core. Courtesy of www.caida.org

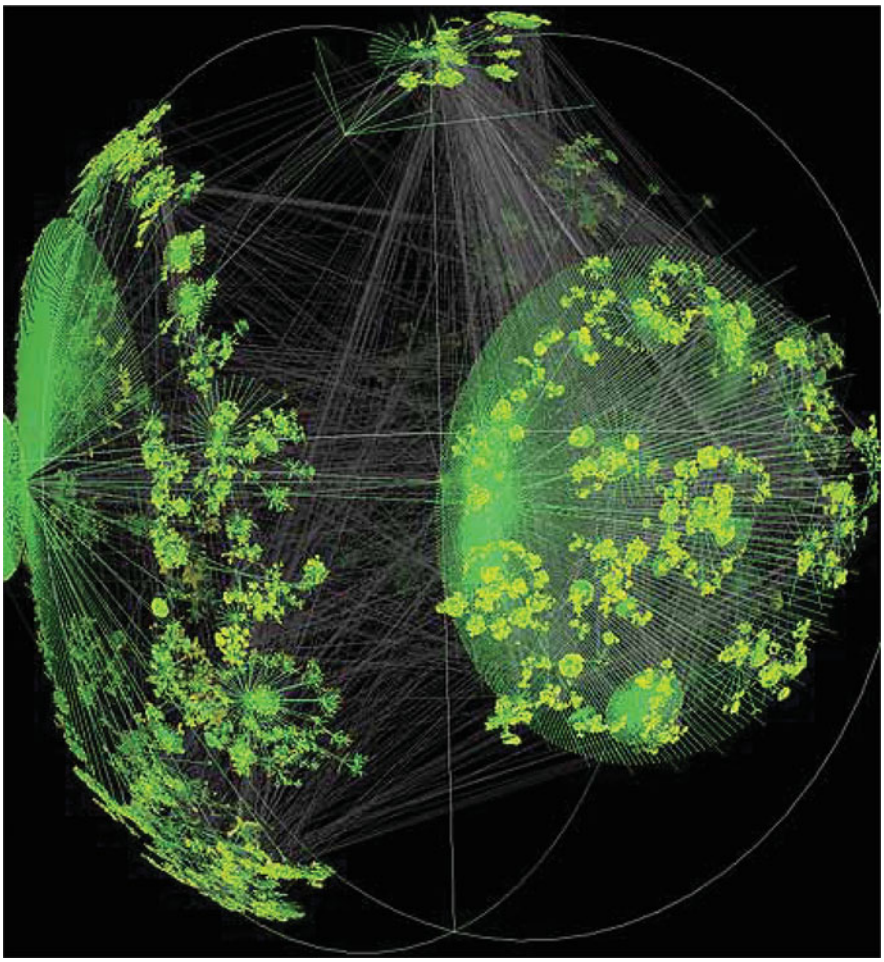


Fig. 7.5 Jelly fish aspect of the world wide web in 2013. Courtesy of www.caida.org

valuable insights into algorithms for searching and identifying communities, and sociological phenomena which characterize its evolution (see, for example, [178], or the recent analysis in [223]). For a very recent mapping of the Internet network, see Figs. 7.4 and 7.5. The connections between users and providers are modeled as branches of a tree. Models of the Internet social network mimic or are inspired by natural systems like blood vessels or river networks. The study of this type of network is not only of scientific relevance; it also addresses technological questions such as finding which cost function has to be minimized in order to improve net features. For the Internet social network, this should help us to extend wiring to developing countries, and improve the quality of net connections in countries already connected [210].

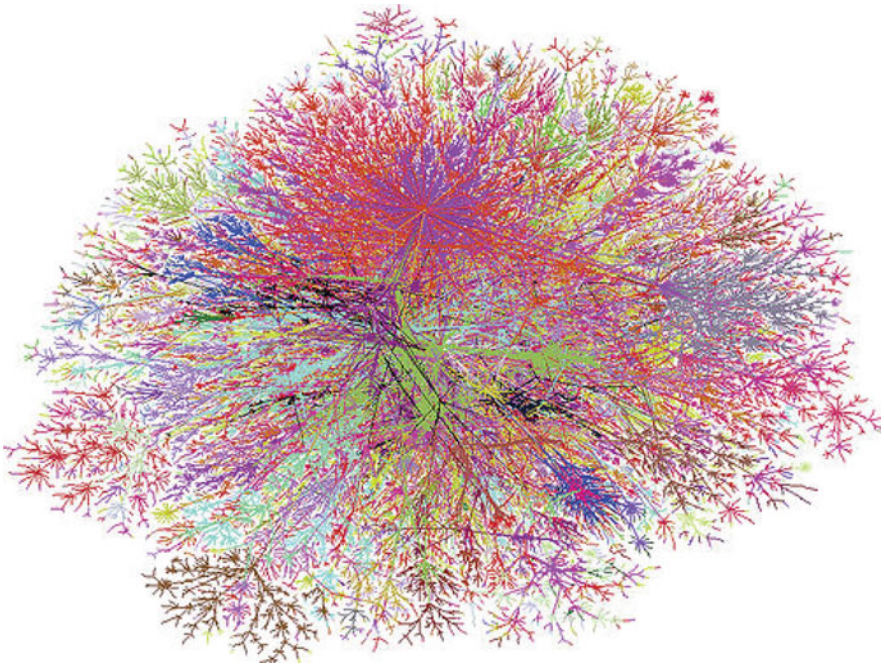


Fig. 7.6 Example of a graph modeling the Internet network in 2011. Courtesy of Steve Jurvetson

There are different models for the Internet network, including stochastic, fractal, small-world, scale-free, etc. (see Fig. 7.5). Almost all these models show the same complexity at different scales. These properties are defined between the size of a single node up to the whole network.

According to [177], the Internet can be modeled by a very large directed graph whose nodes are the documents, and whose edges are the links (URLs). The topology of the graph determines its connectivity, and consequently its internal effectiveness. Concerning the scale of web traffic, the Visual Networking Index estimates that by 2015, 3×10^{18} bytes (3 exabytes) of information will be transferred every day through the Internet, with a rate of increase of more than 10^{20} bytes/year (100 exabytes/year). Current estimates (at the end of 2013) evaluate the Internet as having about 50×10^9 web pages and 2.2–2.3 billion users, wherein it carries some 2×10^{18} bytes (2 exabytes) of information per day, citing the World Wide Web Foundation.

The continuously changing documents and links make it impossible to completely catalog all nodes and edges. With the data to hand, the Internet can be regarded as a graph with $|G| = 2 - 4 \times 10^{10}$, $\|G\| = 5 \times 10^{10}$, which results in an average length $\langle d \rangle \simeq 3$. It is also assumed that a web page is linked (on average) with 60 other pages. By comparison the human brain has around 10^{11} neurons, and each neuron fires on average thousands of other neurons. If the present rate of growth

of the Internet network is constant or even increases (see Fig. 7.9), it will not take long until its volume is one order of magnitude greater, with the possibility that the connectivity will grow by one order of magnitude, too. At that stage, the Internet will be only one order of magnitude less complex than the brain!

There have been many attempts to evaluate the large scale topology of the Internet network and associated graph. For example, in 1999, Barabasi et al. [177] measured the connectivity of the Internet by building a robot that collected all URLs found on a document and recursively followed them to retrieve the related documents and URLs. This research has shown that the probability of a document having k outgoing/incoming links follows a power law over several orders of magnitude. This law appears to be quite different from the Poisson distribution predicted by the classical theory of random graphs [224], but also from the bounded distribution associated with random network models, e.g., as found by Watts and Strogatz [185]. The power law matching their research has the approximate form:

$$\ln P(k) \simeq -49\,050 \ln k - 9.01 \ ,$$

where $P(k) \simeq P_{\text{out}}(k) \simeq P_{\text{in}}(l)$ represents both incoming and outgoing probabilities. Its tail indicates that the probability of finding documents with a large number of links is still significant. This also implies that the network connectivity is dominated by highly connected web pages. According to the conclusions presented in [177], and in spite of the apparent freedom of web page authors to choose the number of links on a document and the addresses to which they point, it appears that the Internet obeys a 'flocking' type of sociology based on scaling laws. These laws are characteristic of highly interactive self-organized systems and critical phenomena.

This model shows that the average distance is given by the empirical formula

$$\langle d \rangle = 0.35 + 2.06 \ln N \ ,$$

where $N = N(G_{\text{Internet}})$. This conclusion also places the Internet graph model in the category of small-world networks. However, the conclusion in [177] shows that the average distance in 1990 was in the range 20–40 and was slowly increasing, while the latest evaluations show that, in 2013, the average distance is even closer to a small-world model, namely, $\langle d \rangle \simeq 4$.

In a study in 2012 [225], the authors investigate the topological connectivity of the Internet, and compare it with the US highway and railroad systems. Their findings show that, if we are to compare the Internet network with other theoretically studied networks, the closest theoretical model would be grid and mesh networks, as opposed to star or tree networks. The authors used the normalized Laplacian spectrum of (5.6) and the multiplicities of the eigenvalues to compare the distribution of the normalized Laplacian eigenvalue densities versus the eigenvalue for some standard networks: star, linear, tree, ring, grid, and mesh. Comparisons were made between these spectra and the topological properties calculated from these spectra and the distribution for realistic networks such as Internet (AT&T, Sprint), US railroads, and US interstate. These findings further support the idea that

the topology of the 'logical' Internet network (the Internet connections between documents) is similar to the grid or mesh networks, while the 'physical' Internet network topology, as well as the topologies of the US railroads and highways are rather similar to bipartite networks.

A paper by Broder et al. [178] presents a study made by a team of computer researchers from northern California, in which they found, contrary to expectations, a significantly different topological structure for the web, unlike other traditional Internet network models like those supporting the jellyfish, small-world, ultra-small-world, scale-free [226], hub-dominated, *k*-shell [227], and other models (see, for example, Fig. 7.6). Their model became known as the *bow tie model*. It is based on studies of local and global properties of the web graph using 'AltaVista' crawls with over 200 million pages and 1.5 billion links.

The model divides the Internet network into four, almost equal in size, disjoint, and directed sub-networks: *strongly connected core*, *IN*, *OUT*, *tendrils*, plus a smaller number of *tubes* and even smaller disconnected components (see Fig. 7.7). Like the jellyfish model, there is a strongly connected core of about 56 million nodes and two other large groups, IN and OUT, of roughly 44 million nodes. IN consists of all pages that link to the strongly connected core, but have no links from the core towards outside, to reach back to these links. The counterpart to this is the OUT group, consisting of all pages that the strongly connected core links to, but which have no links back towards the core. The fourth group, containing 44 million nodes,

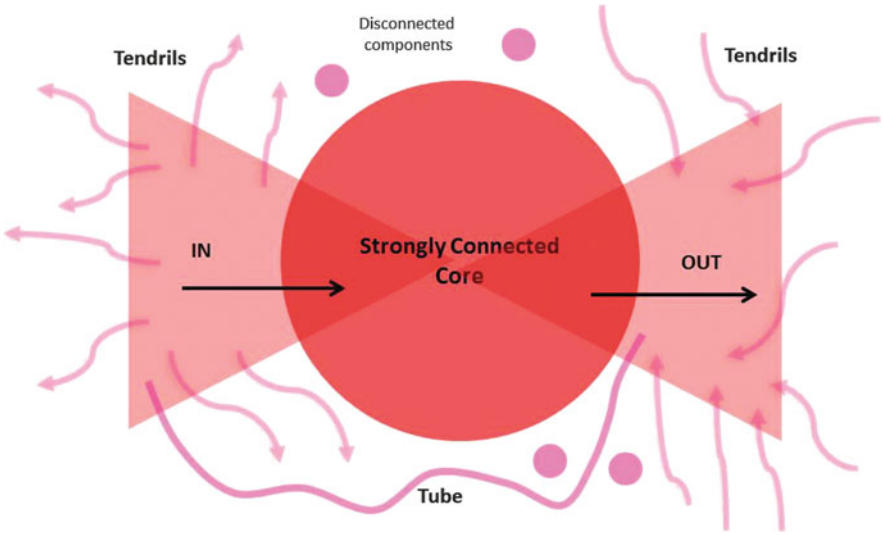


Fig. 7.7 The 'bow tie' model for the connectivity of the web [178] has a strongly connected central core, bounded by two directed IN and OUT subgraphs. These finger out into many tendrils containing nodes that are reachable in the opposite direction from the central directions. Some tendrils coming out of IN may be hooked into others entering OUT, thereby forming a passage from IN to OUT without using the core (a tube)

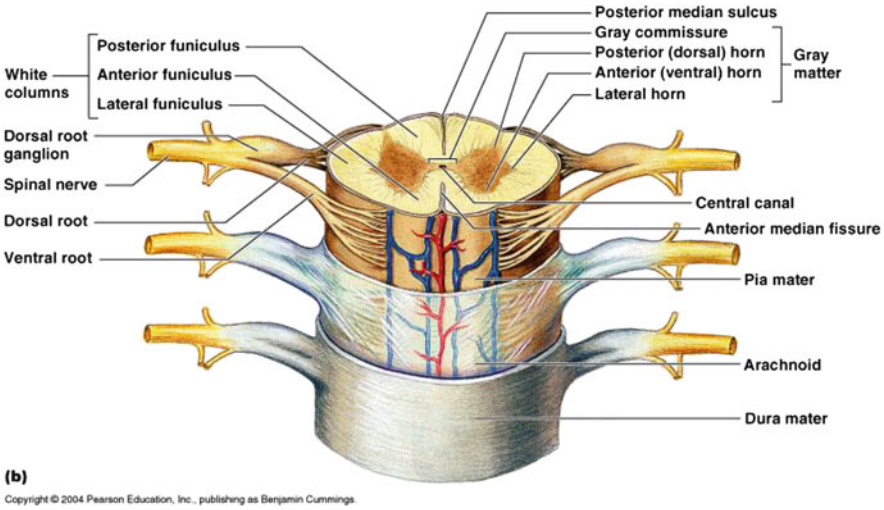


Fig. 7.8 Cross-section of the spinal cord and its connection to the spinal nerves. In one human spinal cord, there are about 32 million neurons, almost the same number of nodes as are found in the central core of the bow tie model of Internet. From http://biol251.wikispaces.com/Ch_13-SpinalCord (http://creativecommons.org/licenses/by-sa/3.0/)

is represented by all other disconnected pages that neither link to the core, nor are pointed at from the core. They can be represented as directed tendrils that always run towards OUT and away from IN (see Fig. 7.7). Some tendrils can accidentally grow and reconnect outside the core, allowing information to flow, but not through the core. There may also be a certain number of smaller, completely disconnected groups.

The procedure followed to obtain this result was accomplished by studying the power law distributions of the nodes and pages. According to Broder et al., two interesting and useful insights were obtained from the analysis. First, the connectivity of the Internet is extremely resilient and does not depend on the existence of hubs or high degree nodes. Second, the high rank nodes (highly ranked web pages, also called 'authorities') were almost all found to be embedded in a very well connected core.

The shape found by Broder et al. [178] for the Internet network has a strong visual similarity with the cross-section of the spinal cord (see Fig. 7.8). Moreover, in one human spinal cord segment there are almost the same number of neurons (32 million) as nodes in the central core of the bow tie model of the Internet.

It is worth noting that the fundamental topological and geometrical properties of the Internet network like its shape and non-homogeneity and its self-similar or scale-free structure do not restrict or even significantly control the dynamics of information, friend selection, or stability and efficiency of the network. Large networks are nevertheless complex systems and their features cannot really be inventoried in a topological or geometrical nutshell. For example, the work of

Kaltenbrunner focuses on the impact of the geometric distance of online social interaction and shows that, even if geometric distance strongly constrains the way social links are established, there seems to be a uniform effect on all user interactions, unrelated to the geographic length they span [228]. A similar conclusion was obtained by Hu et al. in [229], showing that the distribution of geographic distance between friendship is inversely proportional to the geometric distance. In other words, no matter what the topology, the spatial structure of the Internet social network is scale invariant, and the network is equally and uniformly navigable. It may be true that people tend to make friends if they are a shorter distance from each other, but once a remote friend is acquired at a greater distance, this connection will be equally stable.

Results on scale-free characteristics of the Internet network and signatures of its self-organization are put forward in [230]. Here it was found that the probability f that a document has d outgoing (incoming) links follows a power law over many orders of magnitude. In graph theory language, this means that the fraction of nodes with degree d has power law dependence over the whole of its domain $N(G)$:

$$f(d) \sim \frac{1}{d^{\eta}} ,$$

where η is a positive constant in the range 2.1–3.0. Moreover, the authors found, in agreement with several other studies, that the average distance between pairs of nodes follows the law

$$\langle d \rangle = C_1 + C_2 \ln |G| ,$$

where $C_{1,2}$ are constants to be determined by extrapolation data, indicating that the web forms a small-world network of a kind known to characterize social or biological systems. In 2000, Barabási et al. found $\langle d_{\mathrm{www}} \rangle = 19$ [230]. A consequence is that, whenever the network grows, there will be preferential attachment, because there is a higher probability that a new node will be linked to a node that already has a large number of connections.

Facebook is the largest online social network ever used by people to articulate existing offline social connections, as well as to forge new ones [231]. It enables its users to present themselves with an online profile, accumulate ‘friends’, and view each other’s profiles. Users can also join virtual and common interest groups. Today, Facebook has more than 1.2 billion registered users and it is the space on the web where the most networking takes place with respect to total page views (see Fig. 7.9).

In [232], the authors examine the use of an online social networking site created by Michigan State University students and its relationship to social capital formation and integration into college life. The authors showed that participants of a social network that is geographically delimited will be less likely to play with their identities (and therefore to verify others’) due to the bounded nature of the site [232]. The comparative study of the history and properties of this

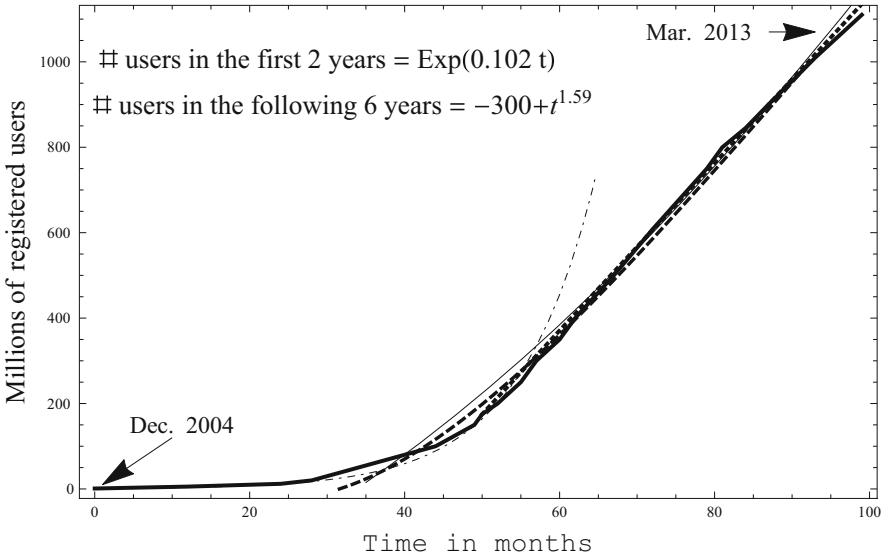


Fig. 7.9 The time dependence of the number of registered *Facebook* users versus time over a period of 8 years is shown by a *thick solid line*. The *thin dot-dash curve* represents an exponential model which fits the first part of the dynamics. The three curves (dashed, thin solid, and dotted) represent interpolations for the late time interval, namely, polynomial, power law, and linear, respectively

network's connectivity showed that the number of stable connections doubled in time owing to the existence of geographic limitations (all members belong to a bounded community where almost everybody knows everybody). In that situation, the existence of superimposed physical (space, location) boundaries on top of a virtual social network can enhance the openness of the exchange of information. Consequently, a bounded network doubled by another channel of communication enhances the network bound and a nonlinear cooperative effect can result.

When online and offline social networks overlap, the direction is from online to offline: online connections result in online meetings; connections made online rarely stay there, and we may wonder whether this is one cause for its instability.

If an online social network serves or is defined through a geographically based community, then its bound coincides with the community's bounds.

Below we list three of the most important parameters that can be used to describe quantitatively an online social network:

1. *Temporal evolution of the network's nodes and connections.* This is the study of the average lifetime of use of a node or a connection, measured by the amount of transferred information, the time spent by a node when the network is used, the use of various network features, the purposes and motivations for using the network, and the perceived critical mass around a certain node or connection. Several studies show that the temporal evolution of one node is

history dependent, e.g., when Facebook users keep their high school nickname as a replacement for their legal name on their Facebook account. In addition to these features, one may consider the measurement of users' perceptions that the people with whom they communicate are also social network users.
2. *Local parameters around a node.* The need for networking and the need for (bridging and bonding) social capital.
3. *Global parameters.* Bridging social capital allows diffusion of information or links to external assets.

In [232], the authors found that members of bounded online social networks (like the one based on geographical association) have a tendency to use links to offline connections rather than grow bridging social capital. It may be that unbounded online social networks have a tendency toward 'poor get richer' as opposed to 'rich get richer', as usually happens in other types of social networks. For example, users with low self-esteem tend to increase their bridging social capital by using online networking rather than offline (geographic) networking.

In a recent article [223], the authors show that the neighborhood function $N(h)$ representing the percentage of user pairs that are within n hops of each other has the form of a sigmoid function of Gompertz type on a logarithmic scale:

$$N(h) = 100 \tanh(0.022 h^{0.698}) \, .$$

7.4 Internet Is a Boundary

Let us model the Internet as a network with a very high degree of connectivity, a hypothesis which works if the network is of small-world type. This also means that almost all nodes are separated on average by the same number of steps. The question is whether it is possible to embed this network in a high-dimensional Euclidean space $\mathbb{E}^n$ in such a way that the average Euclidean distance between two nodes is given by the average number of steps between these nodes. In such a space, almost all points representing the network are placed at the same relative distance, like the 4 vertices of a regular tetrahedron in $\mathbb{R}^3$.

Let us consider a system of N points P_i in $\mathbb{E}^n$ such that the distance $d(P_i, P_j) = d_0$ between any two of them is constant for any $i, j = 1, \ldots, N$. This constraint determines a system of $N(N-1)/2$ equations. It has a solution if the following inequality holds:

$$Nn - n - \frac{n(n-1)}{2} \geq \frac{N(N-1)}{2} \, ,$$

where on the left we have the maximum number of free parameters of N points in a space with n coordinates, minus the arbitrariness of the network origin position, minus the maximum number of possible rotations in the space under which the

geometric structure of the network is invariant. From the solution of this inequality, it follows that the maximum number of points that can be placed at equal relative distance in $\mathbb{E}^n$ is given by $N = n + 1$. This result also shows that, if we want to provide a geometric meaning for the average number of steps between any two nodes, for a very large, highly interconnected small-world type of network, we must embed it in a space with a number of dimensions equal to the number of its nodes minus one. On the other hand, if we place all the nodes inside a spherical hypersurface in $\mathbb{E}^n$ of radius $R \sim N$, we evaluate the ratio between the number of nodes placed on the hypersurface N_Σ and the total number of nodes inside the hypersphere N_V using the formula

$$\frac{N_\Sigma}{N_V} = \frac{k}{N^{1/n}} ,$$

where k is a constant independent of the number of dimensions of the space, and of the order of unity. Since all the nodes are separated by the same average distance and we have $n \sim N \to \infty$, it follows that the right-hand side of the last equation approaches a constant limit k. This in turn means that the fraction of nodes on the left-hand side approaches k, which is almost unity. We are forced to the conclusion that, for a network with such a large number of nodes, almost all the nodes which are separated by the same distance (defined as the number of steps in $\mathbb{R}^3$ and Euclidean distance in $\mathbb{E}^n$) are placed on the surface of a hypersphere of radius n in a space with n dimensions. Consequently, such a network, and this includes the Internet, is its own boundary if the number of nodes increases indefinitely. The Internet is its own boundary.

Chapter 8
Big Data Systems

In the Spring, I have counted 136 different kinds of weather inside of 24 hours

Mark Twain

When your search for hotels and your website becomes busy with promos for car rentals, or when your smart-phone app identifies your location and you receive offers from nearby restaurants, you are in the middle of big data science and data mining in action. In the last decade, the handling of vast amounts of information shifted from the sole responsibility of astronomers and high-energy physicists to that of the life scientist. Starting with fields like meteorology, petroleum exploration, and astronomy, the essential jump in the emergence of big data came in 2003, when the first human genome was completed and a new type of science, big data science, came into being [233].

The sequencing centers, high-throughput analytical facilities and individual laboratories, produce vast amounts of nucleotide and protein sequences, protein crystal structures, gene-expression measurements, protein and genetic interactions, and phenotype studies. For example, a new field of research emerged in biology, viz., *biocuration*, defined as the activity of organizing, representing, and making biological information accessible to both humans and computers. The same situation happened in social networking, in business and communications, where big data science became useful to identify fraud in real time, to score medical patients for health risk, to identify consumer sentiment landscapes, or to explore network relationships. The *Human Connectome Project* collects advanced imaging data in order to explain how spontaneous activity in the brain correlates between different brain regions. The *Citizen Science* or *Galaxy Zoo* projects approach big data in a different manner. For example, in astronomy, *Galaxy Zoo* relies on statistics, multiple viewers, and logic to process and check data by feeding images to volunteers who do basic classifications of galaxies. If the proportion of viewers who agree on the classification of a certain galaxy is constant as more and more people see it, the galaxy would be withdrawn from viewing.

It is estimated that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions. In 2012, almost every day, Google alone processed 24 petabytes of data, but only a small amount of this information is formatted in conventional databases. To have a feeling for the magnitude of this flow of data,

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_8

we mention that the European Bioinformatics Institute in Hinxton, UK, one of the largest biological data repositories in the world, stores all in all about the same amount of data (20 petabytes $= 2 \times 10^{16}$ bytes), and this number almost doubles every year [234]. The exponential growth in the amount of available data in science and media (biology, sociology, knowledge networks, etc.) means that revolutionary measures are needed in the science of curating data, data management, data mining, analysis, and accessibility.

A petabyte is the equivalent of about 20 million filing cabinets' worth of text. To comprehend this amount of information in terms of a more tangible concept, we can convert it into substantial mass. Consider this 24 petabytes of information stored at some point on a magnetic device. The average volume of such a storage device would be, with today's technology performance, about $V \sim 0.05$ m^3. Consider that the bytes are stored through magnetic dipole–dipole interactions in some high quality magnetic material, and consider that writing this information is performed using a spin-transfer torque technology. If the blank memory is taken to have zero potential energy reference level, the encryption of the information will result in an almost random sequence of binaries across the memory. We can evaluate the total magnetic energy needed to write this information to be about 10 GeV. For comparison, this is how much electromagnetic radiation we receive in one second through one square meter on the Earth's surface from the star Alpha Centauri. If we convert this energy into rest mass through mc^2, we obtain the mass equivalent of 10^{10} carbon atoms, which is about the average mass of one human cell.

Big data science, especially when it employs the methods of exploratory data mining and cluster analysis, is becoming more and more useful in many fields, including bioinformatics, DNA microarray technology, information retrieval, pattern recognition, image analysis, and machine learning. The topic is encountered especially in the field of clustering high-dimensional data, where the data set can reach several thousands of dimensions. A straightforward example is provided by clustering of text documents where a word-frequency vector lies in a space with a number of dimensions equal to the size of the vocabulary [235]. Data mining is the set of software algorithms that aim to extract information from huge sets of data.

8.1 Data Dimensionality

The dimensionality of a data set is, loosely speaking, the minimum number of independent variables needed to represent the data without information loss. According to a definition introduced by Fukunaga in the early 1980s [236], a data set $\Omega \subset \mathbb{R}^d$ has *intrinsic dimensionality* $ID = m < d$ if its elements lie entirely within an m-dimensional subspace of $\mathbb{R}^d$. The effective computation of ID can be done using the Fukunaga–Olsen algorithm, which is a local method. If the data vectors were embedded in a linear subspace, ID as defined above would be equal to the number of non-zero eigenvalues of the covariance matrix. In order to use this property, one has to divide the data set into a *Voronoi tesselation* by means of a

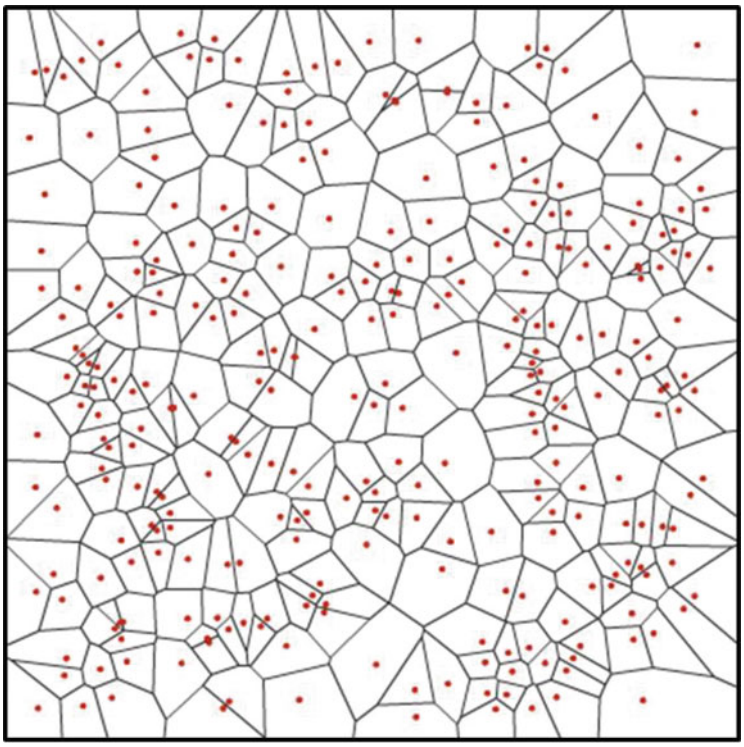


Fig. 8.1 A Voronoi tesselation

clustering algorithm. We remind the reader that a Voronoi tesselation is a special partition of a given domain in a metric space, determined by the initial selection of a collection of L points in the domain, $P_k, k = 1, \ldots, L$. For each k, we define a Voronoi set R_k, $k = 1, \ldots, L$, as the set of points in the domain whose distance to the point P_k is not greater than the distance to any other point $P_j, j \neq k$, from the collection (see Fig. 8.1). In each such Voronoi set, the surface in which the vectors lie is approximately linear and the eigenvalues of the local covariance matrix are computed and normalized by dividing them by the largest one. Then *ID* is obtained as the number of normalized eigenvalues that are larger than a chosen threshold.

As for the clustering algorithms acting on a given set of data points, this involves grouping them into disjoint subsets called clusters, based on some chosen equivalence relation called similarity, and usually inspired by a metric criterion. For example, a hierarchical clustering algorithm starts by considering clusters of the points themselves, and then repeatedly combining the two nearest clusters. The nearness criterion is the distance between the centroids of the clusters, and the clustering algorithm may stop when the points to be added to a given cluster are farther than a given limit. The process continues until some optimization criterion (such as the Bradley–Fayyad–Reina criterion) is satisfied, and the clusters are

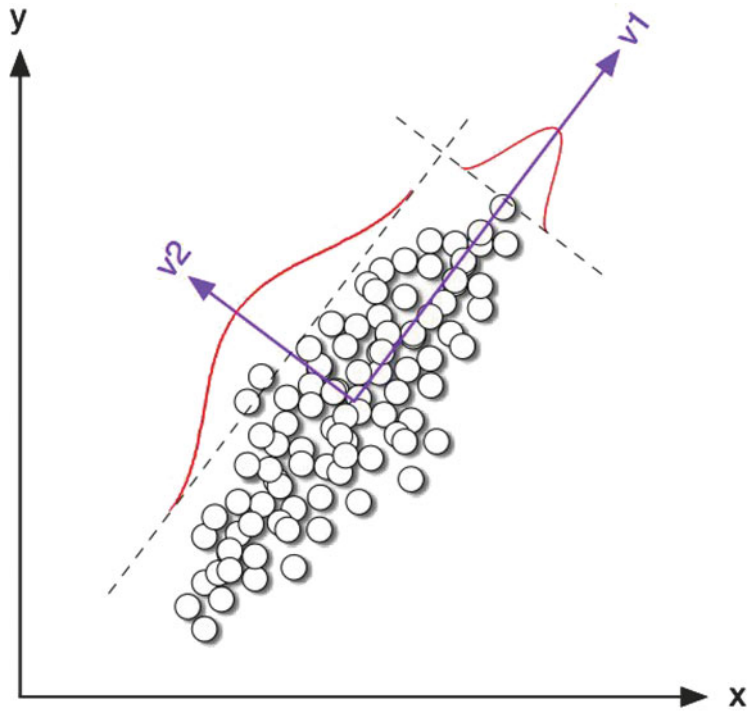


Fig. 8.2 Application of an NLPCA procedure to a 2D data set. The directions $\boldsymbol{v}_{1,2}$ are the principal directions of maximum and minimum data variance

'cohesive' (that is, they satisfy a certain criterion of connectivity, or density, or centroid distribution).

In contrast to local methods, global methods for evaluating *ID* can be divided into projections, scaling, and fractal methods. The most efficient projection method is nonlinear principal component analysis (NLPCA) [236]. The method is not at all new in mathematics or physics, but the computer and software science community has given it a different flavor when applied to large sets of data. The data table regarded as a high-dimensional tensor, or generalized matrix X, is mapped into coordinates of a high-dimensional affine space, and the resulting set of affine points is studied using the idea of principal axes of inertia for a 3D solid (see Fig. 8.2). In this high-dimensional space, a program finds the direction of the largest possible data variance, and then orthogonal directions are constructed. Since the quantity to be maximized can be expressed as a Rayleigh quotient, an eigenvalue procedure can be used. The eigenvector corresponding to the largest eigenvalue is chosen to be the principal direction, and the procedure continues with the other non-zero eigenvalues.

This procedure can be explained very simply as follows. Let $\boldsymbol{v}$ be the direction in which X has its maximum variance, that is, $\boldsymbol{v}^t X \boldsymbol{v}$, which is the variance in direction $\boldsymbol{v}$, must be maximum. Since $\boldsymbol{v}$ is a unit vector, we need to find extremals of the

functional

$$\boldsymbol{v}^{\mathrm{t}} X \boldsymbol{v} + \lambda(\boldsymbol{v}^{\mathrm{t}} \boldsymbol{v} - 1) \ ,$$

where λ is a Lagrange multiplier. Differentiating, we obtain

$$X\boldsymbol{v} - \lambda \boldsymbol{v} = 0 \ ,$$

which is the eigenproblem to be solved. As a practical application, a data set can be embedded in a high-dimensional linear space, a number of principal directions may be constructed (smaller than the space dimension), and the coordinates along these principal directions become the new parameters of the data set. A reduction in the number of degrees of freedom can then be identified. The procedure is similar in some ways to the procedure of finding the nonlinear normal modes of oscillations (NNM) of a coupled nonlinear mechanical system. The normal frequencies of such a dynamical system depend on the energy and on the initial conditions, because the system is nonlinear. An energy–frequency chart of all possible modes of oscillations can then be mapped and the normal modes identified.

The multidimensional scaling method (MDS) is an ordination method, i.e., a procedure for mapping data as points in a space with axes chosen so that the clusterization of data becomes evident and manifest. This procedure is more closely connected to data visualization techniques. Among possible applications, we may mention scientific visualization in cognitive science, psychophysics, psychometrics, marketing, and ecology, and ordination of autonomous wireless nodes in real time.

In the following, we present an example where this method has been successful. Traditionally, the nuclear chart is understood by plotting boxes, each of which represents a unique stable nuclear species, in the periodic table of elements (see Fig. 8.3). No clustering or other structure of data is visible in this traditional representation. However, in nuclear physics, the boundaries for nuclear particle stability are conceptualized as drip lines. Inspired by this phenomenon, plotting nuclear data on a graph where the number of neutrons is increasing on the abscissa (horizontal axis) and the number of protons is increasing along the ordinate (vertical axis) generates a totally different landscape. It is easy now to observe the tendency of the nuclei to cluster along the drip line of nuclear stability. Figure 8.4 is a good illustration of the advantage of using the NLPCA method illustrated in Fig. 8.2, and confirms the importance of this method.

Fractal-based techniques are global methods that have been successfully applied to estimate the attractor dimension of the underlying dynamic system generating time series [236]. In contrast to other global methods, they can provide non-integer values as *ID* estimates, which is why these methods are called fractal. From nonlinear dynamics, box-counting and the correlation dimension are the most popular methods that have been imported into big data global methods.

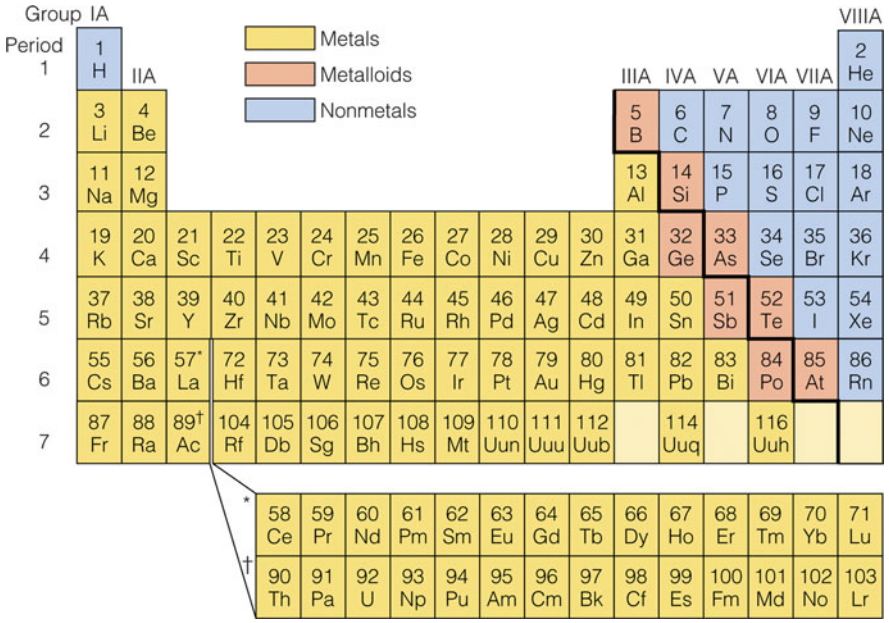


Fig. 8.3 Representation of stable nuclei (atomic nuclei) in the traditional periodic table of elements. The data show no observable clustering tendency when presented like this

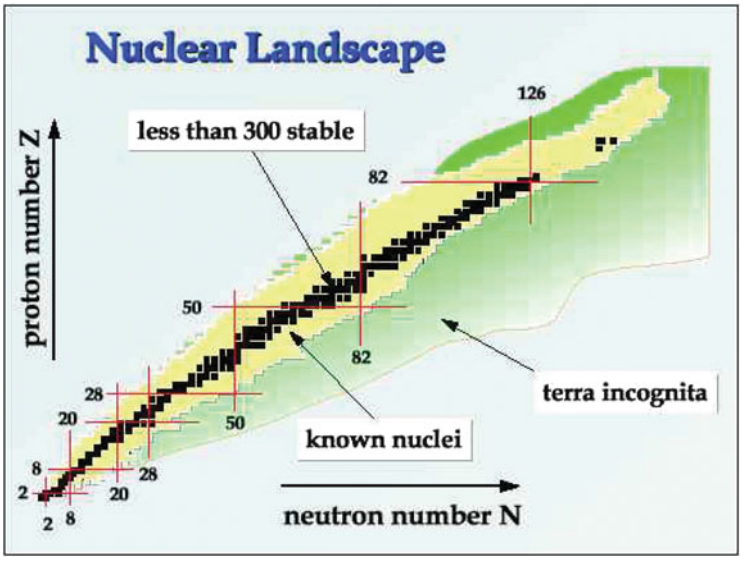


Fig. 8.4 The same nuclear data as in Fig. 8.3, ordered according to their nuclear stability along the neutron drip line. The *dark squares* are stable nuclei, known in nature. *Yellow*: Domain of nuclear metastability. *Green*: Domain of exotic, superheavy nuclei and new types of nuclear matter

8.2 Topology of Big Data: Persistent Homology

The use of algebraic and differential topology and geometry tools in a given scientific field does not need any justification. The richness of operations and structure in geometry has had a powerful influence in science, and at the same time the specific development to meet the needs of physical sciences has led to the identification of new and useful geometrical structures. In particular, algebraic topology, and its more recent cousin, computational algebraic topology, have found a wide field of applications in problems of feature detection and shape recognition in high-dimensional data.

Among other essential mathematical tools used in big data investigations, one important procedure goes by the name of *persistent homology*, i.e., the application of homology theory to point-cloud data sets and the representations of this algebraic characterization using *barcodes*. In the introduction to his paper [237], Ghirst asks the rhetorical question: "How does a topologist visualize a four-dimensional object?" and he writes his response in the form of a Socratic rejoinder: "How do you visualize a three-dimensional object?" Neuroscientists studying the early childhood brain have shown that, in the first year of life, we learn how to infer 3D spatial data from paired planar projections. Years of practice, continues Ghirst, have tuned a remarkable ability to extract global structure from representations in a strictly lower dimension. As a peripatetic would say: *Nihil est in intellectu quod non prius in sensu*.[1]

Another possible answer to the above question is relative homology, Künneth theorems, and cobordism. Put simply, if we know the topology of a lower-dimensional space, that is, if we know all the homology and homotopy groups classifying its structure, number and types of holes, etc., then we can use one of the Künneth theorems to calculate the homology groups of product spaces between these lower-dimensional spaces and a line segment. This is pretty much as simple as constructing a 3D cylinder from the product between a 2D disk and a 1D segment. This is followed by factorizing (shrinking) the cylinder lids into points, and thereby obtaining a 3D disk from a similar 2D one. Both the disk and the segment are contractible, whence a certain chain of homology groups becomes exact and provides the required homeomorphism, and consequently the desired transfer of topological properties (homology) from 2 to 3 dimensions. In this way, a 4D disk (and a 4D sphere as its boundary) can be imagined as the inflation of one point in the 3D space into a 3D sphere, up to a certain radius, followed by its retraction back to a point, in time, for example. In this example, we embed the line segment of the fourth dimension in the time axis, as the fourth axis.

Data is represented in general as an unordered sequence of points in some Euclidean space, $(x_\alpha) \in \mathbb{E}^n$. This data can be generated by time series from sensors, point cloud data from scanned 3D objects, motion capture data, etc. The

[1] There is nothing in the intellect that has not previously been in the senses.

collection of points will generate the vertices of a combinatorial graph whose edges are determined by proximity, e.g., two vertices placed at a distance less than a certain threshold will generate an edge. The graph is then completed to a simplicial complex (see Definition 8 in Sect. 5.5).

Starting from the simplicial complex generated by the data cloud, we can generate a continuous family of topological spaces parameterized by a positive number ϵ whose topological properties change in a discontinuous manner with ϵ. The vast majority of these topological properties appear and disappear with increasing ϵ, and are called topological 'noise'. If, however, some of the algebraic topology properties of this family of spaces remain unchanged over a broad range of variation of ϵ, we call these properties 'persistent' and they can pretty accurately describe the hidden reduced dimensionality of the data set.

To put this simply, imagine a large number of points in the plane placed very close to a circle, i.e., the mean distance from the points to the circle is much smaller than the circle radius. Then start to draw little disks of radius ϵ around each point, considering the intersection of all these disks as a simplicial complex, and keep increasing ϵ. Initially, the resulting set has a disconnected topology, but eventually some of the disks overlap and generate here and there some particular shapes. However, beginning from a certain value of ϵ, and up to very large values of ϵ, all circles overlap in an annulus shape, so the algebraic topology of the resulting simplicial complex is pretty much unchanged over a broad range of values for ϵ. We say that the data topology has a persistent property. It takes very large values for ϵ, compared to the radius of the original surrounding circle, to close the annular simplicial complex into one disk. Since for this large range of ϵ values the data topology remains the topology of the circle S_1, we can say that these data have $ID = 1$.

Based on the data cloud of points $X = \{x_\alpha\} \in \mathbb{E}^n$, and for any $\epsilon \geq 0$, we introduce two types of complex. The *Čech complex* C_ϵ is defined as the abstract simplicial complex whose k-simplices are determined by unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ for which closed $\epsilon/2$-ball neighborhoods have a point of common intersection. The Čech theorem states that C_ϵ has the homotopy type of the union of closed $\epsilon/2$-balls about the point set $X = \{x_\alpha\}$. This means that C_ϵ behaves like a subset of $\mathbb{E}^n$: a point cloud fattened by balls [237].

The *Vietoris–Rips complex* R_ϵ is defined as the abstract simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ that are pairwise within distance ϵ. Two overlapping disks form a 2-simplex (a segment), 3 overlapping disks form a 3-simplex (a triangle), and so on (see Fig. 8.5).

Consider a sequence of Vietoris–Rips complexes R_i associated with a data cloud for an increasing sequence of parameter values ϵ_i. There are natural inclusion maps

$$\iota : R_1 \hookrightarrow R_2 \hookrightarrow \cdots \hookrightarrow R_i \hookrightarrow \cdots \hookrightarrow R_{N-1} \hookrightarrow R_N \ .$$

This sequence generates the maps $\iota_* : H_k R_i \hookrightarrow H_k R_j$, for all $i < j$, where H_k is the homology group of a certain order k. We have the following Lemma [237, 238]:

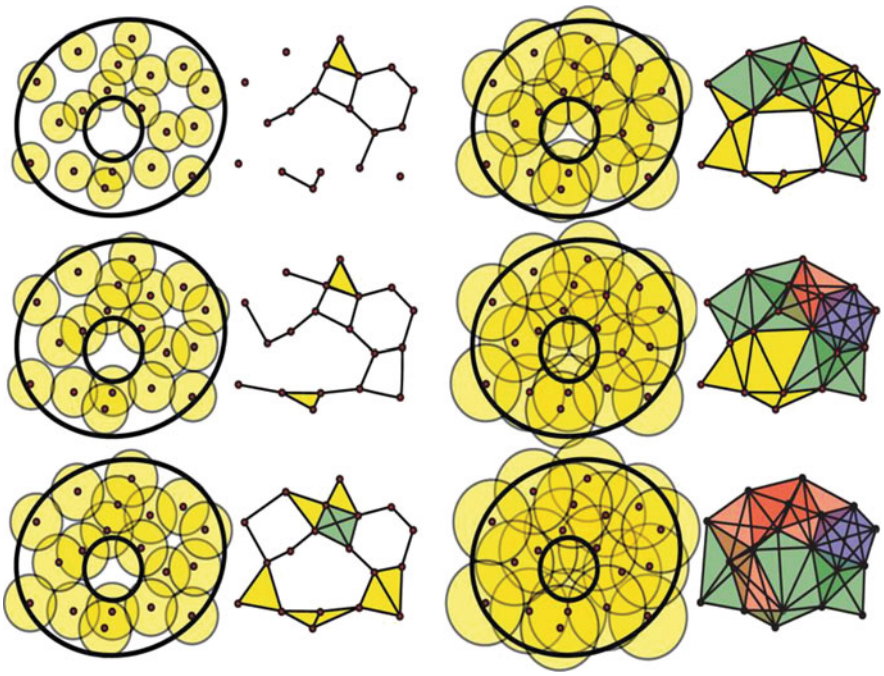


Fig. 8.5 A sequence of Vietoris–Rips complexes for a point cloud data set representing an annulus. Upon increasing ϵ, holes appear and disappear. The only persistent structure which truly describes the data cloud is the upper right C_ϵ. This has the same homology as the circle. All the other complexes for the other values of ϵ contain noise holes. R. Ghrist [237]. Reproduction by courtesy of Bull. Am. Math. Soc.

Lemma 1 *For any $\epsilon_i > 0$, there is a chain of inclusion maps*

$$R_i \hookrightarrow C_{\sqrt{2}\epsilon_i} \hookrightarrow R_j \,, \tag{8.1}$$

with $\epsilon_j = \sqrt{2}\epsilon_i$.

Lemma 1 shows that the topological features that persist under the inclusion map $\iota : R_i \hookrightarrow R_j$ are in fact shared with the topological features of C_{ϵ_j} if $\epsilon_j \geq \sqrt{2}\epsilon_i$. Consequently, the homology of the inclusion $\iota_* : H_k R_i \hookrightarrow H_k R_j$ reveals information that is not otherwise easy to notice from $H_k R_i$ or $H_k R_j$ alone.

By way of illustration, Robins [239] shows how to quantify these topological features using the concept of ϵ-*persistent Betti numbers*. We define

$$\beta_k^{i,j} = \text{rank}\,(Z_k R_i) - \text{rank}\left[\iota_*(Z_k R_i) \cap B_k R_j\right] .$$

In other words, the ϵ-persistent Betti number $\beta_k^{i,j}$ is the number of non-equivalent, non-bounding k-cycles in $H_k R_j$ that are the image of a k-cycle from $H_k R_i$. Geo-

metrically, $\beta_k^{i,j}$ is the number of holes in R_i that do not get filled in by taking a coarser-grain ('fatter') R_j (see again the sequences in Fig. 8.5).

Robins' final theoretical approach in determining the topological structure of the data cloud is to use the inverse limit systems of shape theory and the Čech homology [240]. In order to accomplish this goal, she generalizes Lemma 1 by introducing, instead of $\epsilon_j = \sqrt{2}\epsilon_i$, a continuous parameterization for the finer grain $\epsilon_j > \epsilon_i$. The next step is to choose two sequences satisfying

$$\epsilon_i \to 0 \,, \quad \text{and } \epsilon_i < \epsilon_j \to 0 \,,$$

and to evaluate the limit $\beta_k^{i,j} \to \beta_k^{0,j}$.

These $\beta_k^{0,j}$, called 0-persistent Betti numbers, describe the homology of R_i that genuinely comes from the homology of X, and we have $\beta_k^{0,j} \leq \beta_k^{i,j}$ for any $\epsilon_i \leq \epsilon_j$. From the continuity of the Čech homology [237, 239, 240], we also know that the 0-persistent Betti numbers of R_i converge to those of the original data cloud space, i.e., $\beta_k^{0,j} \to \beta_k(X)$, if X is compact. In order to apply this hole analysis based on persistent Betti number theory to data sets given by finite point patterns, one has to 'fatten' or coarse-grain the set by overlaying a digital mesh and attaching spheres of radius ϵ_i at each point. An appropriate level of coarse-graining can be chosen on the basis of physical reasons. The Betti numbers are then computed using a triangulation that has the same topology as the R_i complex.

Figure 8.6 reproduces such an analysis result from [239] on a fractal set of points obtained by an iterated function system, i.e., similarity transformations of a unit square with contraction ratio $1/2$. The fractal shown in the left frame of Fig. 8.6 is disconnected, and consists of infinitely many line segments. Topologically, therefore, there are no loops in this fractal. However, the homology of the Vietoris–Rips ϵ-complex creates holes in the ϵ-neighborhoods.

The disconnected nature of the data is seen in the staircase growth of β_0 (the blue curve in the right-hand frame of the figure), which counts the number of connected components. In the limit $\epsilon \to 0$, this Betti number approaches the exact number of line segments in the original fractal data cloud. The graph of the number of holes, i.e., β_1 (the red curve), shows that more holes are resolved as $\epsilon \to 0$. One can see towards the $\epsilon \to 0$ limits that three holes remain in the data set, and they are shown by the persistent β_1 Betti number in the right-hand frame.

In Fig. 8.7, we consider a simplicial Čech complex constructed on a collection of 416 points randomly distributed in the Euclidean plane around a figure-eight knot, which has $\beta_1 = 5$. Different insets in the figure represent, from left to right, increasing values for ϵ-ball neighborhoods. The curves represent the Betti numbers calculated for these complexes, namely, β_0 in blue and β_1 (scaled 10:1) in orange. The straight lines represent the exact values.

The main shape is concentrated around this knot, so we have $ID = 1$. The β_0 parameter pretty well describes the evolution of the complex, and ranges from almost the total number of data points to nearly 1, the topological characteristic of a path-connected curve. However, when we compute the loop homology H_1 of this

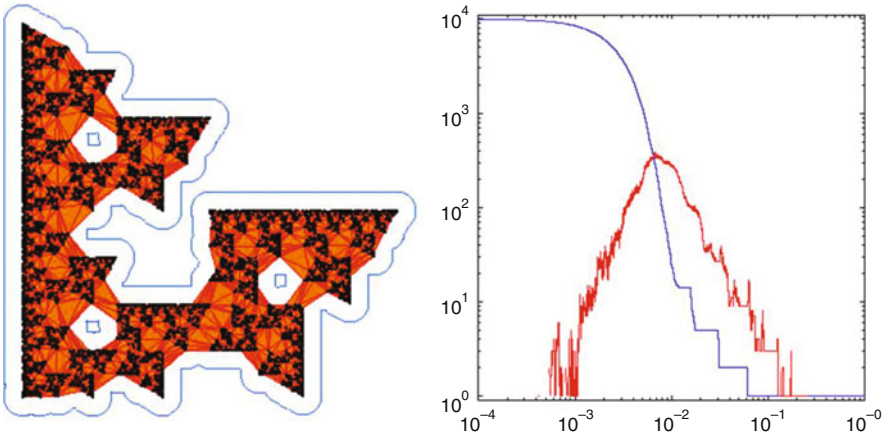


Fig. 8.6 *Left*: Example of a Vietoris–Rips complex with $\epsilon \sim 0.1$, associated with a disconnected fractal made of 10^4 line segments. The original data cloud has no loops. However, the $\epsilon \sim 0.1$ complex creates three artificial holes. *Right*: Betti numbers β_0 in *blue* and β_1 in *orange*, on logarithmic axes. There are spikes due to the geometry of the fractal, but they are non-persistent holes. The three artificial non-persistent holes are also revealed by the *orange curve*. Reproduced from K.R. Mecke and D. Stoyan (Eds), Lect. Notes Phys. **600** (2002) p. 274, by courtesy of Springer-Verlag

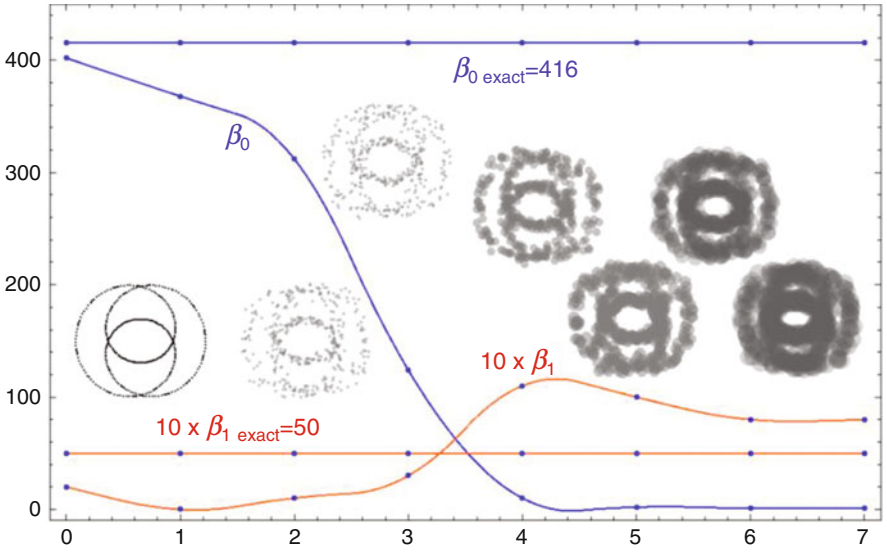


Fig. 8.7 Examples of different ϵ-complexes built around 416 points that describe a figure-eight curve plus white noise. In the main window, we plot the number of connected components $\beta_0(\epsilon)$ in *blue* and the number of holes $\beta_1(\epsilon)$ in *orange*. Homology persistence occurs only after exceeding a certain threshold in ϵ

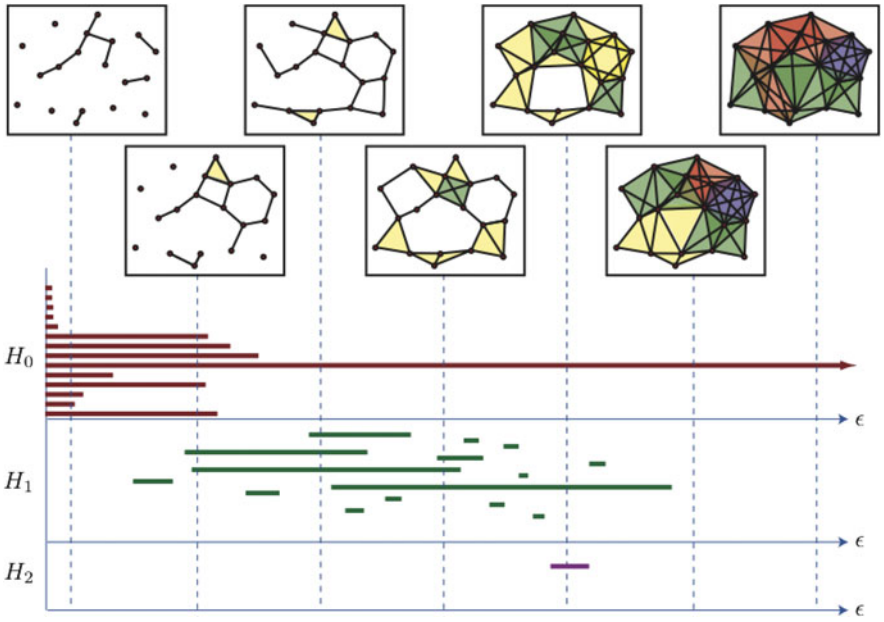


Fig. 8.8 Annulus shape data. When ϵ varies, more simplices are generated (see new *colored triangles*) and topological features like connectivity, Betti numbers, or holes appear or disappear. The features that persist for a long stretch are considered to be real; the small ones are called topological noise. This type of analysis is called persistent homology, and its representation in the bottom frame is called the 'data barcode'. The Betti numbers, i.e., ranks of H_kR_i, equal the number of intervals in the barcode for H_kR intersecting the (*dashed*) line $\epsilon = \epsilon_i$. Image from [237], courtesy of Bull. Am. Math. Soc.

complex for small ϵ values (left part of the graphics), it yields Betti numbers β_1 oscillating between zero and two, randomly including noise holes and too sparse to circumvent the real loops of the figure-eight knot. In the middle range for ϵ, the computational homology provides more accurate values for β_1, even intersecting its correct value of 5 loops. There is still the presence of noise, as can be seen for larger radii of the disks when $\beta_1 \sim 12$. Intuitively, we regard this large number of holes as coming from faulty sampling or other errors in the recovery of the data. Towards even larger ϵ, the β_1 curve tends to stabilize, and most likely asymptotically approaches an almost correct value. In principle, one could then argue that a correct choice of the parameter ϵ generates a complex with the right connectivity structure.

The dependence of the Betti numbers on the parameter ϵ inspires a visual snapshot in the form of a 'barcode' (see Fig. 8.8), that is, a graphical representation of H_kR_i as a collection of horizontal line segments in a plane whose horizontal axis corresponds to the ϵ parameter and whose vertical axis represents an (arbitrary) ordering of homology generators. Figure 8.8 gives an example of barcode representations of the homology associated with point sampling in an annulus.

So far we have presented families of simplicial complexes parameterized by a single parameter ϵ, but there may be other situations when multiple parameters can prove useful, e.g., if we study the probability distribution which gave rise to a point cloud [238]. One way to do this is to estimate the density function using a *density estimator*. The new parameter τ will be a percentage parameter, and $X(\tau) \subset X$ is the subset of points which lie within the τ th percentile of density as measured by the given density estimator. Then it is interesting to study the topological evolution as τ changes.

8.3 Topology of Big Data: Regions with Holes

There may be issues when applying *persistent methods* in cases where big data cloud shapes are not directly topological. If the cloud can still be embedded in a differential manifold, one can study the filtration on the manifold through the value of the scalar curvature, and the evolution of the topology of the level sets of this filtration can reflect interesting properties of the shape (see [238, 241] and references therein). In such cases, one needs to find discrete versions of the curvature and to use *multiple persistence* theory based simultaneously on a geometric quantity related to curvature, for example, and the scale parameter ϵ. Such cases are investigated using a mathematical formalism based on graded modules, but the analysis becomes too deeply involved in algebraic calculations, and hence strays too far from the purpose of our study of the boundary of big data sets in this chapter.

Here, we present an example where simpler, non-homological topological methods can help us to understand and classify big data sets. The Geographic Information System (GIS) is a geoinformatics computer/software network and institution which deals with models of geographical reality and describes geographical objects and spatial relations of the kind 'San Marino is surrounded by Italy' or 'the South Florida Everglades graminoid marshes are straddled between sawgrass marshes and sloughs'.

Sometimes homology methods are not sufficient, and in order to study geographic systems or ecosystems, one should use different scales, i.e., different topologies, for one may consider the effect of one boundary on another (e.g., surge or flood effects on rural populations in south Louisiana, or the effects of coherent oscillations in car traffic at peak times between two large neighboring areas in Houston, etc.), or the effect of one species on another, on a group of other species, or on the whole ecosystem. Network analysis must also be involved in the GIS studies in order to compare ecosystems with different spatial extents, and to understand the different patterns and dynamics that arise (for a review, see [242] and references therein).

The topology of these objects is in general irregular since it contains topological singularities as holes and separations. In order to study the topological relations between such 2D objects with arbitrary holes, connected boundaries, and connected interiors, a few bivalent computational topology models have been developed [242].

Among them, a traditional approach is the '4-intersection' model, in which regions containing holes are generalized using the union between the region and the interior of its holes, and the closure of each hole. Hence, for a region A with holes $h_i \subset A$, $i = 1, \ldots, n$, we define the *generalized region*

$$A^* = A \cup \left(\bigcup_{i=1}^{n} \bar{h}_i \right) . \tag{8.2}$$

The topological relation between two such generalized regions with holes, A^* and B^*, is expressed through the 4-intersection matrix defined by

$$\begin{pmatrix} \partial A \cap \partial B & \partial A \cap \overset{\circ}{B} \\ \overset{\circ}{A} \cap \partial B & \overset{\circ}{A} \cup \overset{\circ}{B} \end{pmatrix} , \tag{8.3}$$

where the circle over a set denotes the interior (an open set). The entries of the matrix in (8.3) can be empty or non-empty, so there are eight possible cases. Egenhofer et al. have shown in [242] that, for two regions with n and m holes, respectively, the total number of topological relations that can be specified between the objects (regions and their holes) is $\mu = (m + n + 2)^2$, and this number includes redundant relations, too. The next step in the analysis of large data clouds associated with such regions, each determined by large sets of points in the plane, is to develop algorithms to minimize the number of necessary relations. The software works using a language based on words describing the only independent possibilities, i.e., disjoint, meet, equal, inside, contains, covers, covered by, and overlap.

The literature in the field of information geometry, and specifically in the computational topology and geometry of big data sets, is vast and increasing by the week. More and more methods are being developed and it is difficult to keep track of them all. A very good source of inspiration for such models is [123], for example. This book lists the main open trends and questions to be tackled, including bivariate families, neighborhoods of Poisson randomness and the duality between independence and randomness, cosmological voids and galactic clustering, amino acid clustering in protein chains, cryptography and signal clustering, stochastic fiber networks, stochastic porous media and applications in hydrology and industry, and quantum chaology.

To end this chapter, we may quote Steve Lohr's wise and documented words from his article *The Age of Big Data* in The New York Times of 11 February 2012. He asks: "So what is big data?" We can add another question: "And how is big data connected to boundaries?" For sure, Lohr says, big data is a marketing term, but also a term in technology that helps us to find new approaches to understanding the world and making decisions. Data increases all the time, more than doubling every two years, and this does not only concern more streaming, but new types of data, e.g., the continuous improvements and inventions created to link sensors in cars, houses, weather stations, airplanes, offices, etc., to computing intelligence.

There is an increasing realization of the enormous potential of data-driven computational social science. The availability of unprecedented amounts of data about human interactions in different social spheres or environments opens the possibility of using those data to leverage knowledge about social behavior. Big data also presents many formidable challenges to governments and citizens, precisely because data technologies are becoming so pervasive, intrusive, and difficult to understand [235]. This is what Adam Jacobs calls *The Pathology of Big Data*. Naive or brute-force incorporation of large-scale data into simulation models may not lead to the expected results in terms of achieving progress in social science. While it is apparent that analysis of the data will certainly contribute to our understanding of mechanisms, it is also clear that further input will often be needed, in particular input obtained from mathematical, geometrical, topological, and other models. As a rapidly developing and successful field, computational social science must be aware of the need to develop its theoretical premises [233].

Lohr comments that "one witnesses the rise of what is called the *Internet of Things* or the *Industrial Internet*". Improved access to information also contributes to the growth of big data itself, for data is not only becoming more available, but also more understandable to computers.

Chapter 9
Physical Boundaries

When you see a fish you don't think of its scales, do you? You think of its speed, its floating, flashing body seen through the water . . . If I made fins and eyes and scales, I would arrest its movement . . . I want just the flash of its spirit . . . What is real is not the external form, but the essence of things . . . it is impossible for anyone to express anything essentially real by imitating its exterior surface.

Constantin Brâncusi

What Brâncusi intended was not to diminish the importance of shape and boundaries. He probably meant that there is more to a boundary than just the appearance of its shape (see Fig. 9.1). The shape can actually represent, through its freedom and its intangible majesty, the essence of reality. Shape is important, it is even essential, and sometimes, as Brâncusi said, it is hard to distinguish the reality it describes, as we can see in Fig. 9.2. Antoine de Saint Exupéry said (*Citadelle* 1948): "A rock pile ceases to be a rock pile the moment a single man contemplates it, bearing within him the image of a cathedral." In other words, even though the content, the internal structure, and the density may be the same, what makes the difference here is the external shape.

In this chapter, we shall describe some examples illustrating the importance of the free boundary, or interface, for the physics of liquids described from a differential geometry point of view. For more detail on the importance of free boundaries for liquid masses, one can consult [121, Chap. 3].

Regular liquids have three interesting properties of great interest for pure and applied mathematics: they experience a wide class of geometrical deformations, they can have free surfaces, and inside their free surface boundaries, they exhibit the property of being incompressible to a very large extent. Incompressibility together with the existence of boundaries results in one of the most fascinating phenomena in the mathematically-explained world, namely, a system of a given dimensionality (and complexity) can be explained by a part of lower dimensionality (complexity), i.e., the overall body dynamics can be inferred from the dynamics of the shape alone. In the following sections, we will ignore the existence of phase changes, or chemical or electromagnetic interactions.

The dynamics of a liquid drop is modeled by considering the internal liquid matter, mostly incompressible, surrounded by its boundary, which performs

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_9

Fig. 9.1 Two artistic representations of a fish. *Left*: Chinese statue of Arowana fish. *Right*: Fish (bronze, metal and wood sculpture, 1926, Tate Modern, London) by Constantin Brâncui. Public domain: https://commons.wikimedia.org/wiki/Category:Constantin_Br%C3%A2ncu%C5%9Fi#/media/File:%27Fish%27_by_Constantin_Br%C3%A2ncusi,_Tate_Modern.JPG

Fig. 9.2 When art copies nature: which sculpture represents what natural object? *Clockwise from top left*: Crystal glass vase, bohemian crystal vase, cut glass bowl, Swarowski pineapple art object, 50 μs collapse of a cavity in water created through impact of a disc, and a crystal vase. The water cavity image is reproduced from Fig. (1b) from reference [243] with permission by the authors and the journal

oscillatory or rotational motion mediated by surface effects, plus external field forces (gravitational, centrifugal, electric forces, magnetic traps, etc.). Interaction of the drop surface with surrounding fluids, or solids like supporting surfaces, containers, or pipes, etc., can also be considered. Even if it does not quite qualify as a complex system, a drop is a complicated nonlinear system. If the macroscopic and microscopic drop properties are combined in one interactive model, a drop can become a complex system, e.g., coupling the fluid mechanics of the drop with its

thermal properties. This way one can easily obtain complexity signatures like pattern generation, fractal surfaces, chaotic behavior, and symmetry bifurcations. If the drop geometry is coupled with external non-mechanical fields, as in the Ginzburg–Landau model for mesoscopic superconductors where vortices can interact with the drop geometry, the resulting system is complex, too.

One feature common to all these drop models is the existence of a surface tension which always acts as restoring force against various shape deformations and internal flows. Surface tension arises in any drop model, from the Planck scale to the cosmological scale, due to the imbalance of forces between the internal constituents of the drop and those at the surface. Particles in the interior receive equal forces from the surrounding particles, whereas the surface particles receive a net force directed towards the interior of the material. The droplet tends toward an equilibrium if the forces between the particles are balanced, but if the droplet is displaced from this equilibrium, oscillations and waves occur as the forces adjust to move the droplet back towards equilibrium.

The interesting behavior of drops consists in the huge variety of shapes they can adopt during this tendency to restore equilibrium, viz., spherical, axisymmetric drops, or multi-lobed drops. A deep understanding of the existence and coexistence of these galleries of shapes results from a deep understanding of the mathematical insights given by the drop model.

When studied in bounded forms, liquids with free boundaries can be modeled by smooth and compact geometrical surfaces which endow them with the privilege of being described by very rich mathematical objects like spectral theory, bounded operators, mode expansions, functional variational principles and stability criteria, etc. We can divide bounded liquid bodies with free boundaries into two main classes: films and drops (where we also include liquid shells, bubbles, bubble clusters, and antibubbles). Liquid films can simply be modeled by surfaces spanned by curves, so even one more step down in complexity from two dimensions to just one. The system under study is 2D and it has a rigid curve as imposed boundary. Drops and bubbles are 3D systems with 2D free boundaries. The boundary is compact, and in general, a homeomorphism of embedded spheres and/or tori, unless they merge into clusters [244].

Liquid film and drop dynamics include interesting phenomena like oscillations (or irregular vibrations or waves crossing their bulk and boundary), rotations, topological transitions (transitions from simple connections to annular shapes), breaking (transitions from connected to disconnected domains), and splashing (transitions in the dimensionality of the supporting manifold). Consequently, we organize this chapter into the following sections: liquid drops in general, 3D drops, and 2D drops.

In order to cover the dimensionless analysis of hydrodynamic equations in this book, Table 9.1 presents the dimensionless ratios between the most important driving terms occurring in the dynamical equations of fluid mechanics (electromagnetic, chemical, and thermodynamic terms are not considered). In the columns and rows, we enumerate expressions proportional to these driving terms: the hydrostatic pressure P, gravitational force $\rho g h$, dynamical pressure (kinetic energy density)

Table 9.1 First order dimensionless numbers in fluid mechanics

	P	ρgh	ρV^2	$\nu V\rho/L$	$\Omega^2 L^2\,\rho$
ρgh	To				
ρV^2	Eu	Bos, Fr, Ri			
$\nu V\rho/L$	Poi	$\frac{\text{Ga}}{\text{Re}}$	Re		
$\Omega^2 L^2\rho$		η	Rigidity	Ek, Wo, Go, $\sqrt{\text{Ro}}$	
$H\gamma$	La	λ_c, Bo, Eo	We	Capillary, $\frac{\text{Mg}}{\text{Re}}$, $\frac{\text{Re}}{\text{La}}$, $\frac{\text{Re}}{\text{Oh}}$	Pt

We consider only macroscopic mechanical effects, and do not include here thermal (equilibrium or non-equilibrium), electromagnetic, chemical, or elastic effects

ρV^2, viscous force $\nu V\rho/L$, centrifugal force $\Omega^2 L^2\rho$, and surface tension $H\gamma$. At the intersection of the columns and rows, we identify the dimensionless number associated with the relative importance of the corresponding driving terms.

The symbols listed in the table represent the dimensionless numbers in fluid mechanics as follows:

- To = Torricelli law
- Eu = Euler number
- Bos = Boussinesq number
- Fr = Froude number
- Ri = Richardson number
- Ga = Galilei number
- Re = Reynolds number
- Ek = Ekman number
- Wo = Womersley number
- Go = Görtler number
- Ro = Rossby number
- La = Laplace number
- Bo = Bond number
- Eo = Eötvös number
- λ_c = capillary length
- We = Weber number
- Capillary = capillary number
- Mg = Marangoni number
- Oh = Ohnesorge number

The dimensionless number $\eta = R\Omega^2/g$ represents the ratio of the maximum centrifugal and gravitational accelerations, as given, for example, in [245]. We also introduce a rigidity term for the rotation of a deformable body through the ratio between how much its rotation has the same angular velocity everywhere (rigid body rotation) and how much its rotation follows other rules, e.g., gravitational or electrostatic rotation, etc. We refer to the drag as the ratio between the drag force and the driving pressure. An example of a second order dimensionless ratio of driving terms is Archimedes' number defined by $\text{Ar} = \text{Ri} \times \text{Re}^2$.

The only dimensionless numbers which are not often described in the literature, but which we would like to introduce here are the *Poiseuille number* Poi and the *Plateau number* Pt defined as follows. The number

$$\mathrm{Poi} = \frac{\gamma P}{\nu V \rho / L} \tag{9.1}$$

represents the ratio between the pressure drop at the ends of a cylinder, and the drag force combined with the flow rate of the fluid through the pipe, at hydrodynamic equilibrium, i.e., the Hagen–Poiseuille law. The Plateau number is defined by

$$\mathrm{Pt} = \frac{\gamma H L}{\rho \nu V} = \frac{\mathrm{Mg}}{\mathrm{Re}}\,, \tag{9.2}$$

and represents the ratio between the surface curvature pressure generated by the surface tension and the centrifugal force density. The Plateau number describes the relative contribution of the surface tension effects versus the centrifugal effects in a rotating drop. A small value for the Plateau number indicates a large centrifugal effect that exceeds the restoring tendency of the surface tension and introduces high instability, super-deforms the drop, or even breaks it. In practice, a large Plateau number indicates an almost spherical drop where the rotation effects can be neglected as compared with a stronger surface tension.

The dynamics of liquid droplets displaced from equilibrium are of interest not only from a purely scientific point of view, but also for the impact they have on human-related activities, various industrial applications, and natural physical processes. Rainfall [246], air pollution [247], printing [248], painting [249], pharmaceutics [250], mixing [251], flying [252], vortices in distillation [253], absorption [254], fermentation [255], liquid–liquid extraction [256], and spray drying [257] are only a few of the phenomena and operations where drops play a primary role. Drop physics is important at every scale, from heavy nuclei models [258] and exotic radioactivity [259], to Bose–Einstein condensate drops [260], quark matter droplets and quark–gluon plasma drops [261, 262], to nanofluidics [263] and labs-on-a-chip [264], to swimming motile cells [265] and cell division [266], to oceanography [267] and tsunami studies [268], when parts of the ocean can be modeled as drops, all the way to astrophysics [269] and neutron stars [270]. Less traditional structures have also been observed like double bubbles, antibubbles, walking bubbles, bubble clouds, etc. [271–278].

9.1 Geometry of Inviscid Fluids

A natural way to model pattern generation and wave propagation phenomena is through conservative models, i.e., based on local conservation laws from which solutions can be derived using symmetries. A typical example is the study of

conservative partial differential equations where the powerful geometric methods of Hamiltonian mechanics work very efficiently. By extension, many attempts have been made to apply the methods of Hamiltonian mechanics to problems from incompressible fluid dynamics (a detailed discussion is provided, for example, in [279]), thanks to the simplicity of the fluid equations in such a case.

Traditionally, the fluid equations can be written in a closed form in terms of velocity, density, and entropy (with pressure given as a function of entropy and density through some equation of state) in the Eulerian frame of reference. In the case of classical mechanics, the dynamics of all dependent variables (all particle paths) are coupled so it is impossible to reduce the configuration space to a smaller subset. In contrast, in the case of incompressible flow, the Eulerian field formalism provides an extraordinarily simple tool for the study of many-particle systems. This advantage has its roots in a special symmetry property of the fluid particle Hamiltonian, namely its invariance under relabeling particles with the same density and entropy.

The particle-relabeling property corresponds to a Noetherian conservation law which actually expresses vorticity conservation. It is thus interesting that all the results obtained using vorticity theorems and formalism are deeply related to the existence of an approach using an Eulerian reference system. At the same time, the vorticity laws (like the circulation theorem or free surface boundary conditions) describe the properties, and the locations, of labeled fluid particles, which relate the concept of vorticity to the Lagrangian frame approach. This example suggests the existence of an (epistemological) principle of uncertainty for fluid dynamics, even at pre-quantum levels. As a consequence of this principle of uncertainty, the price to pay when choosing one frame rather than another is a matter of personal choice.

However, there is a certain bias: a Hamiltonian approach brings many advantages. We may mention succinctness of the formalism, the connection between symmetries and conservation laws, and independence of the form of the laws with respect to any special choice of coordinates. Unfortunately, there is a serious impediment when one tries to use the traditional Hamiltonian method (as inspired by classical mechanics) in fluid dynamics: the Eulerian variables are not canonical. It is geometric Hamiltonian mechanics, together with its new developments, which allows the use of Hamiltonian structure without the need for canonical variables [280–284]. As Salmon writes in his introduction to [279]: “From the geometric viewpoint, the statement that noncanonical (e.g., Eulerian) variables are sometimes useful even though the underlying dynamics is Hamiltonian is closely analogous to the more obvious statement that non-Cartesian (e.g., spherical) coordinates are sometimes useful, even though the underlying geometry is Euclidean.”

During the last few decades of the twentieth century, a considerable amount of literature was published by a number of researchers: Bridges, Benjamin, Marsden, Montgomery, Ratiu, Salmon, and Shepherd, to name only a few [167–169, 279–281, 285]. They use a theoretical approach based on differential geometry for the equations of water waves and outline the critical role of multi-Hamiltonian and multisymplectic structures, especially when associated with the equations of incompressible fluids with free boundary. Such structures contain variational

principles in a natural (geometric) way, and this is essential for the study of pattern formation and wave instability [169], as mentioned earlier.

The main result of all these contributions can be synthesized in the observation that the classical problem of incompressible flow is equivalent to a purely geometric problem of finding a path of minimum length in a multi-dimensional 'landscape' that conserves local volume. Mathematically, this problem can be expressed as the geodesic problem for smooth bijective volume-preserving maps defined on a smooth manifold. These diffeomorphisms can be interpreted as transformations of positions of particles lying in a manifold, under local volume conservation. The diffeomorphisms can be composed, like any other maps, and such a composition law induces a Lie group structure in the space of the diffeomorphisms. The connection between hydrodynamics and the construction of the Lie group of diffeomorphisms is made by specifying an invariant metric on the tangent space to the manifold, and considering the corresponding volume element induced in the manifold by this metric. In other words, for a given smooth vector field defined on the manifold, the geodesic flow of this field satisfies the Euler equation for an incompressible fluid on this manifold. The pressure becomes a scalar field that can be obtained from this flow. The geodesic flow is tangent to the boundary of the manifold, if the manifold has a boundary, and the flow parameters can be adjusted to satisfy smooth initial conditions.

This intriguing but straightforward hydrodynamics construction in terms of Lie groups enables a unified approach to an even broader variety of dynamical systems, from the equation of rotation of a rigid solid to the Navier–Stokes equations for hydrodynamics. There is, for example, a famous application of this theory to the question of non-existence of long-term reliable weather forecasts. Arnold's estimates [286], related to curvatures of diffeomorphism groups, show that the weather is essentially unpredictable after two weeks, as the error in the initial condition grows by a factor of 10^5 over this period.

The topological approach to incompressible fluid dynamics [284, 285] contrasts with that of classical Hamiltonian fluid dynamics by using a Lagrangian frame formalism. The theory begins by defining a configuration space M which is a domain with a smooth boundary embedded in the 3D Euclidean space where the fluid particles move. In general, M can be a smooth compact Riemannian n-dimensional manifold with smooth boundary, and it is called the *reference fluid container*. Because of the incompressibility property, the initial domain occupied by the fluid is in one-to-one correspondence with the cloud of fluid particles at any moment of time, so one can label the particles by their initial positions. The shape of the fluid domain changes, but its local and global volume forms are conserved.

In the following, we construct a fiber bundle formalism for the topological fluid dynamics theory (see Sect. 4.5). Figure 9.3 presents a sketch of the concept. The base space will be $M \times \mathbb{R}$ made of *material points* representing events in space and time $(\boldsymbol{x}, t) = (x^i, t) \in X$. On top of X, we construct a fiber bundle Y whose standard fiber is the manifold M itself, this time labeled by coordinates $(y^i) \in M$ called *spatial points*. The canonical projection is $\pi^{-1}(x^i, t) = y^i = x^i$. The coordinates of a point in the Y bundle are (x^i, t, y^j). In this way, the base space is the reference (or

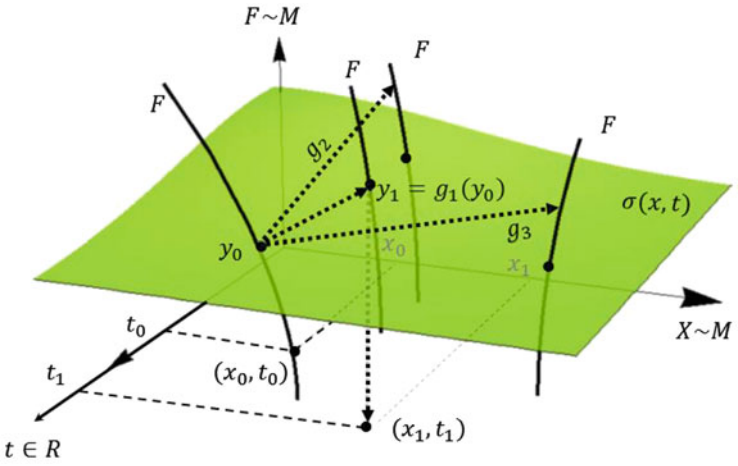


Fig. 9.3 The Euler point of view in the flow bundle. Particles are labeled by the axis $X \sim M$, and $t \in R$ is the time axis. The vertical axis is vertical fiber direction and different F curves are different fibers. At t_0, the particle labeled x_0 occupies the position $(x_0, t_0) \in X$ and $y_0 \in \pi^{-1}(x_0, t_0) \sim F$. We let this particle go, but make measurements at the same space position at different moments of time. We apply $g \in G$ transformations to y_0 and some of the images $g(y_0)$ intersect other fibers at the flow $\phi(X)$, e.g., at $y_1 = g_1(y_0)$. We project $\pi(y_1)$ on X to obtain the new particle labeled x_1 which passes through the same fixed space position y_0 at t_1, and so on

initial) configuration of fluid particles, and its extra dimension in the base labels the time evolution. The bundle Y becomes the particle placement field and any smooth section $\phi : X \to Y$, $\phi^j(x^i, t)$ in the bundle describes later configurations of the fluid. The sections in this bundle represent all possible physical flows. The tangent space to each point of a section is called the *first jet bundle* and it is denoted by $J^1TY_{\phi(\boldsymbol{x},t)}$. If we consider the first jet bundle as a fiber bundle on top of Y, the coordinates of this super-bundle are (x^i, t, y^i, v^i_j) with $v^i_j = \partial_j \phi^i(\boldsymbol{x}, t)$.

For a given flow, i.e., a section ϕ, the Lagrangian point of view for the fluid is determined by the restriction $\phi(\boldsymbol{x}_0, t), \boldsymbol{x} = \boldsymbol{x}_0 = \text{const.}$, namely, the positions $y^i(\boldsymbol{x}_0)$ occupied in the particle position space (in the bundle Y) by the particle labeled $\boldsymbol{x}_0$ (or, in other words, by the particle which was initially at the given fixed point $\boldsymbol{x}_0$ in the base space).

The Euler point of view is described by the structure group of the bundle Y. Because we want to observe the flow (the section) at the same position for any moment in time, we need to transport this fixed space position through all the frames at all moments in time. This means smoothly moving some initial point from the intersection of the initial fiber with the section, viz., $(x_0, t_0, y_0) = \pi^{-1}(x_0, t_0) \in F \cap \phi(X)$, to another fiber, by the action of the structure group of the bundle. This way, the position in the spatial position space is maintained constant. When we apply all the group elements $g \in G$ in the structure groups to y_0, the resulting points $g(y_0)$ touch all the other fibers, but not all points belonging to the section image $\phi(X)$ (see Fig. 9.3). We choose only those points which belong to the section image, taking

them in order of increasing times. The uniqueness of each such point is guaranteed by the smooth fiber bundle structure. The resulting curve

$$\Big\{(x,t,g(y_0)) \in \phi(X) | g \in G\Big\}$$

lies entirely on the $\phi(X)$ section image, so represents the flow. Moreover, since all the points $g(y_0)$ belong to the same relative position in each fiber, they represent observations made at the same position in the reference container. This is the Euler point of view of the flow. The projection of the curve $\big\{g(y_0) \in \phi(X) | g \in G\big\}$ represents the particles that intersect this fixed position at various times.

Now we can define the two types of fluid velocities. The Lagrangian (material) velocity is

$$\boldsymbol{v}_L = \boldsymbol{v}(\boldsymbol{x},t) = \frac{\phi(x,t)}{\mathrm{d}t} \,,$$

while the (spatial) Euler velocity is

$$\boldsymbol{v}_E = \boldsymbol{v}(\boldsymbol{y},t) = \frac{y\Big(x\big(\pi^{-1}(y)\big),t\Big)}{\mathrm{d}t} \,.$$

In order to write the dynamical equations for the fluid, we introduce the incompressibility criterion in the form of invariance of the Lie derivative of the volume form. In the following, we denote the Euler (spatial) velocity components by $\boldsymbol{v}$:

$$\mathcal{L}_{\boldsymbol{v}} \Omega = 0 \,, \tag{9.3}$$

where $\mathcal{L}_{\boldsymbol{v}}$ is the Lie derivative in the direction of the fluid velocity field and Ω is a differential 3-form, the volume form, possibly induced by the Riemannian metric on M. In principle, on 3D domains, $\Omega = \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z = \text{const.}$, but when looking for more complicated solutions for liquid drop or soap film bubbles, it may be convenient to use a non-constant volume form, and this makes (9.3) non-trivial.

The dynamical equation for an incompressible force-free inviscid fluid has the form

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} + \nabla_{\boldsymbol{v}} \boldsymbol{v} = -\mathrm{d}P \,, \tag{9.4}$$

where $\nabla_{\boldsymbol{v}}$ is the covariant derivative along the flow field direction, and the 0-form P is the fluid pressure, defined by

$$P = \lambda + \frac{1}{2}(\Pi; \boldsymbol{v}, \boldsymbol{v}) \,,$$

with λ the Lagrange multiplier needed in the Lagrangian form in order to consider the volume conservation constraint, and Π the second fundamental form, defined on the tangent bundle of the flow, that is, on $T\phi(X)$.

We mention that similar expressions can be obtained if one includes external force fields (potential fields are of course easier to handle) and viscosity, since the viscous drag operator is introduced through a linear elliptic operator (Laplace operator on flat manifolds, and Laplace–Beltrami operator on Riemannian manifolds), which can be absorbed in the differential geometry formalism. There are of course more details defining these equations in the most rigorous form, but such an elaborate geometrical approach would exceed by far the topics studied in this book. More references can be found in [167–169, 279, 282, 283, 286].

We mention that if the flow is bounded by a fixed boundary ∂M, then the field $\boldsymbol{v}$ must be tangent to the boundary. The section describing the flow in the flow bundle Y is defined by the Euler velocity field and one can write

$$\frac{\partial \phi(\boldsymbol{x},t)}{\partial t} = \boldsymbol{v}\big(t,\phi(\boldsymbol{x},t)\big) \ ,$$

with initial condition $\phi(\boldsymbol{x},0) = \boldsymbol{x}$, that is $F \sim M$. The chain rule applied to the above equation allow us to write the Euler equation in the form

$$\frac{\partial^2 \phi(\boldsymbol{x},t)}{\partial t^2} = -\mathrm{d}P\big(\phi(\boldsymbol{x},t),t\big) \ .$$

From the last form of the Euler equation, we notice that the acceleration of the flow (curvature of the section) is an exact form and that it is orthogonal (in some scalar product induced by the Riemannian metric) to the tangent space of the volume-preserving diffeomorphisms, namely, the divergence-free fields. This means that the fluid motion is a geodesic in the set of such diffeomorphisms. The same equation describes the motion of an ideal incompressible fluid filling an arbitrary Riemannian manifold M equipped with a volume form [286]. In the latter case, $\boldsymbol{v}$ is a divergence-free vector field on M, while $\nabla_{\boldsymbol{v}}\boldsymbol{v}$ stands for the Riemannian covariant derivative of $\boldsymbol{v}$ in its own direction.

We mention that the differential geometry/topological formalism as applied to incompressible hydrodynamics has promoted research directions that might not otherwise have been easily discovered. This description of the Euler equation as a geodesic flow on a Lie group and its resulting Hamiltonian formulation reveals new knowledge and investigation in geometry and topology, especially in the case of infinite-dimensional spaces. One famous problem is *long-time existence and uniqueness* for the Cauchy problem of 3D Euler hydrodynamics, one of the six unsolved Clay Millennium Problems. Such situations in which the new mathematical tools needed to solve physical problems have returned to pure mathematics with enriched results have happened before. To give an example, during the early 1930s, the development of Dirac's approach to quantum mechanics brought revolutionary

new concepts to the theory of distributions, not to mention linear operators and their representations.

Regarding the differential geometric/topological description of incompressible fluids, the invariants of the Lie group of diffeomorphisms (which are more easily calculated as Casimir elements) provide a source of first integrals for the Euler equation, which in their turn help us to find new criteria for stability of flows, onset of turbulence, and long lifetime analytic solutions. In the case of 3D manifolds M, the invariants are, of course, the energy integral

$$E(\boldsymbol{v}) = \int_M (\nabla \boldsymbol{v}, \nabla \boldsymbol{v}) \mathrm{d}\omega \ ,$$

the helicity (Hopf) invariant

$$J(\boldsymbol{v}) = \int_M (\nabla \times \boldsymbol{v}, \boldsymbol{v}) \mathrm{d}\omega \ ,$$

which describes the mutual linking between the trajectories of the vorticity field $\nabla \times \boldsymbol{v}$. However, for 2D flows, there is an infinite number of such invariants. They are called *enstrophy invariants* and are expressed in the form

$$J_k(\boldsymbol{v}) = \int_{M^2} (\nabla \times \boldsymbol{v})^k \mathrm{d}x \wedge \mathrm{d}y \ .$$

Serre, Tartar, Ovsienko, Khesin, and Chedanov explained in a series of independent articles (see, for example, the review in [286]) that having an infinite number of invariants is a property of this type of volume-form-conserving flow in manifolds with an even number of dimensions. A very simple explanation is that, in odd dimensions, the vorticity field is frozen into the fluid and transported with the flow, while in even dimensions, the fluid transport of the infinitesimal invariants is pointwise. Indeed, these are very interesting mathematical speculations, but there will never be an experimental observation of an n-dimensional fluid in this observable universe. There is not enough evidence for realistic systems of this kind, and even if such a higher-dimensional fluid manifold were to model some physical, biological, or social system, the evidence that the laws of such systems will still be Newton's laws, and in particular that they will obey the law of inertia, is very slim.

Finally, we note that the differential geometry of volume-preserving diffeomorphisms of a bounded domain of a 2D manifold differs drastically from that of higher dimensions. This difference occurs because, in more than two dimensions, there is enough space for particles to move to their final positions without hitting each other. But the motion of the particles in the plane might necessitate the braiding of particles into much longer paths, in spite of the boundedness of the domain. The

diameter of the group of all possible diffeomorphisms of a bounded manifold with $n > 2$ dimensions is given by the Shnirelman theorem:

$$\operatorname{dia}\big(\operatorname{SDiff}(M)\big) \leq \frac{2}{\sqrt{3}} n^{1/2} , \quad n > 2 .$$

In other words, the set of all such paths has no infimum (there is no shortest path). Actually, the diameter of the group of diffeomorphisms on a manifold of dimension 2 is infinite (there is no longest path).

9.2 Geometry of Viscous Fluids

In this section, we present a geometrical approach to the Navier–Stokes equations in the absence of external forces ($\boldsymbol{f} = 0$). This differential geometry formalism is based on solution representation formulas. The advantage of this formalism over other mathematical approaches is that it generates a direct and compact formula to work with, and also that it brings a geometrical intuition to the solution. In contrast with other mathematical models for fluids that cannot describe domains with boundary, the topic we are concerned with in this book, the geometric representation model presented here can handle flows with either fixed or free boundaries. The trick is to modify the Navier–Stokes equations in order to keep the boundary as an invariant set for the velocity field. Such methods are described in detail in [282, 283], where the authors use contact topological methods on domains with boundary. For free boundaries, the procedure is to allow the shape of the boundary to vary rather than to vary the metric. Formally, this approach is still an open problem.

The equation of continuity and Navier–Stokes equation without external forces have the local form

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0 , \quad \frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{v} = -\frac{1}{\rho} \nabla P + \nu \triangle \boldsymbol{v} ,$$

for a fluid velocity field $\boldsymbol{v}(\boldsymbol{r}, t)$, pressure $P(\boldsymbol{r}, t)$, density $\rho(\boldsymbol{r}, t)$, and kinematic viscosity $\nu > 0$. The traditional way to express the Navier–Stokes equations in a differential form approach leads to the Navier–Stokes equation in its global form

$$\mathcal{L}_{\boldsymbol{v}} \rho = 0 , \quad \frac{\partial \boldsymbol{v}}{\partial t} + \nabla_{\boldsymbol{v}} \boldsymbol{v} = -\mathrm{d}H ,$$

where $\mathcal{L}_{\boldsymbol{v}}$ is the Lie derivative in the direction of the flow, and $\nabla_{\boldsymbol{v}}$ is the covariant derivative along $\boldsymbol{v}$, which also depends on the metric g. Here H is the energy

0-form obtained from Cartan's formula (4.15) (see Sect. 4.4, and also Friedlander and Serre's basic book [287]):

$$H = \frac{1}{\rho}P + \frac{\nu}{2} * \boldsymbol{v} * \boldsymbol{v}g \ ,$$

where $*$ represents the dual operation transforming vector fields into dual 1-forms, and g denotes the metric of the space.

In order to use the representation formalism, we limit ourselves to the study of the linearized Navier–Stokes equation and incompressible fluids. The generalization of this approach to compressible flows (based on the geometric form of the equation of continuity $\mathcal{L}_{\boldsymbol{v}}$) is just a matter of more extensive calculations. However, it is more challenging to approach the full nonlinear Navier–Stokes equations using this representation method [287], and we will not present this aspect here.

We consider the linearized incompressible Navier–Stokes equation for a force-free fluid ($\rho = \text{const.}, \nabla \cdot \boldsymbol{v} = 0, \boldsymbol{f} = 0$):

$$\frac{\partial \boldsymbol{v}}{\partial t} = -\frac{1}{\rho}\nabla P + \nu \triangle \boldsymbol{v} \ . \tag{9.5}$$

A solution of (9.5) can be given in the form (see [288, 289] or [121, Chap. 10]):

$$\boldsymbol{v} = C\big[\nabla \times \nabla \times (\boldsymbol{r}\beta) + \nabla \times (\boldsymbol{r}\alpha)\big] + \nabla \Phi \ , \tag{9.6}$$

$$P = -\rho \frac{\partial \Phi}{\partial t} \ , \tag{9.7}$$

where C is a constant and $\alpha(\boldsymbol{r}, t)$, $\beta(\boldsymbol{r}, t)$, and $\Phi(\boldsymbol{r}, t)$ are smooth functions satisfying

$$\nu \triangle \alpha = \frac{\partial \alpha}{\partial t} \ , \qquad \nu \triangle \beta = \frac{\partial \beta}{\partial t} \ , \qquad \triangle \Phi = 0 \ .$$

The proof is by straightforward calculation. We will study the solution in more detail and try to explain it in a more geometric way.

Because of the incompressibility condition, the last term can be written in the form

$$\frac{\partial \boldsymbol{v}}{\partial t} + \frac{1}{\rho}\nabla P = \nu \triangle \boldsymbol{v} = -\nu \nabla \times \nabla \times \boldsymbol{v} \ .$$

Let us study the first term in the solution (9.6). It can be rewritten in the form

$$\nabla \times \nabla \times (\boldsymbol{r}\beta) = \big(2 + \boldsymbol{r} \cdot \nabla - \boldsymbol{r}\nabla \cdot\big)\nabla \beta = \mathcal{O}\beta \ , \tag{9.8}$$

where we have denoted this action on the function β with a generic solution operator $\mathcal{O}$. This term, when plugged into the Navier–Stokes equation, generates

the Laplacian of the velocity on the right-hand side of the equation. The total lack of symmetry of the solution expressed as in the second term of (9.8) is somehow surprising. This property of being a solution of $\mathcal{O}$ can be written

$$[\mathcal{O}, \triangle] = 0 \,, \tag{9.9}$$

that is, the solution operator $\mathcal{O}$ commutes with the Laplacian operator. In a way, this is the meaning of a solution as an eigenvector, since the Laplace operator works like the Hamilton function, and the solution operator like a conservation law. More symmetries of the operator $\mathcal{O}$ can be revealed if we rewrite this solution operator in terms of differentiable forms.

Let us consider β as a 0-form on $D \subset \mathbb{R}^3$, with D connected, contractible (i.e., we can find a smooth family of smooth maps defined on D which can shrink D smoothly to a point), and compact. We assume the existence of the canonical scalar product of vectors $\langle \, , \rangle$ in $\mathbb{R}^3$, which is also the Riemannian metric. Consider

$$\boldsymbol{r}_{\text{cov}} = x_i \mathrm{d}x^i \,,$$

which are the components of a closed 1-form on D. Indeed,

$$\mathrm{d}\boldsymbol{r}_{\text{cov}} = \mathrm{d}(x_i \mathrm{d}x^i) = \frac{\partial x_i}{\partial x^k} \mathrm{d}x^k \wedge \mathrm{d}x^i = \delta_{ik} \mathrm{d}x^k \wedge \mathrm{d}x^i = 0 \,.$$

If $\delta = - * \mathrm{d} *$ is the *codifferential* operator acting on k-forms defined on D, $0 \leq k \leq 3$, $*$ the Hodge operator [a linear operator on the exterior algebra of forms mapping k-forms into $(3-k)$-forms, as described below], and $\mathcal{L}_{\boldsymbol{r}}$ the Lie derivative in the direction of the smooth contravariant vector field

$$\boldsymbol{r} = x^i \frac{\partial}{\partial \mathrm{d}x^i}$$

on D, then we have

$$(2 + \boldsymbol{r} \cdot \nabla - \boldsymbol{r} \nabla \cdot) \nabla \beta = (1 + \mathcal{L}_{\boldsymbol{r}} + \boldsymbol{r}_{\text{cov}} \delta) \mathrm{d}\beta = (1 + \mathcal{L}_{\boldsymbol{r}}) \mathrm{d}\beta + \boldsymbol{r}_{\text{cov}} \triangle \beta \,, \tag{9.10}$$

where we have used the definition $\triangle = * \mathrm{d} * \mathrm{d}$ for the Laplace operator. This is a highly symmetric form and shows explicitly the Lie derivative transport and the Laplace representation of a solution.

To prove this geometric expression, we use the definition of the Hodge operator and the action of the Lie derivative on a k-form ω given by the *Cartan formula*

$$\mathcal{L}_{\boldsymbol{r}} \omega = \boldsymbol{r} \perp \mathrm{d}\omega + \mathrm{d}(\boldsymbol{r}; \omega) \,,$$

where (;) is the action of a vector field on a form, by definition. In this case, setting $\omega = \beta$ in the above relation, we can finally write

$$(\boldsymbol{r}\cdot\nabla)\nabla\beta = \mathrm{d}(\boldsymbol{r}\perp\mathrm{d}\beta) = \mathrm{d}\big[\mathcal{L}_{\boldsymbol{r}}\beta - \mathrm{d}(\boldsymbol{r};\beta)\big] = \mathrm{d}(\mathcal{L}_{\boldsymbol{r}}\beta) = \mathcal{L}_{\boldsymbol{r}}\mathrm{d}\beta\ ,$$

which proves the middle term in brackets in (9.10).

In a similar way, the second term in the solution (9.6) can be rewritten in the form

$$\nabla\times(\boldsymbol{r}\alpha) = -\boldsymbol{r}\times\mathrm{d}\alpha = -\boldsymbol{r}_{\mathrm{cov}}\wedge\mathrm{d}\alpha\ .$$

Finally, we can express the Navier–Stokes linearized force-free equation for incompressible flow and one of its solutions in completely geometrical form:

$$\mathcal{L}_{\boldsymbol{v}} = -\frac{1}{\rho}\mathrm{d}P - \nu * \mathrm{d} * \mathrm{d}(\boldsymbol{v})\ , \tag{9.11}$$

which has the solution

$$\boldsymbol{v}(\alpha,\beta,\Phi,\boldsymbol{r}) = C\big[-\boldsymbol{r}_{\mathrm{cov}}\wedge\mathrm{d}\alpha + (1+\mathcal{L}_{\boldsymbol{r}}+\boldsymbol{r}_{\mathrm{cov}}\delta)\mathrm{d}\beta\big] + \mathrm{d}\Phi\ . \tag{9.12}$$

9.3 Soap Films with Boundary

The problem of finding the area-minimizing surface with a given boundary was posed by Lagrange in 1760, and investigated by the Belgian scholar J.A.F. Plateau (1801–1883) [290] in the second half of the nineteenth century. For his experiment, Plateau used soapy water mixed with glycerine and dipped wire contours into it, noting that the surfaces formed were minimal surfaces. Mathematically, the problem was tackled by Weierstrass, Riemann, and Schwarz, and finally solved in an acceptable way by Douglas and Radó, and later enriched and fully solved by Jenny Harrison [291]. The experiments initiated by Plateau showed that an area minimizing surface can be obtained in the form of a film of incompressible liquid stretched on a rigid frame (the so-called soap film system).

Apparently, there is a straightforward way to evaluate the stationary equilibrium condition for a soap film from thermodynamic considerations. In the stationary case $\boldsymbol{v} = 0$, for a fluid with free boundary Σ, the Euler equation reads

$$-\frac{1}{\rho}\nabla P + \boldsymbol{f} = 0\ , \tag{9.13}$$

where ρ is the fluid density, P is the pressure, and $\boldsymbol{f}$ is the mass density of the force field acting inside the fluid. If the force field derives from a potential, i.e., $\boldsymbol{f} = -\nabla\Phi$, the stationary Euler equation reduces to the simplest Bernoulli type of equation, viz., $P = P_0 - \rho\Phi$.

Let us treat the separating surface as a parameterized regular geometrical surface $\boldsymbol{r}(u, v) : U \subset \mathbb{R}^2 \to \Sigma$ with outward unit normal $\boldsymbol{n}(u, v)$. We consider a normal variation of this surface (see the appendix in Appendix 1):

$$\boldsymbol{r}^t(u, v, t) = \boldsymbol{r}(u, v) - th(u, v)\boldsymbol{n}(u, v) \ ,$$

where $h(u, v)$ is a real differentiable function. For each value of t, the map $\boldsymbol{r}^t : U \times (-\varepsilon, \varepsilon) \to \mathbb{R}^3$ is a regular parameterized surface, and for $t = 0$, the normal variation reduces to the original surface. If the surface suffers a normal variation determined by the function $h(u, v)$, there will be some work produced by the compression in any elementary volume in the vicinity of the surface:

$$W_{\text{vol}} = -t \int_{\bar{U}} Ph\sqrt{EG - F^2}\, \mathrm{d}u\, \mathrm{d}v \ ,$$

where P is the change in pressure across the surface. The total change in the free energy of the system is given by δW_{vol} plus the work associated with the variation of the area of the separating surface, that is, the surface energy given by the product between the coefficient of surface pressure γ and the variation of the area $\delta\mathcal{A}$. The total variation in the free energy becomes

$$\delta\mathcal{F} = -t \int_{\bar{U}} Ph\sqrt{EG - F^2}\, \mathrm{d}u\, \mathrm{d}v + \gamma\, \delta\mathcal{A} \ .$$

From the equilibrium condition $\delta\mathcal{F} = 0$, we obtain the expression for the surface tension across the surface

$$P_\Sigma = -\gamma(\kappa_1 + \kappa_2) = -2\gamma H \ ,$$

where $\kappa_{1,2}$ are the two principal curvatures of the surface at p and H is the mean curvature calculated using the convention wherein the sphere S^2 has $H_{S^2} = -1$. The above equation is the Young–Laplace formula for the capillary pressure.

Now, we know that soap films spanned by embedded curves are represented by surfaces with boundary, so the surrounding pressure is the same at any point, and in consequence, such surfaces must be minimal surfaces. If the soap film is represented by a surface without boundary (like drops, bubbles, shells, double bubbles, bubble clusters, or antibubbles), the pressure inside is different from the pressure outside the surface. However, in a stationary equilibrium situation, the pressures inside and outside are uniform and constant, so the surface must be described by a constant mean curvature (CMC) surface. Minimal surfaces and Delaunay surfaces are good examples.

The equilibrium condition for static soap films can be expressed in a simpler form if the coordinate system used to parameterize the surface Σ is orthogonal, i.e., if $F = \boldsymbol{r}_u \cdot \boldsymbol{r}_v = 0$. It is always possible to choose such an orthogonal parametrization for a regular surface. Moreover, if in addition we have $\boldsymbol{r}_u \cdot \boldsymbol{r}_u = \boldsymbol{r}_v \cdot \boldsymbol{r}_v$ and $\boldsymbol{r}_u \cdot \boldsymbol{r}_v = 0$,

the surface is said to be isothermal. Isothermal parameterized surfaces are endowed with orthogonal, but not normalized curvilinear coordinates.

If the surface Σ parameterized by $\boldsymbol{r}(u, v)$ is isothermal, the mean curvature is expressed by a simple equation of the form

$$\boldsymbol{H} = H\boldsymbol{n} = -\frac{1}{2\boldsymbol{r}_u \cdot \boldsymbol{r}_u} \triangle_\Sigma \boldsymbol{r} \ ,$$

where $\triangle_\Sigma = \partial_{uu} + \partial_{vv}$ is the Laplace–Beltrami operator in the curvilinear coordinates on the surface.

It follows that the study of soap films with boundary is the study of minimal surfaces $H(\Sigma) = 0$, and for a closed soap film without boundary, it reduces to the study of CMC surfaces. Minimal surfaces have a lot of interesting topological properties. The zeros of the Gaussian curvature of a minimal surface are isolated. In other words, there is no straight line on a minimal surface. Furthermore, there are no compact minimal surfaces, because all the points of a regular minimal surface are hyperbolic. It follows that all (regular) minimal surfaces are unbounded. In addition, if Σ is a regular closed minimal surface and not a plane, the image of the Gauss map is dense in the sphere S^2. If Σ is minimal and has no planar points ($K_{\text{Gauss}}(\Sigma) \neq 0$), then the angle of intersection of any two curves on Σ and the angle of intersection of their spherical images through the tangent map to the Gauss map are equal up to a sign. This means that the directional derivatives of the pressure along two perpendicular directions of the tangent plane are also perpendicular.

Under deeper analysis, Plateau's original problem of finding the minimal surface involves some geometrical issues which occur when one tries to solve this problem with a unique solution. The main mathematical question becomes the existence and uniqueness of the solution, and the question is whether there exists a surface with minimal area Σ that spans a prescribed smoothly embedded closed curve $\gamma \in \mathbb{R}^3$. The most recent systematic study of this problem was reported by Harrison in [291]. She explains that the existence and uniqueness of solutions depend on the definitions of surface, area, and span. From previous reports, the following set of questions were raised:

1. Does there always exist a surface Σ spanning a given curve γ?
2. Is the infimum of these areas (of surfaces spanning the curve) nonzero? Figure 9.4 gives an example of a non-spanning surface of the unit circle.
3. Does there exist a surface Σ_0 spanning γ with prescribed area m?
4. What is the structure of Σ_0 away from its boundary γ and near the boundary?

The results in the study [291] provide exact answers to the first three questions above by proving the following theorem:

Theorem 10 *Given a smoothly embedded closed curve $\gamma : S^1 \to \mathbb{R}^3$, there exists a surface Σ_0 spanning γ with minimal area $|\Sigma_0|$. This surface is an element of a large set of surfaces which includes representatives of all types of observed soap*

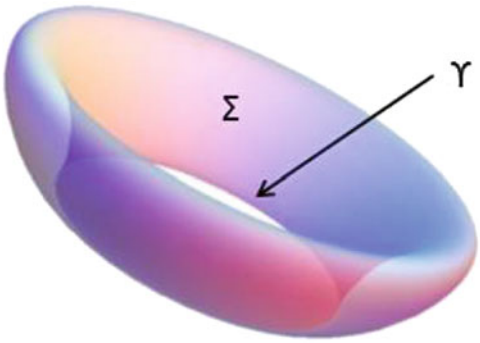


Fig. 9.4 A non-spanning surface of the unit circle. Figure inspired by Harrison [291]

films as well as all smoothly immersed surfaces of all genus types, orientable or nonorientable, including those with possibly multiple junctions.

This result not only closes a 250 year old problem, but it provides an elegant geometrical solution which eliminates the ambiguity and traps of previous attempts. Despite the tedious constructions [244, 291], we shall try to take the reader through the key concepts, but only for the 3D case, since the general case is too complicated to be examined in this book.

Let U be an open subset of $\mathbb{R}^n$ and Ω_R an open $(n-1)$-sphere of radius R, containing the origin. We define a *k-element* in U, denoted by $(p;\alpha)$, to be a pair comprising a point $p \in U$ and a finitely supported section of the k th exterior algebra of the tangent bundle of U at p, $\alpha \in \Lambda_k(T_p(U))$, $0 \leq k \leq n$. In other words a k-element can be a function, a vector, or a contravariant skew-symmetric tensor of order k with a given orientation, which is different from zero only at a finite number of points. For example, if $n = 3$, a 1-element can be $(0;\boldsymbol{X})$ with $0 = (0,0,0)$ and $\boldsymbol{X} = (x,y,z)$ as a vector with origin at $p = (0,0,0)$ and end at $\boldsymbol{X}$. A 2-element can be $\big((0,0,1),\boldsymbol{X} \wedge \boldsymbol{Y}\big)$ with $\boldsymbol{X} = (x,y,z)$ and $\boldsymbol{Y} = (x',y',z')$, where $\boldsymbol{X} \wedge \boldsymbol{Y}$ is the oriented parallelogram $\boldsymbol{X} \times \boldsymbol{Y}$, and so on.

A *Dirac k-chain* in U is a formal sum $A = \sum(p_i;\alpha_i)$ of k-elements, each term of the sum being defined at a different point p_i of U (see Fig. 9.5 upper left frame for an example with $n = 3$, $k = 2$). All the differential objects defined in this section and all the continuous operators have these topological and calculus properties based on special technical choices of the norms which can only be introduced and described through long procedures that go beyond our present scope. We refer the reader to the classic papers on this topic (see [244, 291] and the references therein for more detail).

The closure of the vector space of all the Dirac k-chains of U can be structured as a Banach space $\hat{\mathcal{B}}_k^r(U)$ with a Fréchet-type norm $\|\cdot\|_r$ defined by uniform boundedness on each directional derivative, up to a certain order r. The elements of this Banach space are called *differentiable k-chains of class B^r in U*. The dual

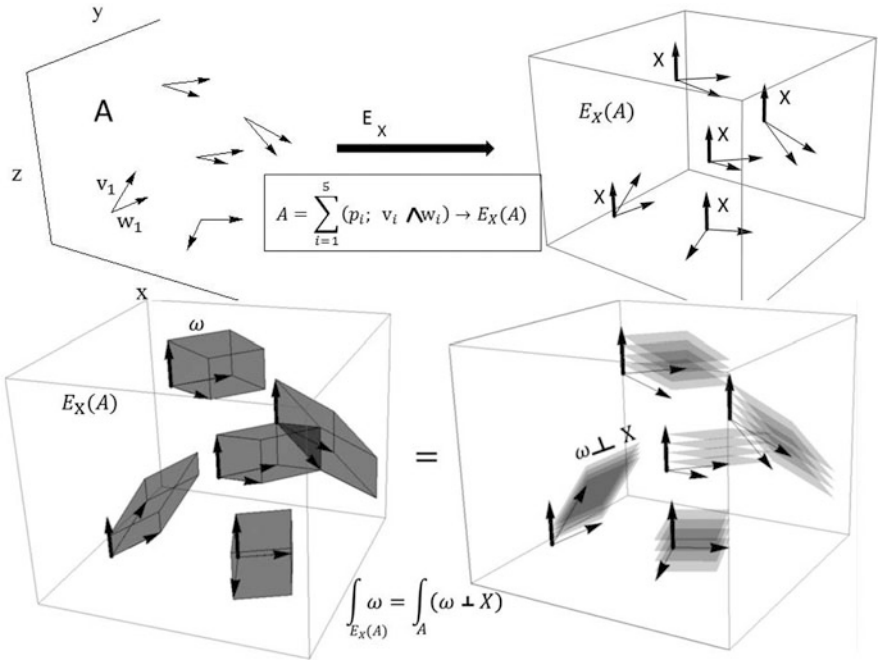


Fig. 9.5 *Clockwise from upper left*: Example of a differential 2-chain A in $\mathbb{R}^3$ with 5 elements; application of the extrusion operator $E_X(A)$ on it; pictorial representation of (9.15) applied to the differential k-chain A; *left-hand side* of the equation, and *right-hand* terms

is the space $\mathcal{B}_k^r(U)$ of differentiable k-forms on U. The two spaces are dual and canonically isomorphic by the integral pairing

$$\int_A \omega = \omega(A) \ , \tag{9.14}$$

where $\omega \in \mathcal{B}_k^r(U)$ and $A \in \hat{\mathcal{B}}_k^r(U)$, and the right-hand side is defined in the sense of the action of differentiable forms on vector fields defined in Sect. 4.4 [see (4.9)].

In the following paragraphs, we skip the upper index r which indicates the degree of smoothness or boundedness, and in addition we shall not mention the open U in brackets, but consider them as understood.

In order to study integration theory using differentiable k-chains, we need the equivalent of a simplicial complex from algebraic topology. So we define an *affine n-cell* in $\mathbb{R}^n$, denoted by σ, as the intersection of finitely many affine half-spaces in $\mathbb{R}^n$, whose closure is compact. This is basically a prism. A cell is not necessarily closed or open, and in general an *affine k-cell* in $\mathbb{R}^n$, $0 \le k \le n$, is an affine k-cell in the subspace $\mathbb{R}^k$ of $\mathbb{R}^n$. An oriented cell has a given orientation for its edges and faces, like any simplex. Below, we introduce one of the major elements in the proof of the existence and uniqueness of a non-trivial solution for the Plateau problem:

Theorem 11 *If σ is an affine oriented k-cell in $\mathbb{R}^n$, then there is a unique differentiable k-chain $\tilde{\sigma} \in \hat{\mathcal{B}}_k$ such that*

$$\int_{\tilde{\sigma}} \omega = \int_{\sigma} \omega \ ,$$

for any differentiable k-form $\omega \in \mathcal{B}_k$, where the right-hand side of the above equation is taken in the Riemann integral sense.

Proof of Theorem 11 can be found in [244, 291]. This theorem says that, instead of the Riemann integration on complexes made of affine hyperplanes, i.e., cells, we can use the formalism of duality and integral pairing for k-chains, e.g., from (9.14). If $n = 3$ and we have a 2-cell in the shape of a rectangle, then $\tilde{\sigma} = [0, 2] \times [0, 3]$ in the (x, y) plane. If we apply the area 2-form $\omega = \mathrm{d}x \wedge \mathrm{d}y$ (to the right-hand side of the above equation) to the vectors $\boldsymbol{v} = (2, 0, 0)$ and $\boldsymbol{w} = (0, 3, 0)$ generating this cell, we have

$$\omega(\boldsymbol{v} \wedge \boldsymbol{w}) = (\mathrm{d}x \wedge \mathrm{d}y; \boldsymbol{v}, \boldsymbol{w}) = (\mathrm{d}x \wedge \mathrm{d}y) v^1 w^2 = 6\, \mathrm{d}x \wedge \mathrm{d}y \ ,$$

which is exactly the left-hand side of the formula above, namely, an area of 6 units. We define the support of a differentiable k-chain J to be the set

$$\operatorname{supp} A = |A| = \inf \left\{ E \subset U \mid \int_A \omega = 0, \forall \omega \in \mathcal{B}_k, \ \operatorname{supp}(\omega) \cap E = \emptyset \right\} \ .$$

Obviously, the support of any differentiable k-chain exists, and it is unique.

Below, we combine a differentiable k-chain with a vector field and define four *primitive geometric operators* acting on differentiable chains: extrusion, retraction, boundary operator, and prederivative. If X represents a differentiable vector field on U, we define a continuous bilinear map called the *extrusion operator* as acting on differentiable k-chains A by (see Fig. 9.5 upper right)

$$E_X(A) = E_X\left(\sum_i (p_i; \alpha_i) \right) = \sum_i \left(p_i; X(p_i) \wedge \alpha_i \right) \ ,$$

which maps from $\hat{\mathcal{B}}_k$ to $\hat{\mathcal{B}}_{k+1}$ and has the integral property

$$\int_{E_X(A)} \omega = \int_A (\omega \perp X) \ , \tag{9.15}$$

for any $(k + 1)$-differentiable form in $\mathcal{B}_k$ (see bottom row of Fig. 9.5). Again, $\perp$ represents the interior product between a vector field and a differentiable form. This is called the *change of dimension I* integral formula between the matching pairs (A, ω) (see Fig. 9.5).

The operator called *retraction* is also defined with the help of a vector X. The operator acts on any differentiable k-chain $A = \sum(p_i; \alpha_i) \in \hat{\mathcal{B}}_k$, but it is easier to show its action on a differentiable k-element $(p; \alpha)$ and then extend the action by linearity. We denote the skew-symmetric exterior product of vectors α by $\alpha = v_1 \wedge \ldots \wedge v_k$. We then have

$$E_X^\dagger(p; \alpha) = \sum_{j=1}^{k} (-1)^{j+1} \langle X, v_j \rangle (p; \hat{\alpha}_j) \ , \tag{9.16}$$

where we denote $\hat{\alpha}_j = v_1 \wedge v_{j-1} \wedge v_{j+1} \wedge \ldots \wedge v_k$. The retraction operator lowers the degree of the k-chain, i.e., $E_X^\dagger : \hat{\mathcal{B}}_k \to \hat{\mathcal{B}}_{k-1}$ (see Fig. 9.6).

The *prederivative operator*, denoted by $P_X : \hat{\mathcal{B}}_k \to \hat{\mathcal{B}}_k$, is determined by an arbitrary vector field X and its action upon a k-element $(p; \alpha)$ according to

$$P_X(p; \alpha) = \lim_{t \to 0} \left[\left(p + tX; \frac{\alpha}{t} \right) - \left(p; \frac{\alpha}{t} \right) \right] . \tag{9.17}$$

Note that the prederivative does not change the order of a differentiable k-chain. Here we calculate a simple example to see how the prederivative operator works (see also Fig. 9.7).

With the help of the prederivative, we further introduce the *boundary operator* acting on differentiable k-chains, denoted ∂, which can be expressed in the form of a sum of *directional boundary operators* along all directions of an orthonormal basis in $\mathbb{R}^n$:

$$\partial = \sum_{i=1}^{n} P_{e_i} E_{e_i}^\dagger : \hat{\mathcal{B}}_k \to \hat{\mathcal{B}}_{k-1} \ , \tag{9.18}$$

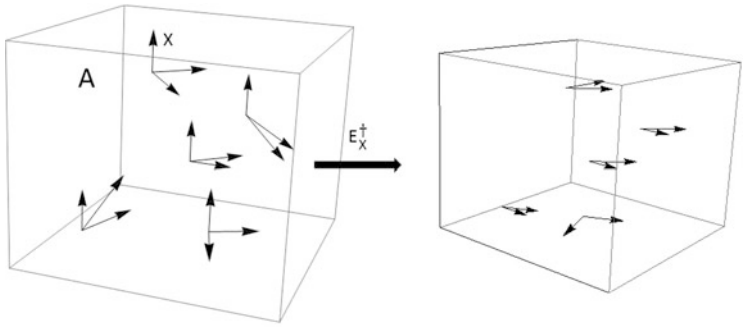


Fig. 9.6 Action of the retraction operator on a 3-chain with 5 elements, with $X = (0, 0, 1)$

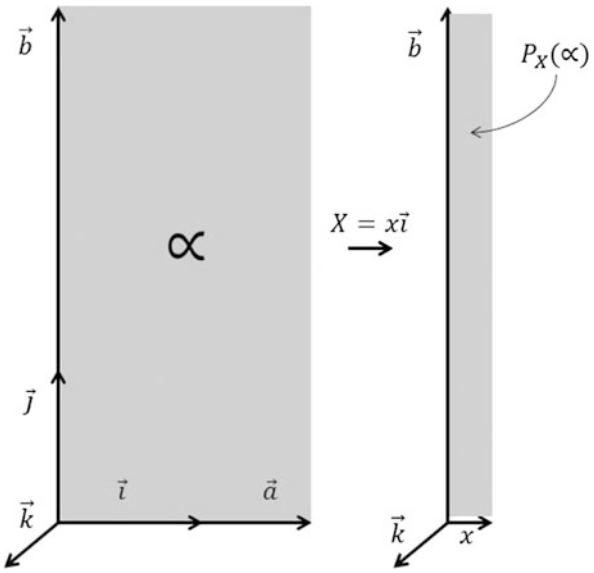


Fig. 9.7 Action of the prederivative upon a 2-chain in $\mathbb{R}^3$, as obtained through (9.20)

where $\{e_i\}_{i=1,\dots,n}$ is an orthonormal basis in $\mathbb{R}^n$, while P_{e_i} are the prederivative operators of (9.17) along these basis directions, and $E^{\dagger}_{e_i}$ is the retraction operator of (9.16).

The boundary and prederivative operators allow one to write down various useful constructions, including the *generalized Stokes theorem*:

Theorem 12 *For any differentiable k-chain $A \in \hat{\mathcal{B}}_k$ and any differentiable k-form $\omega \in \mathcal{B}_{k-1}$, we can define a continuous boundary operator ∂ by the equations*

$$\int_{\partial A} \omega = \int_A \mathrm{d}\omega \ , \quad \textit{and} \ \ \partial \cdot \partial = \partial^2 = 0 \ . \tag{9.19}$$

If a differentiable k-chain A has the property $\partial A = 0$, it is called a *differentiable k-cycle* in U.

Moreover, there is a reverse definition of the prederivative operator obtained from the boundary operator, known as *Cartan's magic formula for differentiable chains*:

$$P_X = E_X \partial + \partial E_X \ . \tag{9.20}$$

Equation (9.20) actually defines the prederivative as the dual of the Lie derivative (presented in Sect. 4.4), through the Cartan formula (4.15), between a vector field X and a differentiable form ω :

$$\boldsymbol{X}(\omega) = \boldsymbol{X} \perp \mathrm{d}\omega + \mathrm{d}(\boldsymbol{X}; \omega) \ ,$$

or written in another common notation, $\mathcal{L}_X(\omega) = i_X \mathrm{d}\omega + \mathrm{d}(i_v) \cdot \omega$.

Finally, using all the above definitions, we can write the *dual relation between the prederivative and the Lie derivative* in the form

$$\int_{P_X(A)} \omega = \int_A X(\omega) \,, \tag{9.21}$$

or in another commonly used notation,

$$\int_{P_X(A)} \omega = \int_A \mathcal{L}_X(\omega) \,,$$

where $A \in \hat{\mathcal{B}}_k$ and $\omega \in A \in \mathcal{B}_k$ once again. The prederivative is a way to 'geometrically' differentiate a differentiable k-chain A in some infinitesimal direction determined by a vector field X, even if the support of the differentiable chain is highly non-smooth, like the sharp edges of cells, or corners, etc. In Fig. 9.7, we represent such an action in the space $\mathbb{R}^3$.

With these prerequisites, we can now approach the Plateau problem. Consider a closed and smoothly embedded curve $\gamma : S^1 \to \Omega_R \subset \mathbb{R}^3$, with two points $q, p \in \Omega_R$, $p \in \gamma$, but q not on γ. We construct the differentiable vector field

$$Y(p,q) = \frac{q-p}{\|q-p\|} \,,$$

in a (tubular) neighborhood of any point p of the γ curve. The construction of the minimal area surface is based on two important constraints. The first refers to the definition of 'spanning'. Here, Harrison defines the *surface Σ spanning γ* as a surface whose boundary operator $\partial\Sigma$ generates the prederivative of a 1-chain representing γ along the vector field $Y(p,q)$. That is,

$$\partial\Sigma = P_{Y(p,q)}\tilde{\gamma} \,, \tag{9.22}$$

where $\tilde{\gamma}$ is a differentiable 1-chain representing γ in the integral duality sense of Theorem 11. The second constraint introduced in this theory is to insist that, for any smoothly embedded curve $\gamma' \in \Omega_R$ linking with γ with linking number 1, Σ should be such that $\gamma' \cap |\Sigma| \neq \emptyset$. In other words, a surface Σ qualifies for a Plateau solution if it is punctured at least at one point by any smoothly embedded closed curve which links the given curve γ once. The second constraint does not consider Plateau solutions behaving like the annulus torus in Fig. 9.4. We present some examples in Fig. 9.8.

Now we gather the theorems and definitions previously discussed in this section and come up with a complete construction. First, we have a closed and smoothly embedded curve $\gamma : S^1 \to \Omega_R$. Such curves cannot be a simple point (smooth image of S^1 which is not contractible to a point) and cannot be unbounded because they are contained in the sphere of radius R. Next, we identify an arbitrary point q in the sphere, but not on the curve, and construct the differentiable vector field $Y(p,q)$,

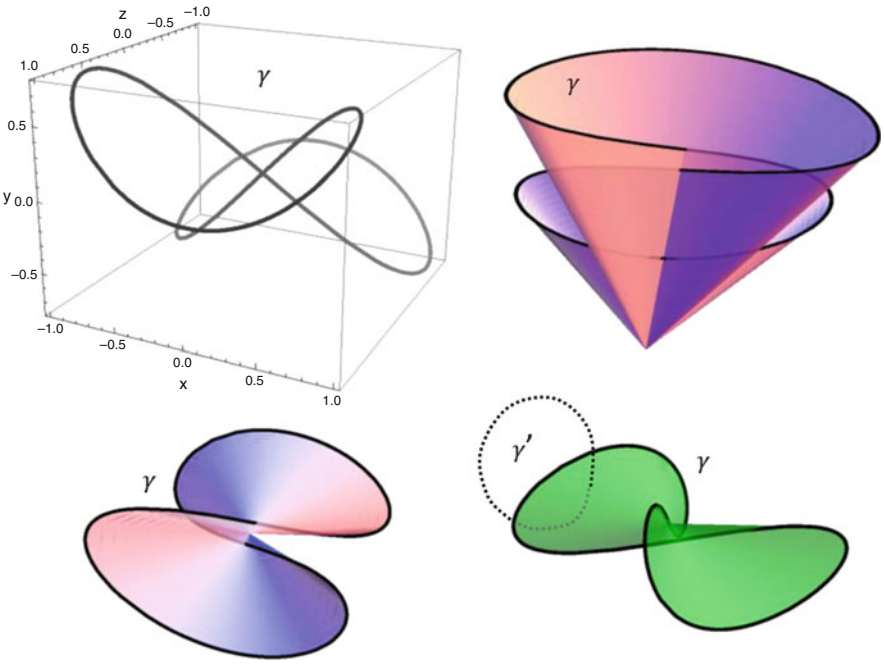


Fig. 9.8 For the smoothly embedded closed curve γ presented in the *upper left corner*, we can create 2-cells $\tilde{\sigma}$ based on various points q [see (9.22)]. However, there exists a unique surface (*bottom right corner*) Σ spanning the curve γ with minimal area and such that any other curve γ' linking γ only once will intersect it at least once

$p \in \gamma$. Then we define the set of all differentiable 2-cells $\tilde{\tau}$ (which are nothing but smoothly embedded surfaces with a boundary) having the dual cell of the γ curve as topological boundary, i.e.,

$$\partial_{\text{topological}} \tilde{\tau} = \tilde{\gamma} \ . \tag{9.23}$$

Let τ be the dual differential 2-chain associated with the 2-cell $\tilde{\tau}$ through Theorem 11, and consider the same duality between $\tilde{\gamma} \sim \gamma$.

The next step is to find a procedure for calculating the areas of these differentiable 2-chains $\tilde{\tau}$, and to find the properties of the resulting areas. Let us apply the prederivative operator along $Y(p, q)$ on the boundary (9.23), which converted into the corresponding differential chains reads

$$P_{Y(p,q)} \partial \tau = P_{Y(p,q)} \gamma \ .$$

From Cartan's magic formula (9.20), we observe that

$$\partial P_Y = \partial E_Y \partial + 0 \ , \quad \text{and} \quad P_Y \partial = 0 + \partial E_Y \partial \ ,$$

so the boundary operator and the prederivative operator commute, that is, $\partial P_Y = P_Y \partial$. If we use this property on the formula we have obtained, it follows that $P_{Y(p,q)} \partial \tau = \partial (P_{Y(p,q)} \tau) = P_{Y(p,q)} \gamma$. In other words, we have built an object $P_{Y(p,q)} \tau$ whose mapping under the boundary operator is the prederivative of the differential 1-chain dual to the supporting curve γ. For each τ, we define this geometric object $P_{Y(p,q)} \tau$, which is a differential 2-chain, the starting set for the solution of the Plateau problem, and denote it by Σ. Of course, it is not unique, because on the one hand the chain–cell duality relations do not provide uniqueness, and on the other hand there is a whole set of differentiable 2-chains τ spanning γ to start with. We mention that the elements Σ have the property

$$\partial \Sigma = P_{Y(p,q)} \gamma \ ,$$

and because all the operators involved in this closed equality relation are continuous, we know that the set of Σs is a closed set under the topology of the norms $\| \cdot \|_r$ defined and used throughout this section.

In order to define the area of these elements Σ, we introduce a differentiable 2-form ω and write

$$\mathcal{A}(\Sigma) = \int_{P_{Y(p,q)} \tau} \omega = \int_\tau Y(p,q)(\omega) = \int_\tau \mathcal{L}_{Y(p,q)}(\omega) \ ,$$

where we have used (9.21) and where $\mathcal{L}_{Y(p,q)}$ is the Lie derivative. To continue, using the property of the volume form $\mathrm{d}V$ and the properties of the vector field $Y(p,q)$ defined correspondingly, we have $\mathcal{L}_{Y(p,q)}(\omega) = Y(p,q)(\omega) = \mathrm{d}V \perp Y(p,q)$. By substituting this last relation in the integral equation above, we have finally

$$\mathcal{A}(\Sigma) = \int_\tau Y(p,q)(\omega) = \int_\tau \mathrm{d}V \perp Y(p,q) = \int_{E_{Y(p,q)}(\tau)} \mathrm{d}V \ , \tag{9.24}$$

where in the last integral we used (9.15). This last integral is the volume of the differential 3-cell obtained by extrusion of the differentiable 2-cell with the vector field $Y(p,q)$, and it is a Riemann volume integral. Hence, it is a continuous bounded operator, and consequently the area of the Σ object is a continuous function of Σ, so the set of areas of all such possible Σ objects is closed and positive (it is defined by positively oriented area forms). Such a set of positive numbers always has a lower bound, so according to the theory of real numbers, it has a greatest lower bound (an infimum) and this one is unique. Again, by continuity, the element Σ which guarantees this infimum area is itself unique.

We have sketched the proof and construction of the existence and uniqueness of the surfaces Σ minimizing the area of all surfaces spanning the given curve γ. There are a few more technical details for a complete and rigorous proof, e.g., for any closed curve γ' linking with γ with link number 1, the 'second' constraint (mentioned above) requires γ' to intersect the support of such surfaces at a point. This constraint prevents this area $\mathcal{A}$ from being zero, so it eliminates

the trivial empty solution. In addition, a complete proof should show that all these constructions are independent of the choice of the point q.

9.4 3D Drops

The best theoretical models and lab experiments for the study of free interfaces are liquid droplets. In all their possible forms, viz., drops, shells, bubbles, and antibubbles, they represent a fruitful symbiosis of topological, geometrical, and dynamical features in one compact system. Drops can be studied through simple physical models endorsed by innumerable natural examples, and from a theoretical standpoint, they provide favorable working spaces for mathematical approaches. The drop surface is in general smooth, compact, non-self-intersecting, and regular. Liquid drops, as well as bubbles and liquid shells, have always been a good source of inspiration and a straightforward experimental resource.

Through the help of topological fluid dynamics (see Sect. 9.1), researchers have discovered that Lie group constructions enable a unified approach to a wide variety of different hydrodynamical systems, from the deformation of the water surface in a rotating bucket to the Euler equation, to mixing problems, chaos, and turbulence. Other more recent applications include extinct neutron stars or accelerators in plasma dynamics.

Drops offer a natural framework for studying flows in Riemannian manifolds with free boundary, with or without force fields (spherically symmetric like self-gravity or electrostatic, axisymmetric like rotating drops, or uniform like gravity, etc.). The flow in such domains with boundaries involves modification of the governing Euler equations in order to keep the boundary an invariant set for the velocity field. The mathematical procedure when approaching flows within free boundaries is to handle the variation of the metric through the variation of the shape of the boundary [164, 286, 287].

The minimal ingredients of most general drop models are a smooth, compact, and connected manifold with boundary, embedded in physical space (usually $\mathbb{R}^n$), and the existence of a flow inside this manifold, tangent to the boundary and with the property of *incompressibility*. These ingredients are enough to build a mathematical model that can classify the allowable boundaries, showing that the flow inside depends only on the boundary and that the boundary can be obtained from stability and extremum principles.

The fundamental mathematical features of liquid drops can be found in the developments of modern geometrical hydrodynamics during the 1980s. As Modin and coauthors show in [284], beginning in 1966 [142, 292], Arnold and his followers demonstrated that Euler's equation for an ideal fluid is the geodesic equation on the group of volume-preserving diffeomorphisms with respect to the right invariant L^2 metric.

The differential geometry model of 2D and 3D incompressible liquid drops were established in rigorous Hamiltonian form by sustained work due to Zakharov [293],

Miles, Benjamin, Arnold, Marsden, Ratiu, Lewis, Shepherd, Bridges, Crawford, and others. More information can be found in the two main articles [167, 294]. A comprehensive review of the topic is given in Morrison's paper [295]. Good support for Hamiltonian systems, bifurcations, and Lyapunov procedures can be found in [296] (a more differential geometry orientation), or in the book [297] (a more functional analysis orientation) and [169].

In order to introduce a Hamiltonian structure for free surface incompressible fluids (essentially free drops), one needs to generate a Poisson bracket. The dynamic variables for this free surface Hamiltonian system are the Euler (spatial) divergence-free velocity field $\boldsymbol{v}(\boldsymbol{r}, t)$, $\nabla \cdot \boldsymbol{v} = 0$, and the boundary itself as a compact surface without boundary smoothly embedded in $\mathbb{R}^3$, denoted by Σ and with unit normal $\boldsymbol{n}$. The free surface is the boundary of a compact domain $\Sigma = \partial D_\Sigma$. Let us denote their space by $\mathcal{N} = \{\boldsymbol{v}, \Sigma\}$. According to the Weyl–Hodge theory [167], we can always express

$$\boldsymbol{v} = \boldsymbol{w} + \nabla \Phi \,, \quad \nabla \cdot \boldsymbol{w} = 0 \,, \quad (\boldsymbol{w} \cdot \boldsymbol{n})_\Sigma = 0 \,, \quad \triangle \Phi = 0 \,, \quad \frac{\partial \Phi}{\partial \boldsymbol{n}} = (\boldsymbol{v}, \boldsymbol{n}) \,.$$

We have $\mathcal{N} = \mathcal{N}' = \{\boldsymbol{w}, \Phi, \Sigma\}$. For any function $F : \mathcal{N} \to \mathbb{R}$, we define the following functional derivatives:

$$D_{\boldsymbol{v}} F \cdot \delta \boldsymbol{v} = \int_{D_\Sigma} \left(\frac{\delta F}{\delta \boldsymbol{v}}\bigg|_\Sigma , \delta \boldsymbol{v} \right) \mathrm{d}^3 \boldsymbol{x} \,, \tag{9.25}$$

$$\frac{\delta F}{\delta \Phi} = \left(\frac{\delta F}{\delta \boldsymbol{v}}, \boldsymbol{n} \right) , \tag{9.26}$$

$$D_\Sigma F \cdot \delta \Sigma = \int_\Sigma \frac{\delta F}{\delta \Sigma} \delta \Sigma \, \mathrm{d}A \,, \tag{9.27}$$

where the subscript Σ inside the volume integral means that the functional derivative holds Σ fixed. $\delta \Sigma$ is the normal variation of Σ (see the Appendix 1), and by the incompressibility condition, we also have

$$\int_\Sigma \delta \Sigma \, \mathrm{d}A = 0 \,.$$

In these and future equations mentioned in Sect. 9.4, we use the scalar product in the $L_2(D_\Sigma), L_2(\Sigma)$ sense, with the symbol $(\cdot\,, \cdot)$.

The Hamiltonian considered for the liquid drops is the traditional kinetic energy plus surface potential energy, viz.,

$$H[\boldsymbol{v}, \Sigma] = \frac{\rho}{2} \int_{D_\Sigma} |\boldsymbol{v}|^2 \mathrm{d}^3 \boldsymbol{x} + \gamma \int_\Sigma \mathrm{d}A \,, \tag{9.28}$$

where γ is the coefficient of surface tension and ρ a constant density. The functional derivatives of the Hamiltonian have the form

$$\frac{\delta H}{\delta \boldsymbol{v}} = \boldsymbol{v} \ , \qquad \frac{\delta H}{\delta \Phi} = \left(\frac{\delta H}{\delta \boldsymbol{v}}, \boldsymbol{n} \right) = (\boldsymbol{v}, \boldsymbol{n}) \ ,$$

and

$$\frac{\delta H}{\delta \Sigma} = \frac{\rho}{2} |\boldsymbol{v}|^2 + 2\gamma H \ .$$

It is straightforward to prove that the Hamilton equation for (9.28) has the form of the Euler equation for incompressible flow:

$$\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} = -\nabla P \ , \qquad \nabla \cdot \boldsymbol{v} = 0 \ , \tag{9.29}$$

together with the kinematic boundary conditions:

$$\frac{\partial \Sigma}{\partial t} = (\boldsymbol{v}, \boldsymbol{n}) \ , \tag{9.30}$$

and dynamic boundary condition

$$P_\Sigma = 2\gamma H \ , \tag{9.31}$$

where H is the mean curvature of the surface oriented inwards. A similar Poisson structure can also be introduced in terms of the independent variable $\boldsymbol{w}, \Phi, \Sigma$.

9.5 Rotation of 3D Drops

If we look for the origins of interest in rotating liquid surfaces, we find that there is probably no lower bound on the fluid research timeline. The subject probably became of interest with Newton's model for the shape of the Earth (1687) as a homogeneous gravitating liquid drop. He proposed the oblate spheroid as the shape of a gyrostatic equilibrium configuration. But Plateau was the first to raise the problem of finding the minimal area surface supported by a given curve, i.e., soap film problems [290], stationary or in rotation. The starting point for this investigation was observation of the evolution of rotating drops of oil in a mixture of water and alcohol, some time around 1840 (Plateau was nearly blind and he was interested in the physics of eyes and colors).

A large number of studies came in the wake of Plateau's experiments, seeking to clarify the dynamics and stability of rotating liquid droplets. On its 150th anniversary, the problem of finding equilibrium configurations for a rotating mass

of liquid shows a record of continuous progress, initially motivated by genesis of the Earth, eye and color studies, and later on, understanding the nuclear fission (and radioactivity) of heavy nuclei. More progress came with investigations into the evolution of stars, and later the study of neutron star dynamics and black holes, not to mention the development of nanoscience and nanofluids, printing technologies, and finally the field of hot dense nuclear matter, and especially the shapes of quark–gluon drops. However, in the last decade, the problem was still receiving new and unexpected geometrical solutions, motivated by rigorous mathematical questions from geometric analysis, as well as huge computer codes [291, 298–300].

The drop shapes initially considered were axisymmetric and simply connected, while later on non-axisymmetric shapes and non-simply-connected (toroidal) shapes were discovered to be stable. Pioneering papers from Plateau, Lord Rayleigh [290], Chandrasekhar [301], Ross [302], Brown and Scriven [303], and Ungar and Brown [304], etc., made numerical studies of the stability of different rotating drop shapes and compared their results with more and more sophisticated experiments. The experimental challenges of isolating the drop in rotation were met by suspending drops in buoyancy-free liquids at the price of increasing viscosity interactions, diamagnetic or acoustic levitation, gravity-free experiments in space labs, and now contemporary experiments on Leidenfrost drops. Current reviews and the latest results in computation and experiments can be browsed in [291, 299, 300, 305, 306].

In this section, we limit our study to the description of equilibrium shapes of liquid drops in rotation without considering gravitational (self or exterior) or electromagnetic forces, evaporation, or other nonlinear surface effects. In the case of the drop, shapes are governed by the equation of equilibrium at the interface between internal pressure, surface tension, and centrifugal forces. It is known that no steady motion can exist unless the liquid rotates as a rigid body with a constant angular velocity Ω, so we will only consider this case. We also note that it has been proved in several papers (see, for example, [307]) that the full dynamics of a 3D axisymmetric Euler system (inviscid liquid) in which the fluid does not move like a rigid solid, and any solution initiated from axisymmetric smooth initial data, blows up in finite time at points on the axis of rotation. This situation arises mainly because the radial pressure increment is not consistent with the global regularity of the classical solution. The Euler equation for axisymmetric flow can be rewritten as a complex Riccati type of equation plus pressure term for $\boldsymbol{v} = v_r \boldsymbol{e}_r + v_z \boldsymbol{e}_z \to v_r + \mathrm{i} v_z$, whose derivative in the radial direction blows up in a finite amount of time. But we will not consider these situations in the remainder of this book.

Before continuing with globally rotated drops, we recommend the reader to begin with a simple study, namely a liquid in a rotating cylinder [245]. The physics is the same as in the drop case, except that the geometry is simple (planar 2D), but even in this case the ordinary differential equation for the liquid shape cannot be integrated exactly, and numerical computations are necessary. The axisymmetric gyrostatic shapes obtained are stable, as opposed to the vast majority of rotated drops, where the shapes are unstable.

As a general criterion it has been proved [298, 302, 303] that the drop shape is symmetric with respect to a plane perpendicular to the axis of rotation, usually intersecting the center of mass of the drop. This results from the need of the curvature to have absolute extremum points at all points placed farthest from the rotation axis. In the following, we denote by ρ the difference $\triangle\rho = \rho_{\text{liquid}} - \rho_{\text{gas}}$, if the drop is surrounded by a gas. We will obtain the equation for the equilibrium surface of a gravity-free incompressible drop in uniform rotation with angular velocity Ω. Let us parameterize the drop surface Σ with a smooth function $(u, v) \to \boldsymbol{r}(u, v) \in \Sigma \subset \mathbb{R}^3$. According to the equation for the second fundamental form in differential geometry (see more details in the Appendix 1 at the end of this chapter) we can compute the mean curvature of the surface as

$$H = -\frac{1}{2}\text{Tr}(d\boldsymbol{N}) = \frac{eG - 2fF + gE}{2(EG - F^2)} , \tag{9.32}$$

where Tr is the trace of the matrix representing the linear operator $d\boldsymbol{N}$, considering the unit normal as a smooth map $\boldsymbol{N} : \Sigma \to S^2$ (the so-called Gauss map).

Let us assume the existence of another surface Σ^ϵ, which is an infinitesimal 'normal' deformation from the original surface Σ along its normal, that is, a diffeomorphism

$$\boldsymbol{r}^\epsilon(u, v, \epsilon) = \boldsymbol{r}(u, v) + \epsilon h(u, v)\boldsymbol{N}(u, v) ,$$

with $0 < \epsilon \ll 1$ a small parameter and $0 < |h(u, v)| < 1$ a smooth bounded function defined on the original surface. After a short calculation, we obtain the area element of the deformed surface in the form

$$E^\epsilon G^\epsilon - (F^\epsilon)^2 = (EG - F^2)\big[1 - 4\epsilon hH + \mathcal{O}(\epsilon^2)\big] .$$

The last equation is just another expression of the formula for the *first variation of area*:

$$\frac{\text{d}}{\text{d}t}\text{d}A = H\text{d}A ,$$

where t, in this case, labels a smooth flow of the surface directed everywhere in the direction of the normal. In other words, the relative rate of change of the area as the surface evolves in the outward normal direction is the mean curvature.

From the thermodynamics side, it is known that the variation δU of the total energy U of the liquid drop (for our case in the absence of gravitational, electromagnetic, chemical, and phase transition effects) is given by

$$\delta U = T\delta S - P\delta V + \delta U_\Sigma + \delta U_{\text{cf}} ,$$

where T, S, P, V are temperature, entropy, pressure, and volume, respectively. The last two terms are responsible for the surface and centrifugal potential energies, respectively. With γ the coefficient of surface tension, we have $\delta U_\Sigma = \gamma \delta \mathcal{A}$. In the (noninertial) system of reference of the rotating drop, the fluid mass is subjected to the inertial centrifugal force $\mathrm{d}\boldsymbol{F}_i = -\boldsymbol{a}_i \mathrm{d}m$ and $\mathrm{d}F_i = \Omega^2 r_\perp \mathrm{d}m$, where $r_\perp$ is the distance from the rotation axis to the elementary mass $\mathrm{d}m$, e.g., given in spherical coordinates (r, θ, ϕ) by $r_\perp = r \sin\theta$. Because the effective centrifugal potential energy is minus the work done by the centrifugal force, which is $\mathrm{d}W_{\mathrm{cf}} = \Omega^2 \mathrm{d}(r_\perp^2)\mathrm{d}m$, we have

$$U_{\mathrm{cf}} = \Omega^2 \int \rho \mathrm{d}V \left(- \int_0^{r_\perp} r_\perp \mathrm{d}r_\perp \right) = -\frac{\Omega^2}{2} \int \rho r_\perp^2 \, \mathrm{d}V \, .$$

When the drop shape changes infinitesimally, what drives the variation of different terms in the energy of the drop is mainly concentrated at the drop surface, as a consequence of incompressibility and isentropy. Therefore, it makes sense to express all integral quantities of the energy in terms of surface integrals, and for arbitrary local shape variations to infer from there a differential condition holding at the drop surface. The procedure is based on the exterior derivative applied to the volume element $\mathrm{d}V$, namely $*\mathrm{d}\mathcal{A} = \mathrm{d} * \mathrm{d}V$, where $*$ is the Hodge dual operator.

Since the normal deformation of the surface is directed along the normal at any point of the surface, the fluid is incompressible, and changes of shape are rapid (adiabatic, $\mathrm{d}S = 0$), it follows that

$$\delta U = -\delta \int P \mathrm{d}V + \gamma \delta \int \mathrm{d}\mathcal{A} - \delta \int \frac{\Omega^2 r_\perp^2 \rho \mathrm{d}V}{2} \, ,$$

and from here the total energy for an incompressible adiabatic drop is

$$U = \gamma \int_\Sigma \mathrm{d}\mathcal{A} - \frac{\Omega^2}{2} \int r_\perp^2 \rho \mathrm{d}V \, . \tag{9.33}$$

We can convert the volume integral into a surface integral:

$$\delta \int P \mathrm{d}V = \iint \epsilon h P \sqrt{EG - F^2} \, \mathrm{d}u \, \mathrm{d}v \, .$$

At the same time,

$$\delta U_\Sigma = \gamma \int (\mathrm{d}\mathcal{A}^\epsilon - \mathrm{d}\mathcal{A}) \, ,$$

$$\delta U_\Sigma = \iint \sqrt{EG - F^2} \left(\sqrt{1 - 4\epsilon h H} - 1 \right) \mathrm{d}u \, \mathrm{d}v + \mathcal{O}(\epsilon^2) \, .$$

Combining all of these variations under the surface integral, at equilibrium ($\delta U = 0$), we obtain the Young–Laplace equation for the rotating (incompressible, gravity-free) liquid drop surface:

$$\left(P + 2H\gamma + \frac{1}{2}\Omega^2 r_\perp^2 \rho\right)_\Sigma = 0 \,, \tag{9.34}$$

where for P one can take the value of the pressure difference across the drop surface at the point $r_\perp = 0$ [302, 303].

Note that, according to the above definition (9.32) of the mean curvature, its value is negative for a convex surface, i.e., $H < 0$, and the mean curvature vector always points into the direction in which the area becomes smaller by deformation (i.e., towards the center if the surface is spherical [121]). From here, a common confusion occurs in almost all physics literature when (9.34) has a different sign in front of the term $2\gamma H = \gamma(\kappa_1 + \kappa_2)$, but we can resolve it in the following equation.

We introduce the scaling parameters L_0 and V_0 for distance and velocity, respectively, so that $x = \tilde{x}L_0$, $V = \tilde{V}V_0$. Then we have $t = \tilde{t}L_0/V_0$, and we can rescale the pressure $P = \tilde{P}(\rho V_0^2)$ (or $P = \tilde{P}(\nu\rho V_0)/L_0$) and the angular momentum $\mathcal{L} = \tilde{\mathcal{L}}(2R^3\sqrt{2\gamma\rho R})$. Consequently, we can rewrite the Laplace–Young equation (9.34) in the dimensionless form

$$\tilde{P}|_\Sigma = \frac{2}{\mathrm{We}}|\tilde{H}| - \frac{\tilde{\Omega}^2}{2}(\tilde{r}_\perp)_\Sigma^2 \,, \tag{9.35}$$

where

$$\mathrm{We} = \left|\frac{\rho V^2}{\gamma H}\right| = \frac{\rho V_0^2 L_0}{\gamma} = \frac{\rho\Omega_0^2 L_0^3}{\gamma} \sim \frac{\Omega_0^2 m}{\gamma} \tag{9.36}$$

is the dimensionless Weber number (see Table 9.1), defined as a measure of the relative importance of the fluid's inertia compared to its surface tension. We also mention that the Weber number is $\mathrm{We} = 8\Sigma$, with Σ the rotational Bond number used in traditional analysis of rotational shapes [301, 304]. In the rotational case, the third term on the right-hand side of (9.36) actually shows proportionality with the mass m of the drop. This quantity is useful in analyzing thin film flows and the formation of droplets and bubbles.

For slowly rotating (2 Hz) water drops of diameter around $R = 10\,\mathrm{mm}$, we have $\mathrm{We} = 10^{-2}$ and surface tension dominates (9.35), so the drops are pretty spherical. At 10 Hz, we have $\mathrm{We} = 10$, and at 100 Hz, we have $\mathrm{We} = 200$, in which case the centrifugal term almost completely balances the surface energy and the shapes become highly deformed. Similarly, bigger drops have stability shapes shifted towards less symmetric configurations.

In order to study the stability of different rotating shapes, one needs to introduce volume conservation through a Lagrange multiplier, and also to emphasize the imposed condition of constant rotation, in which the angular momentum of the drop

is not conserved and the Lagrangian functional is the one given by (9.33) plus the Lagrange multiplier term, viz.,

$$L_\Omega = \gamma \int_\Sigma \mathrm{d}\mathcal{A} - \frac{\rho\Omega^2}{2} \int r_\perp^2 \mathrm{d}V - k\left(\int \mathrm{d}V - V_0\right) . \tag{9.37}$$

Here, k is to be determined from the minimization procedure and V_0 is the initial (constant) volume of the drop. According to [303], the Lagrange parameter k from (9.37) represents the pressure difference across the drop surface at the axis of rotation, and the whole volume difference term $-k(V - V_0)$ measures the pressure energy of the liquid in the drop.

However, if the drop is in free rotation (as in gravity-free experiments in space, drops suspended by magnetic or ultrasonic levitation, or maintained frictionless by a flow of air, or Leidenfrost drops), the angular momentum $\mathcal{L}$ is conserved, and the Lagrangian functional to be minimized is no longer the one in (9.37). It must be changed to the corresponding Routhian [303] by a Legendre transform

$$L_0 = L_\Omega - \Omega \frac{\partial L_\Omega}{\partial \Omega} = \gamma \int_\Sigma \mathrm{d}\mathcal{A} + \frac{\mathcal{L}^2}{2\mathcal{I}} - k\left(\int \mathrm{d}V - V_0\right) , \tag{9.38}$$

where the angular momentum is

$$\mathcal{L} = \Omega\rho \int r_\perp^2 \mathrm{d}V ,$$

and the moment of inertia is given by

$$\mathcal{I} = \rho \int r_\perp^2 \mathrm{d}V .$$

It is simpler to study the energy expressions and their Lagrange multipliers in dimensionless form. In the fluid mechanics of rotating drops, the most important dimensionless parameter is the Weber number

$$\frac{\mathrm{We}}{2^{3/2}} = \sqrt{\frac{\rho\Omega^2 R^3}{8\gamma}},$$

where R is the mean radius of the undeformed drop, and Ω is the angular velocity of rotation of the drop as a solid body.

If we consider the rotation of a free drop involving the conservation of its angular momentum

$$\mathcal{L} = \frac{\Omega\rho R^5}{5} \int_0^{2\pi} \mathrm{d}\phi \int_0^\pi \tilde{f}^5 \sin^3\theta \, \mathrm{d}\theta , \tag{9.39}$$

we need to reduce the dimension of the angular momentum by dividing it by $2R^3\sqrt{2\gamma\rho R}$, that is,

$$\tilde{\mathcal{L}} = \frac{1}{5}\sqrt{\frac{\rho\Omega^2R^3}{8\gamma}}\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\tilde{f}^5\sin^3\theta\,\mathrm{d}\theta = \frac{\mathrm{We}}{10\sqrt{2}}\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\tilde{f}^5\sin^3\theta\,\mathrm{d}\theta\ .$$

By (9.38), the dimensionless Routhian becomes

$$\tilde{L}_0 = \frac{1}{4R^2}\int_\Sigma \mathrm{d}\mathcal{A} + \frac{\mathcal{L}^2}{8\gamma R^2}\left(\rho\int r_\perp^2\,\mathrm{d}V\right)^{-1} - \tilde{k}\left(\frac{1}{R^3}\int \mathrm{d}V - \tilde{V}_0\right), \tag{9.40}$$

where the term $(\dots)^{-1}$ is the moment of inertia of the drop. In spherical coordinates, the Routhian reads

$$\begin{aligned}\tilde{L}_0 = {} & \frac{1}{4}\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\tilde{f}\sqrt{(\tilde{f}^2+\tilde{f}_\theta^2)\sin^2\theta+\tilde{f}_\phi^2}\,\mathrm{d}\theta \\ & +5\tilde{\mathcal{L}}^2\left(\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\tilde{f}^5\sin^3\theta\,\mathrm{d}\theta\right)^{-1} \\ & -\frac{\tilde{k}}{3}\left(\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\tilde{f}^3\sin\theta\,\mathrm{d}\theta - 4\pi\right).\end{aligned} \tag{9.41}$$

In order to study the stability configurations of the drop, one needs to apply the calculus of variations to the resulting Lagrangian and/or Routhian, using the formulas from Appendix 2. To accomplish this, one must apply Theorem 16 and calculate the first Gâteaux (or the Euler equation) and the second Gâteaux differentials of these functionals, then study the conditions

$$\tilde{L}_\Omega' = 0\,, \quad \langle\tilde{L}_\Omega''[u]h,h\rangle \geq 0\,, \tag{9.42}$$

or

$$\tilde{L}_0' = 0\,, \quad \langle\tilde{L}_0''[u]h,h\rangle \geq 0\,, \tag{9.43}$$

for any h in a given space of functions and for u a solution of the corresponding Euler equation (9.35). Rotational shapes obtained by (9.42) and (9.43) have been studied in different ways in the literature.

One of the first rigorous approaches was introduced by Ross in his article [302], mainly based on Lord Rayleigh's earlier results in [290]. Ross considered only axisymmetric shapes, but he kept the possibility of having multiply-connected shapes. For a drop rotating around the z-axis at a fixed angular velocity Ω, his procedure is to choose a finite set of single-valued functions $r = f_j(z)$ to describe the single-valued components of the vertical cross-section through the drop. Here

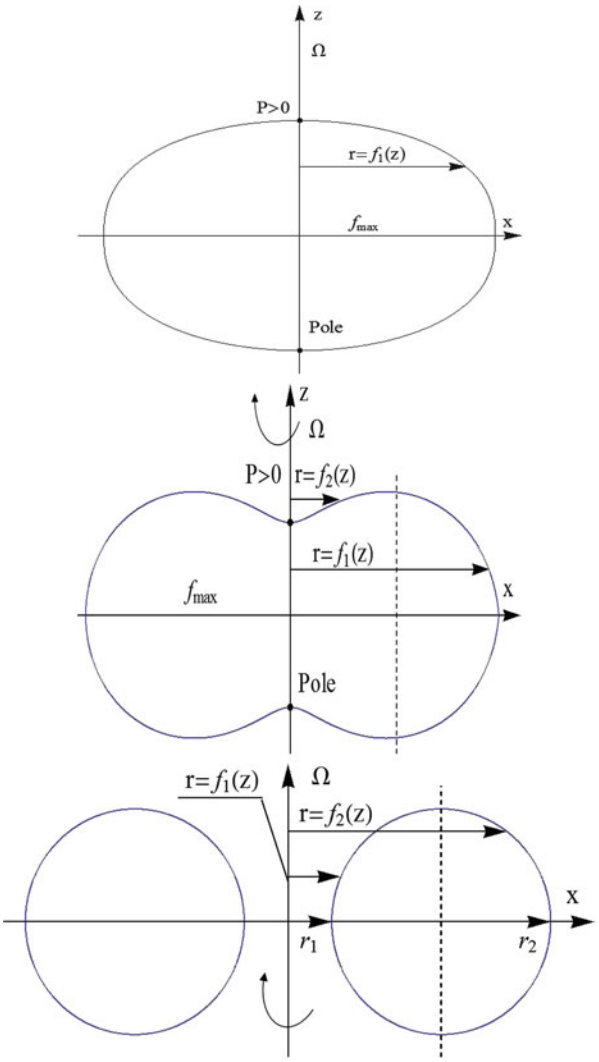


Fig. 9.9 Axisymmetric shapes of rotating drops. In the *upper right frame* the drop is concave at its poles, the Weber number is larger, and the pressure is negative at the poles. The *bottom row* represents a toroidal shape when the drop is one-connected

(r, z, ϕ) are the polar coordinates versus the rotation axis (see Fig. 9.9). From the first condition in (9.42), he obtained the nonlinear ordinary differential equation

$$\frac{1}{\sqrt{1+f'^2}} = \frac{kf}{2\gamma} + \frac{\rho\Omega^2 f^3}{8\gamma} - \frac{A}{f}\,, \tag{9.44}$$

where A is an integration constant which becomes zero if the free surface intersects the axis of rotation. In this equation, k is again the Lagrange multiplier from (9.37), which is equal to the pressure change across the drop surface at the intersection between the rotation axis and the drop (at the poles). There are two mathematically distinct cases. One is when the drop is simply connected, the rotation axis has nonempty intersection with the drop volume, and $A = 0$ (presented in the upper row in Fig. 9.9). The second case is when the drop is multiply-connected (see the bottom row in Figs. 9.9 or 9.12) and the rotation axis does not intersect the drop. In this case, we have ($A \neq 0$) and we will discuss it further in this section.

In the upper drawing of Fig. 9.9, when the drop contour physically intersects the rotation axis, the condition of uniqueness of the intersection point with the z-axis results in the occurrence of a Lagrange multiplier $k = P$, with the physical meaning of the external pressure at that point. Different solutions of this differential equation can be classified with the help of the Weber number. The pressure across the drop surface at its poles (intersection with the rotation axis) satisfies the condition

$$P = \frac{\gamma(8 - \mathrm{We}^2)}{4f_{\max}} , \tag{9.45}$$

where $f_{\max}$ is the largest radius of the drop, usually in its middle equatorial plane. For $\mathrm{We} < 2^{3/2}$, the pressure at the poles is positive and the drop surface at the poles is convex, similar to an oblate ellipsoid (see Fig. 9.9 upper left). For Weber numbers greater than $2^{3/2}$, the pressure at the poles is negative, so the drop is concave in a neighborhood of its intersection with the axis, and if the Weber number increases, it tends to break up (see Fig. 9.9 upper right or Fig. 9.11). This study is rather qualitative and does not investigate the specific stability of various shapes. However, some important conclusions can still be inferred from such a simple model.

By solving (9.44) for $A = 0$, one obtains the following four important numerical observations:

1. There exists one surface of revolution corresponding to each angular momentum, with less than a critical value:

$$\mathcal{L}_c = 2.8506\pi\sqrt{2\rho\gamma r_0^7} , \quad \text{with } V_0 = \frac{4\pi r_0^3}{3} .$$

2. The maximum angular speed occurs when the difference between kinetic and surface energies is minimal, and this limit depends only on constants determined by the material and the drop volume. Ross observed the following:

$$\Omega_{\max} = 0.7540\sqrt{\frac{8\gamma}{\rho r_0^3}} .$$

3. For each value of the angular velocity less than its maximum, but greater than the limit

$$\Omega_c = \frac{\pi}{4}\sqrt{\frac{\rho r_0^3}{\gamma}}\,,$$

there are two possible surfaces of revolution.
4. The kinetic energy increases with increasing angular momentum and the drop collapses when its kinetic energy becomes

$$K_c = 4.0316\pi\gamma r_0^2\,.$$

When the drop does not intersect the rotation axis, we have $A \neq 0$ and the shape becomes toroidal (see the lower row in Figs. 9.9 or 9.12). Let r_1 be the radius of the hollow part of the drop (major torus radius), and r_2 the maximum equatorial radius of the drop (major radius plus minor diameter). According to Ross, the condition for the breakup of a simply-connected drop into a toroidal shape is given by

$$\mathrm{We}^2\left(1-\frac{r_1}{r_2}\right)\left(1+\frac{r_1}{r_2}\right)^2 < 32\,.$$

A more detailed numerical investigation for driven and isolated rotating drops, also considering non-axisymmetric shapes, is presented in [303], where stability is taken into consideration through (9.42) and (9.43). Brown and Scriven numerically tested the stability of various shapes by describing them in polar coordinates using a set of orthonormal functions $\Phi^i(\theta,\phi)$ generated by a Hermite bicubic basis. They proved that, as the angular velocity rises, the shapes evolve from the perfect sphere, when the drop is at rest, through oblate shapes to biconcave shapes. The results of the stability calculations in the case of drops rotated with constant angular velocity are shown in Fig. 9.10. They identify three bifurcation points and two turnaround points. Hence, the points on the solid curve of increasing angular velocity are shapes stable to axisymmetric perturbations. The points indicated by dashed curves are unstable to axisymmetric deformations that shift liquid away from the axis of rotation and therefore increase the angular momentum, at constant angular velocity. Two such branches meet in a turning point at a certain maximum value of the angular momentum ($L_{\max}$), where the axisymmetric equilibrium shape is neutrally stable to a certain axisymmetric perturbation. There are limits in angular velocity above which all axisymmetric shapes become unstable to two-lobe perturbation, three-lobe perturbation, and so on, toward higher lobe multiplicities. Still within this limit for angular velocity, but for angular momentum values higher than $L_{\max\,\mathrm{III}}$, the drop becomes unstable to multiply-connected shapes like the torus.

When the drops are isolated and there is conservation of angular momentum, the axisymmetric perturbations that cause instability on the descending solid-line branch in Fig. 9.10 at fixed values of the angular velocity do not conserve angular

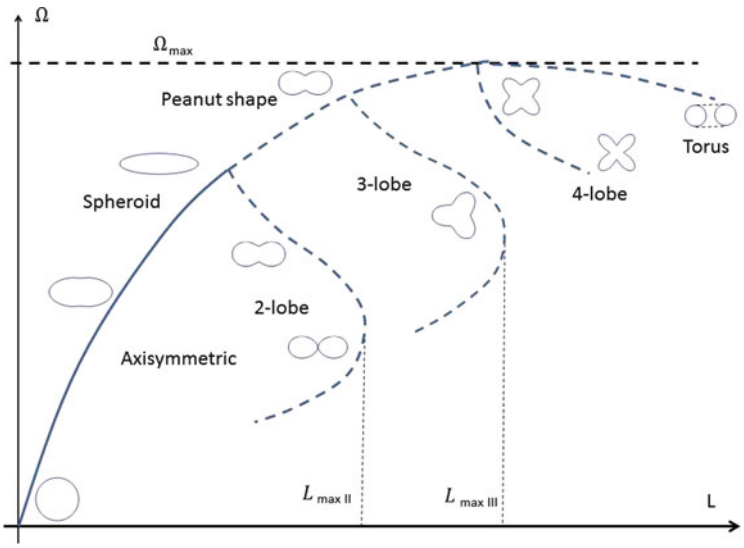


Fig. 9.10 Shape families, their stability and bifurcations for drops with a constant angular velocity are presented in an angular momentum–angular velocity diagram. From data in [298, 303]

momentum, and so do not arise on an isolated drop. However, the isolated drop also proves to be unstable to the two-, three-, and four-lobed perturbations. The lower limit of the unstable multiple-lobe shapes points toward breakup of the drop, but these limiting situations are not studied in the papers cited above. It seems that, without including viscosity, it is difficult to prove the breakup limit theoretically [308].

According to Brown and Scriven's calculations [303], none of these perturbations change either the angular momentum or the moment of inertia of the drop, so they leave its angular velocity unaltered. Results show that the stability criteria for these perturbations for axisymmetric shapes are identical. The dashed curves that primarily bifurcate under the main axisymmetric shapes (solid curve) show secondary bifurcation points connecting shapes with different numbers of lobes. These types of predicted shapes for rotating drops (see, e.g., Figs. 9.11 and 9.12) have been observed since the 1980s and previously recorded [298, 305], including studies of the flattening of slowly rotating drops and the generation of toroidal and multi-lobed shapes at higher rotation rates.

A recent and strictly analytical proof for the bifurcation behavior of the equations governing the rotating droplets problem shows that there is an interesting and rich variety of solution families [299]. Some of the results are also presented in Figs. 9.16, 9.17, 9.18, and 9.19. In [299], Ω is used as bifurcation parameter,

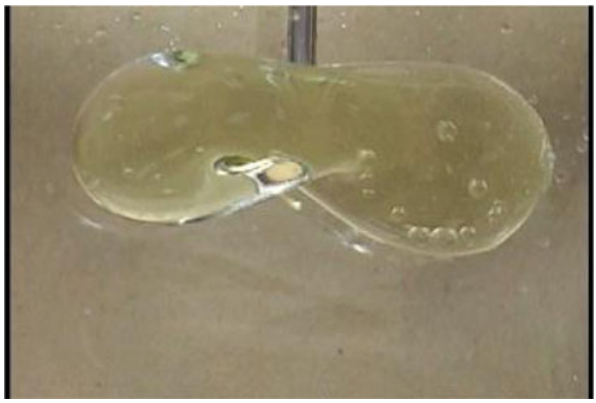


Fig. 9.11 Experiments with rotating drops of castor oil: peanut shape. Photo courtesy of the ACE-REU Program, Georgia Institute of Technology

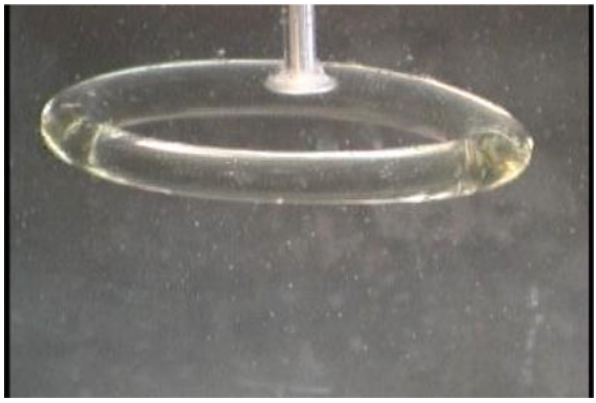


Fig. 9.12 Experiments with rotating drops of castor oil: toroidal shape. Photo courtesy of the ACE-REU Program, Georgia Institute of Technology

and (9.42) and (9.43) are used to compute stationary points of L_Ω and L_0. The dynamical equilibrium condition for the rotating surface in (9.34) can be rewritten in the form

$$2H\gamma = -\gamma \triangle_\Sigma = -k - \frac{1}{2}\Omega^2 r_\perp^2 \rho \,, \tag{9.46}$$

where we use the relation between the mean curvature and the surface Laplace operator, and the fact that the pressure evaluated at the poles is the Lagrange multiplier k used for the volume change. The first variation of the two functionals in question, i.e., L_Ω, L_0, in the Gâteau differential sense (see Appendix 2) is calculated by considering that the Lagrangian and the Routhian are functionals depending on the shape of the drop Σ (or its parameterization function $\Phi : D \subset \mathbb{R}^2 \to \Sigma \subset \mathbb{R}^3$) and the Lagrange multiplier k, so that we have $L_\Omega[\Sigma, k]$ and $L_0[\Sigma, k]$, respectively,

while Ω is the bifurcation parameter. For an infinitesimal variation $0 < \epsilon \ll 1$ of Φ with a deformation $\epsilon\varphi$, and $k + \epsilon K$, we have

$$L'_\Omega = \frac{\mathrm{d}}{\mathrm{d}\epsilon} L_\Omega \big[\Sigma(\Phi + \epsilon\varphi), k + \epsilon K\big]_{\epsilon=0} \tag{9.47}$$
$$= -\int_\Sigma \left(\gamma \triangle_\Sigma \Phi \cdot \varphi + \frac{\rho\Omega^2 r_\perp^2}{2}\varphi_\perp + k\varphi_\perp\right)\mathrm{d}A - \frac{K}{3}\left(\int_\Sigma \Phi_\perp \mathrm{d}A - 4\pi\right),$$

and as shown previously, a stationary solution (Σ, k) also satisfies the associated (nonlinear) Young–Laplace equation (9.34), or (9.46). In a similar fashion, one can calculate the second variation with respect to two orthogonal perturbations φ, ψ, K, J with $\varphi \cdot \psi = 0$:

$$\begin{aligned} L''_\Omega = &-\gamma \int_\Sigma \psi \cdot \triangle_\Sigma(\varphi)\mathrm{d}A - 2\gamma \int_\Sigma \psi_\perp \varphi_\perp \Big[\triangle(\varphi_\perp) + \varphi_\perp(\triangle_\Sigma \boldsymbol{n})_\perp\Big]_\perp \mathrm{d}A \\ &-\rho\Omega^2 \int_\Sigma \varphi_\perp \psi_\perp \boldsymbol{r} \cdot \boldsymbol{n}\, \mathrm{d}A - J\int_\Sigma \varphi_\perp \mathrm{d}A - K\int_\Sigma \psi_\perp \mathrm{d}A \\ &+2\int_\perp H\psi_\perp\varphi_\perp \left(2H - \frac{1}{2}\rho\Omega^2 r_\perp^2 - k\right)\mathrm{d}A\,, \end{aligned} \tag{9.48}$$

where $\boldsymbol{n}$ is the unit normal to Σ. Actually, the last integral on the right-hand side of the last equation is zero if the second variation is evaluated for a solution to the Young–Laplace equation. Similar formulas hold for the first and second variation of the Routhian functional, but we do not present them here since they can be found in detail in [299].

Using several numerical algorithms, the authors in [299] obtained a large collection of most intriguing and nonlinear drop shapes, sub-bifurcating from one another in an almost fractal cascade. For each calculation, the perturbations apply to some initial type of shape, which can be spheroidal or multi-lobed. The results obtained by starting with a unit sphere shape are shown in Fig. 9.13. From the 'main' axisymmetric branch (as it is called in [299]), several branches bifurcate, and further sub-bifurcate into sub-branches. In this article, the authors found multi-lobe shapes that were not identified by previous researchers, because in those previous studies the authors considered (artificially) only shapes with meridional reflective symmetry. Moreover, it seems likely that six lobes is the highest lobe asymmetry for rotating drops, because all numerical computations show no more intersections with the seven-lobe branch than any other branch or sub-branch.

A little bit below the value $\tilde{\mathcal{L}} = 2$, the two-lobed sub-branch sub-bifurcates around $\tilde{\Omega} = 0.3$–0.4 into an attractor configuration having non-axisymmetric shapes with one-lobe only, like the one exemplified in Fig. 9.16, a situation called 'winding-up'. The studies [303] reported reconnections of the peanut-shaped branch to the two-lobed branch (shapes looking like the example in Fig. 9.17) at a smaller

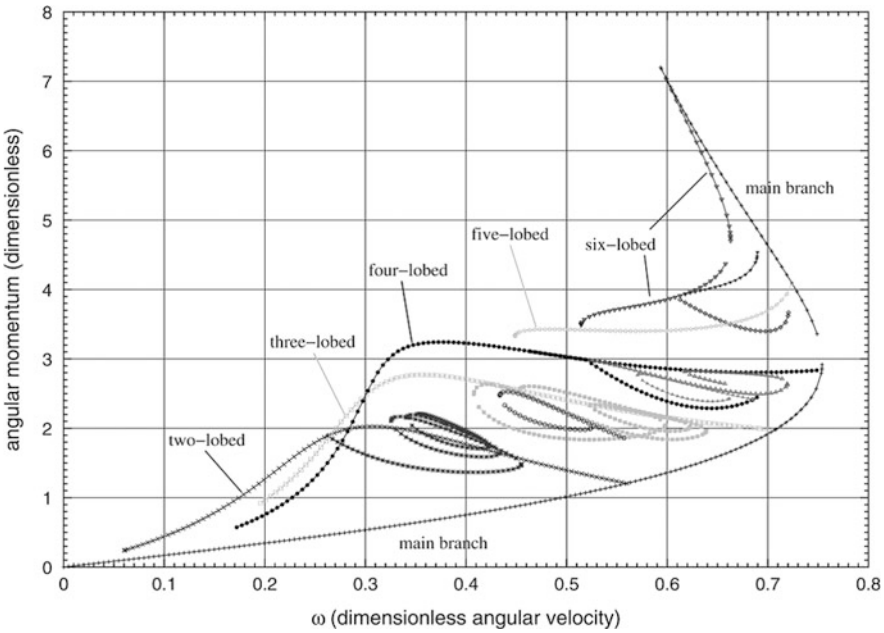


Fig. 9.13 Bifurcation diagrams for the spheroidal family in the (dimensionless angular momentum–dimensionless angular velocity) phase diagram $(\tilde{\Omega}, \tilde{\mathcal{L}})$. In this figure, $\tilde{\Omega}$ is denoted by ω. From Fig. 7 in [299], by permission of Oxford University Press (Oxford Journals)

value of the angular velocity. In the same way, in [299], it was shown that a similar bifurcation further branches off towards larger values of Ω.

Like the two-lobed branch, the three-lobed (Fig. 9.14) and higher-lobed branches each seem to approach a singular limit consisting of spheres of equal size, as Ω tends to zero, pretty much following the conjecture made in [303]. Of more interest is the behavior of the sub-branch which bifurcates from the three-lobed family at $\tilde{\Omega} \sim 0.42$. This sub-branch leads to both smaller and larger values of Ω and seems to be tangential to the three-lobed branch in the $(\tilde{\Omega}, \tilde{\mathcal{L}})$ phase diagram. The sub-branch repeats its winding behavior toward the sub-branch of the two-lobed family.

The five-lobed (Fig. 9.18) and six-lobed (Fig. 9.19) bifurcations differ from the two-, three-, and four-lobed variants because they do not seem to approach limit surfaces consisting of spheres of equal size. Instead, they wind up in the phase space of Fig. 9.13, similarly to the two-lobed sub-branch. The six-lobed shape intersects itself at the bifurcation point: the spheroidal family can be extended past its meeting point with the annular family where the thickness of the drops at the axis of rotation approaches zero. Past that point on the main spheroidal branch, the upper half of the drops reaches below the equatorial plane and intersects the lower half.

These multi-lobed shapes obtained numerically bear a strong resemblance to experimental shapes of 2D rotating drops and splash events of liquid drops, like those presented in Figs. 9.16, 9.18, and 9.19.

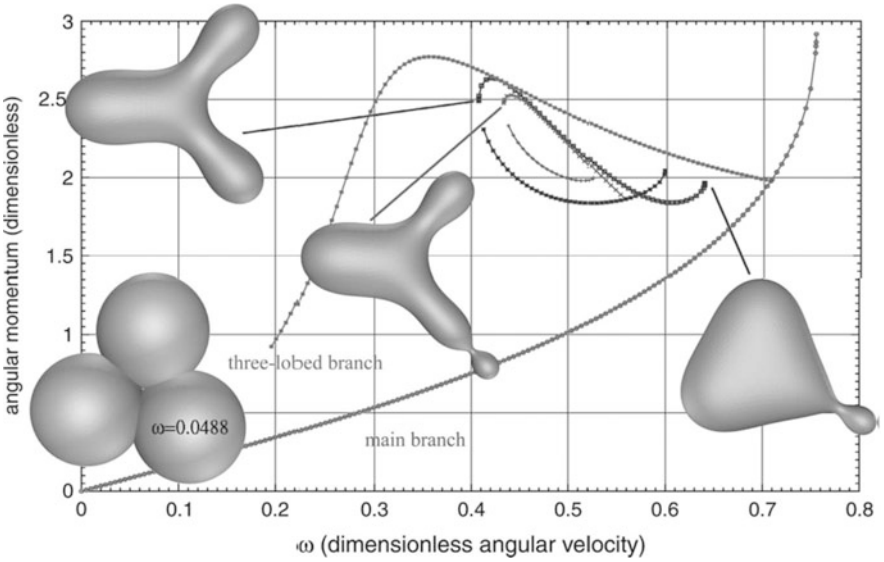


Fig. 9.14 A part of the bifurcation diagram Fig. 9.13 which shows a sub-sub-branch of the three-lobed family breaking the meridional reflective symmetry. From Fig. 13 in [299], by permission of Oxford University Press (Oxford Journals)

In addition to simply-connected shapes, the same approach generates a multitude of new toroidal or annular shapes (see Fig. 9.15). We also mention a remarkable new type of shape called the pearl necklace, i.e., the ring of tangential spheres shown in Fig. 9.15. According to a conjecture stated in [299], it keeps dividing into more and more tangential spheres until it reaches a singular limiting surface.

Equilibrium rotating shapes like those obtained by numerical computations (and presented in Figs. 9.16, 9.17, 9.18, and 9.19) have a similar purely theoretical counterpart in the mathematical paper by Kapouleas [300]. He provides analytical proof of the existence of shapes for small enough angular momentum [close to the origin of the $(\tilde{\Omega}, \tilde{L})$ phase plane] with an aspect of multiple spherical lobes connected by thin necks. Such shapes belong to the same class as the one with three tangent spheres shown in the phase diagram of Fig. 9.14 for $\tilde{\Omega} = 0.0488$, or the pearl necklace shape appearing with the annular shapes in the phase diagram of Fig. 9.15. This study provides a rigorous proof for the existence of even more complicated equilibrium configurations, all symmetric with respect to the plane orthogonal to the axis of rotation.

These exotic shapes look like a central spherical lobe around which m strings, each containing n spherical lobes, are symmetrically distributed. These shapes are said to be of type (m, n) (see Fig. 9.20). In order to prove that such (m, n) surfaces with $1 \leq m \leq 6$, $n \geq 1$ exist as solutions of the equilibrium equation (9.34), one needs to use building blocks made of pieces of spheres and necks that form together so-called Delaunay surfaces, i.e., surfaces with constant mean curvature.

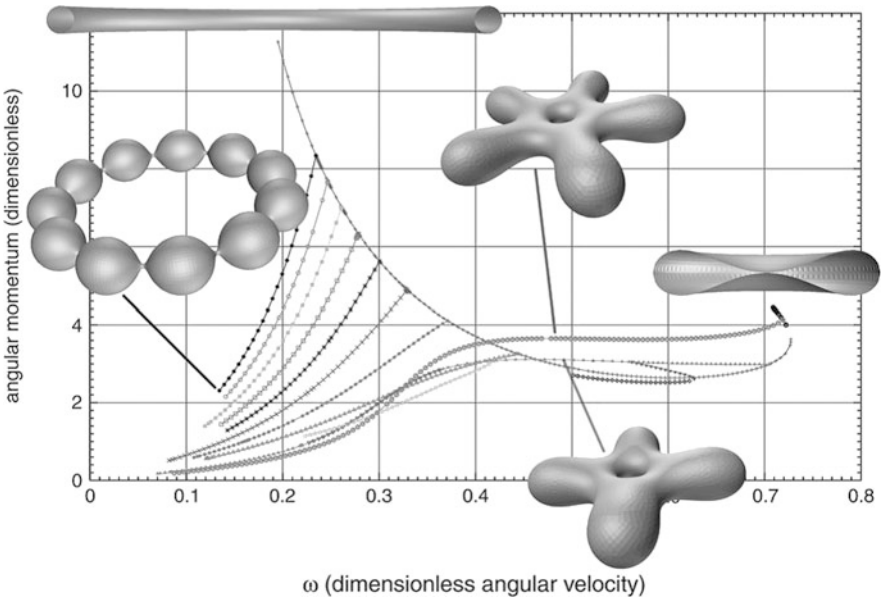


Fig. 9.15 Bifurcations from the axially symmetric family of annular shapes. The meeting point of the annular and spheroidal families where the inner radius of the torus approaches 0 occurs at $\tilde{\Omega} = \sqrt{2}/2$ (denoted ω in this figure). From Fig. 15 in [299], by permission of Oxford University Press (Oxford Journals)

Fig. 9.16 Non-axisymmetric shape with one lobe. Courtesy of the *Drop Gallery* of Dr. Claus-Justus Heine and Dr. Gerd Dzuik, Albert-Ludwigs-University of Freiburg, Dept. Applied Mathematics

These shapes are then combined with transitional 'annuli' of zero mean curvature, and finally the parameters are adjusted to satisfy the equilibrium equation (9.34). We use the same notation Σ for the drop surfaces and $\mathcal{D}$ for the domain defining the drop body, $\Sigma = \partial\mathcal{D}$, also $|\Sigma|$ for the area of the drop surface, and $V_0 = |\mathcal{D}| = 4\pi R^3/3$ for the constant drop volume. In order to simplify the calculations, we rewrite the Young–Laplace equation in a slightly different dimensionless form, using the angular momentum instead of the angular velocity:

$$-H(\boldsymbol{r}) = \frac{P}{2\gamma} + \frac{\mathcal{L}^2\rho}{4\mathcal{I}^2\gamma} r_\perp^2 \,, \tag{9.49}$$

Fig. 9.17 Two-lobed shapes. Courtesy of the *Drop Gallery* of Dr. Claus-Justus Heine and Dr. Gerd Dzuik, Albert-Ludwigs-University of Freiburg, Dept. Applied Mathematics

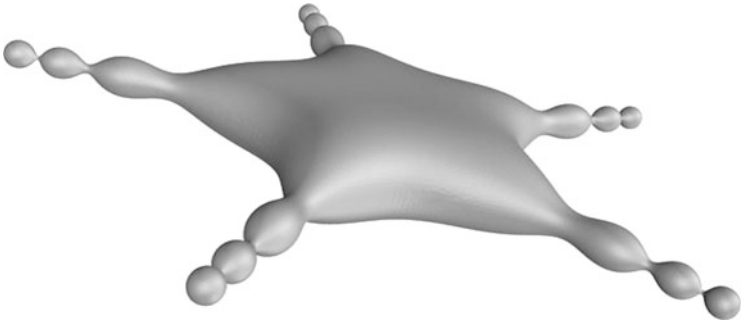


Fig. 9.18 Five-lobed shapes. Courtesy of the *Drop Gallery* of Dr. Claus-Justus Heine and Dr. Gerd Dzuik, Albert-Ludwigs-University of Freiburg, Dept. Applied Mathematics

where $\mathcal{I}$ is the moment of inertia of the drop configuration about the rotation axis, and the rest of the terms have the same meaning as before, i.e., H is the mean curvature with the convention that $H = -1$ for the unit sphere, $\mathcal{L}$ is the angular momentum, P is the pressure across the drop at its poles, and $r_\perp$ is the distance from the rotation axis to the point $\boldsymbol{r} \in \Sigma$.

The theorem in [300], which proves that the lobe-plus-neck surfaces are solutions to the Young–Laplace equation, does not contain an explicit procedure for constructing the solutions. More details and the expressions for the solutions in terms of coordinates can be found in [309]. Let us denote by M the first approximant surface for the solution of (9.49), parameterized as an immersion $X : \Sigma \to \mathbb{R}^3$, and by $\epsilon \in C^\infty(M)$ a smooth deformation of Σ in the direction of the normal to Σ. If we

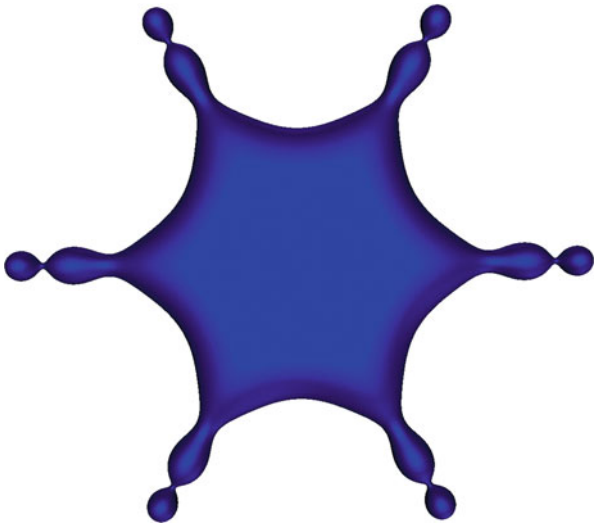


Fig. 9.19 Six-lobed shapes. Courtesy of the *Drop Gallery* of Dr. Claus-Justus Heine and Dr. Gerd Dzuik, Albert-Ludwigs-University of Freiburg, Dept. Applied Mathematics

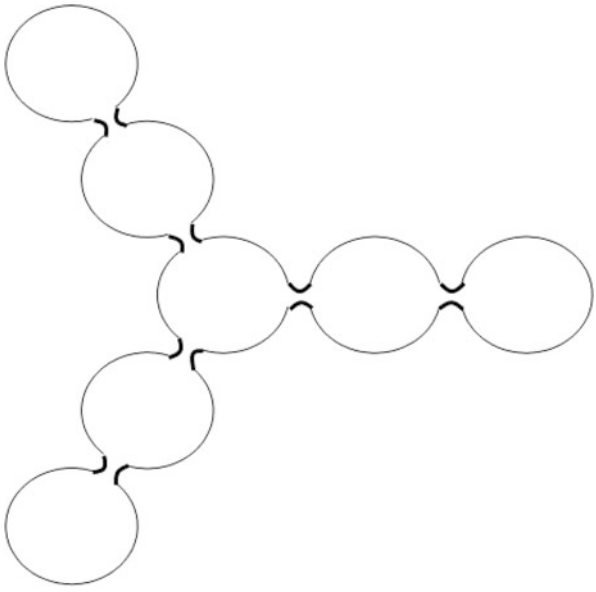


Fig. 9.20 Unstable gyrostatic equilibrium shapes of type $(3, 2)$. See [300]

introduce this new surface equation $X + \epsilon \boldsymbol{n} : \Sigma \to \mathbb{R}^3$ in (9.49) and if we restrict to linear terms in $\mathcal{L}^2/\mathcal{I}^2$ and ϵ, we obtain

$$\triangle_\Sigma \epsilon + |\Pi|^2 \epsilon = \frac{\mathcal{L}^2 \rho}{2\mathcal{I}^2 \gamma} r_\perp^2 + 2H + \frac{P}{\gamma} , \tag{9.50}$$

where $|\Pi|^2 = M^2 + 2M^2 + N^2$ is the square of the length of the second fundamental form

$$\Pi(\boldsymbol{V}, \boldsymbol{W}) = L\|\boldsymbol{V}\|^2 + 2M(\boldsymbol{V}, \boldsymbol{W}) + N\|\boldsymbol{W}\|^2$$

on Σ (see Appendix 1 for more detail). A more complete proof of this linearized formula can be found in [309, Appendix E]. The procedure to build solutions for (9.50) as (m, n) surfaces as in Fig. 9.20 can be briefly explained by the following steps:

1. Denote the expression $\mathcal{L}^2\rho/(2\mathcal{I}^2\gamma) < 1/4$ by a parameter τ and consider it to be small.
2. Build a family of Delaunay surfaces $DS(\tau)$ embedded in $\mathbb{R}^3$, each with mean curvature -1 (equal to a unit sphere S^2), with rotational symmetry, and also satisfying a specified reflection symmetry. These surfaces are basically diffeomorphisms of the unit sphere S^2, and of catenoids (surfaces of revolution of hyperbolic cosine curves), glued together in a differentiable way by circles where the Gauss curvature passes through zero.
3. Show that there are infinitesimal homothety transformations (expansion or dilation with respect to a fixed point p_0, $p \to p_0 + s\boldsymbol{d}_{p_0p}$, $s \in \mathbb{R}$, where $\boldsymbol{d}_{p_0p}$ is the displacement vector from point p_0 to point p) of these surfaces which preserve their geometric properties, viz., the spectrum of the Laplace–Beltrami operator on the surfaces, their symmetries, and the property of having constant mean curvature.
4. The spectrum of the linearized left-hand side of (9.50), as an operator acting on deformation functions, is obtained for some infinitesimal deformations of the Delaunay surfaces. In a next step, it is shown that, in the limit of small eigenvalues, one can choose a space of eigenfunctions orthogonal to the right-hand side of the same equation.
5. We need to define a functional mapping (9.50) from its value initially calculated from the geometry defined in (2) above, to a new value calculated for a hypothetical solution satisfying the Young–Laplace equation.
6. When one performs homothety transformations on the initial shapes as prescribed in (2) and (3) above, the functional defined in (5) must have all the necessary smoothness and isometric properties to be a contraction, so that it has a fixed point on a compact and convex set of hypothetical solutions.
7. From this procedure, we obtain a series of inequalities proving that the functional defined in (5) does indeed have a fixed point which is the solution of the required equation. These inequalities are valid if the control parameter τ is small enough.

In conclusion, the main result can be expressed as follows:

Theorem 13 *For an incompressible liquid drop of density ρ, volume V_0, and surface tension coefficient γ in rigid rotation Ω around a fixed axis, in the absence of any other forces, and for any value of the magnitude of its angular momentum less than a given limit $\mathcal{L} < \bar{\mathcal{L}}(V_0, \rho, \gamma)$, there is a family of shapes of type*

(m, n) *(see Fig. 9.20) and a deformation of these shapes followed by a homothety until the resulting shape contains the volume* V_0 *and satisfies the Young–Laplace equation (9.50).*

This rotating drop is in unstable gyrostatic equilibrium. The stability of gyrostatic equilibrium shapes can be studied using the same Gâteau second differential equations (9.43) and (9.48) for the dimensionless Routhian in (9.40), since the drops in question are free drops:

$$\tilde{L}_0 = \frac{1}{4R^2}\int_\Sigma \mathrm{d}\mathcal{A} + \frac{\mathcal{L}^2}{8\gamma R^2 \mathcal{I}^2} - \tilde{k}\left(\frac{1}{R^3}\int \mathrm{d}V - \tilde{V}_0\right) \tag{9.51}$$

$$= \frac{|\Sigma|}{4R^2} + \frac{\mathcal{L}^2}{8\gamma R^2 \mathcal{I}^2} - \tilde{k}\left(\frac{|\mathcal{D}|}{R^3} - \tilde{V}_0\right). \tag{9.52}$$

The second variation of the Routhian has the form

$$\tilde{L}_0'' = \int_\Sigma \left[\gamma\left(|\nabla\Phi|^2 - \Phi^2|\Pi|^2\right) + \frac{\mathcal{L}^2\rho}{\mathcal{I}^2}\Phi^2(\boldsymbol{r}\cdot\boldsymbol{n})\right]\mathrm{d}A + \frac{\mathcal{L}^2\rho}{\mathcal{I}^3}\left(\int_\Sigma \Phi r_\perp^2\right)^2 > 0\,, \tag{9.53}$$

where $\boldsymbol{r} \in \Sigma$ is the surface parameterization and $\boldsymbol{n}$ is the unit normal on Σ. Rigorously mathematical, the surface Σ is an immersion in a 3D Euclidean space with $\boldsymbol{n}$ its Gauss map. The test functions Φ are defined on Σ and must satisfy the following constraints: they must have zero mean on the surface, have compact support, and be square integrable and continuously differentiable on Σ. In other words, they must belong to the Sobolev space $\Phi \in H_0^1(\Sigma)$. By following a long sequence of inequalities, it can be proved that the second Gâteau variation in (9.53) has negative value, so the solutions are unstable. Moreover, we may mention an interesting kind of behavior here: the more complicated the drop (the larger the number of lobes), the more linearly independent functions violating the stability can be found. This means that the larger the number of lobes, the more unstable the drop is.

It is interesting to estimate the limiting value of the angular momentum below which such (m, n) lobe–neck structures are expected to form. The smallness parameter τ can be extracted from [300] and from the dimensionless formulas (9.35) and (9.49):

$$2H\gamma = P + \frac{\rho\mathcal{L}^2}{2\mathcal{I}^2}r_\perp^2\,,$$

which can be written, after the homothetic expansion, in the form

$$H = 1 + \tau r_\perp^2\,.$$

Table 9.2 Estimates for the maximum allowable angular velocity for several physical systems involving gyrostatic states

Physical system	$\Omega_{\max}$ from theory $[s^{-1}]$	$\Omega_{\exp}$ experimental $[s^{-1}]$	Occurrence of (m, n) shapes
Water drops	10^4	$10 \div 100$	Possible
Mercury drops	10^6	$100 \div 10^3$	Possible
Heavy nuclei	10^{44}	10^{21}	Possible
Neutron stars	10^{-11}	100	Impossible

It follows that $\tau = \rho\mathcal{L}^2/4\gamma\mathcal{I}^2$, but at the same time $P/2\gamma$ must be of the order of unity. The term

$$\frac{P}{2\gamma} \sim \frac{\text{volume energy density}}{2 \times \text{surface area energy}} \sim \frac{|\Sigma|}{V_0} \sim \frac{1}{R} \sim 1 \,,$$

which asks for the distances to be measured roughly in units of the radius of the drop. Because τ has dimensions m^{-3}, its limiting value obtained from the main theorem concerning existence of the solution must be scaled with $\mathcal{O}(R^3)$. In [300], no procedure is given to obtain the minimum value for τ, but in principle one can choose $\tau < 0.1$ from the average limitations of the majority of the lemmata, and from here we have $\Omega < \Omega_{\max} = \sqrt{2\gamma/5\rho}$. In Table 9.2, we present estimates for the maximum allowable angular velocity for several physical systems involving such gyrostatic states.

9.6 Rotation of 2D Drops

In comparison to 3D drops, 2D drops have a stronger tendency to bifurcate their shapes from one symmetry to another, especially if they are in rotation. The stability and symmetry breaking of systems are studied using bifurcation equations that arise from the Hamiltonian formalism. It is well known, and we provide many examples in this book, that the higher the angular velocity, the more unstable the shape becomes. This situation can be proved with great accuracy in terms of mathematical theorems using, for example, the *energy–Casimir method* and singularity theory [167, 294]. The final result of this sophisticated analysis is neither unexpected, nor comprehensive, but it is exact. It applies only to rigidly rotating axisymmetric 2D drops.

For example, an axisymmetric (rather oblate) spheroidal drop of radius R in rigid rotation with angular velocity Ω becomes unstable as soon as $\Omega^2 > 12\gamma/(\rho r^3)$. Hence, for a millimeter-size raindrop, rotational instability breakup would occur around $\Omega \sim 10$–50 Hz, while for a soap bubble it would occur for a slightly higher value, towards the kilohertz range, and for a heavy nucleus liquid drop model,

in the range $\Omega_{\text{nucl}} \sim 2 \times 10^{21}\ \text{s}^{-1}$, which gives a nuclear angular momentum in the range of $\mathcal{L}_{\text{nucl}} \sim 200\hbar$. Actually, this theoretical limit is close to recent experimental measurements on superdeformed/hyperdeformed nuclei. The ^{132}Cs nucleus is known to have a terminating rotational velocity at angular momentum $78\hbar$.

In order to apply the energy–Casimir method, we recall the concepts introduced in Sect. 9.4 on the Hamiltonian description of a free boundary liquid drop in rotation about the Oz axis. Let the Hamiltonian of the drop be described by the Hamiltonian (9.28). The energy–Casimir method seeks to find another conserved quantity C satisfying $\{H, C\} = 0$, where the Poisson brackets are as defined in Sect. 9.4. If C is chosen to be a functional depending on the vorticity $\boldsymbol{\omega} = \nabla \times \boldsymbol{v}$, this guarantees that the Poisson bracket will commute with the Hamiltonian. Let this function be $\Phi(\boldsymbol{\omega})$ which, given the rigid type of rotation and spherical shape (hence major limitations of this model), is just a function of the constant angular velocity Ω, i.e., $\Phi(\Omega)$. In addition, one can always add the angular momentum to C, as this is conserved for such rigidly rotating drops. We now have the *modified* Hamiltonian in the form

$$H_C = H + C = \int_{D_\Sigma} \left[\frac{\rho}{2} |\boldsymbol{v}|^2 - \mu(\boldsymbol{x} \times \boldsymbol{v})_z + \Phi(\boldsymbol{\omega}) \right] \mathrm{d}^3\boldsymbol{x} + \gamma \int_\Sigma \mathrm{d}A \,, \tag{9.54}$$

where μ is an arbitrary parameter for the moment. By computing the first variation for this modified Hamiltonian [see (9.25) and the appendix in Appendix 2], we obtain the equilibrium condition at the critical point of the modified Hamiltonian determined by $\mathrm{d}\Phi(\Omega)/\mathrm{d}\Omega = 0$. Therefore, at the critical points, we can always choose this function to be constant or even zero, i.e., $\Phi = 0$.

In order to determine the stability, we need the positive definition of the second variation, given in (9.25) and Appendix 2. After some vector algebra, neglecting the quadratic terms and retaining only linear terms in $\delta\Sigma$ (another loss of generality in the model), we obtain for the second variation the approximate form

$$\gamma \int_\Sigma \left[-\frac{1}{R^2} (\delta\Sigma)^2 - (\triangle \delta\Sigma)\delta\Sigma \right] \mathrm{d}A > \left(\frac{\Omega}{2\rho} \right)^2 R \int_\Sigma (\delta\Sigma)^2 \mathrm{d}A \,, \tag{9.55}$$

for all normal infinitesimal variations $\delta\Sigma$ of the area, all preserving the volume (the surface mean normal variation is zero). The eigenvalues of the Laplace operator on the circle (2D drop) are given by

$$\lambda_k = \frac{k^2}{R^2} \,, \quad k \in \mathbb{Z} \,,$$

and by choosing the smallest eigenvalue of the first deformed shape $k = 2$ (since $k = 1$ is just a translation of the sphere), we obtain the constraint for stability

$$\Omega < \sqrt{\frac{12\gamma}{\rho R^3}} \,. \tag{9.56}$$

Above this value for the angular velocity bifurcation, a new branch of solutions bifurcates from the axisymmetric solutions. Actually, the latest results show that there is a whole tree of ramifications and branches of star-shaped or bead-shaped 2D drops, as we will show in the following sections.

9.7 Leidenfrost Drops[1]

There are three reasons why the 2D drop-like systems are now so widely studied: they lead to the greatest variety of patterns and waves on the free surface, they are relatively easy to handle theoretically, and experimental procedures are readily available [271–278]. The understanding of the formation, propagation, stability, bifurcation, breakup, and clustering of waves, patterns, and shapes on compact fluid boundaries of isolated drops are of fundamental importance in fluid dynamics problems, printing and painting technologies, and design in the medical and pharmaceutical industries. This field generates an equally diverse array of uses that include paint applications, drug delivery, point-of-care diagnostic chips, neutron star tides, organic synthesis, and droplet-based microfluidics [310]. More generally, the study of the nonlinear and complex dynamics of almost incompressible systems with free surface can help us to understand phenomena in tsunamis and tides, mixing problems, atmospheric dynamics, Bose–Einstein condensates, cosmology, brain structure, motile cells swimming, shear flows, etc.

The first systematic studies on liquid drop oscillations and rotations were reported around the same time as the introduction of differential geometry and topology in hydrodynamics, i.e., in the 1960s and 1970s, in the historical work of Eben, Marsden, Arnold, Kato, Abraham, Smale, Foias, and many others [167–169, 279–281, 285]. Topological fluid dynamics, still a young mathematical discipline even today, is most likely the best theoretical approach for the study of the complex features of flows with complicated trajectories. The theoretical power of differentiable topological methods is consistently supported by more and more experimental results on drop-like systems.

Ideal theoretical models that can be used to test the predictions of the exact Hamiltonian formalism require total isolation of the fluid from any type of external interaction: containers, surrounding walls, or other contact surfaces. Consequently, experiments able to test the Hamiltonian theories need gravity-free labs and

[1] Section written with, and experiments performed by Ajay Raghavendra and Benjamin Dillahunt

completely isolated liquid surfaces. Such experiments can be performed either in parabolic zero-gravity flights and in labs in Earth orbit, or by isolating drops using diamagnetic levitation, ultrasound levitation, air cushion flow from beneath, etc [311].

An alternative experimental method for isolating liquid drops has emerged recently. It consists in using 2D drops wrapped in their own 'hot' vapors. Such drops tend to shrink to disk shapes under their gravity, and the 2D free surface is replaced by the simpler 1D dynamics of the contour. For such systems, the contribution of gravity becomes trivial because of the low dimensionality of the system. The liquid in such a drop is maintained separated from the solid interface for long enough to treat the drop as isolated.

These drops have fully free liquid surface and levitate on their own vapor cushion when brought into the neighborhood of a hot solid, the so-called *Leidenfrost phenomenon*. This is a dynamical and transient effect in which the vapors press against the weight of the drop and the adhesion and friction forces. The absence of solid/liquid contact provides unique mobility for the levitating liquid, contrasting with the usual situations in which contact lines induce viscous forces. Leidenfrost drops of different sources exhibit a frictionless motion with the possibility of bouncing after impact. During the rapid rotation, the resulting instability can lead to polygonal shapes, star-like shapes, and oscillations with various exotic shapes (see Fig. 9.21).

The effect was studied scientifically for the first time in 1756 by the German physician Johann Gottlob Leidenfrost, who published a treatise in which he described the remarkable behavior of liquid drops on a very hot plate, such as water on steel at 300°C [312]. His drops were very mobile, they did not boil, and they lasted a long time despite the very high temperature of the substrate. All this happens because the liquid drop sits on a cushion of its own vapors which prevents contact with the solid surface. For example, a millimetric water drop has a lifetime of several

Fig. 9.21 Leidenfrost star. Water on a hot superhydrophobic surface

minutes on a substrate kept at 250°C, while it evaporates in less than one second when deposited on the same substrate at 150°C (for a very good and recent review, see [313]). The effect has actually been known since earlier times, from Herman Boerhaave's experiments in Leiden in 1732. He reported that alcohol poured on a hot plate does not catch fire, but instead forms "a gleaming drop resembling quicksilver" [313]. Interesting and instructive, the effect is also discussed in literature such as Jules Verne's *Michel Strogoff* (1876) and Marcel Proust's *Swann's Way* (1913).

In spite of its early discovery in the nineteenth century, the *Leidenfrost phenomenon* is still the subject of numerous studies. We may mention at least four important reasons leading to this special interest.

The first is related to the engineering of droplet-based microfluidic systems compatible with many chemical and biological reagents and capable of performing a variety of 'digital fluidic' operations that can be rendered programmable and reconfigurable [264, 314]. Thanks to their dimensional scaling benefits and the possibility of rapid mixing of fluids in a droplet reactor (resulting in decreased reaction times), shape controlled droplets, coupled with the precise generation and repeatability of droplet operations, have made the droplet-based microfluidic system a potent platform for biomedical research and applications. Ranging from the nano- to femtoliter range, droplet-based systems are also used to directly synthesize particles and encapsulate many biological entities for biomedicine and biotechnology applications.

The second reason is the design of self-propelling fluidic devices that have received special attention over the past few years because of their unique ability to displace liquid at small scales without an external force. These devices can be used to chemically treat a solid, to direct and concentrate liquid, for example in condensers, or to drive compounds, as observed with the phalarope, a bird that drives its prey mouthward encapsulated in water.

The third reason is based on the property of a Leidenfrost drop to oscillate spontaneously. For each elementary rebound, part of the kinetic energy can be transferred from the vertical to the horizontal direction because of the asymmetries in the neighboring surfaces.

The fourth reason consists in the ability of Leidenfrost systems to simulate the behavior of super-non-wetting materials. Such non-wetting systems have attracted a lot of attention because they lead to unusual behavior of liquids: a drop impinging on such a solid bounces off, and the film of vapor allows a significant slip of a drop along the solid, which dramatically reduces its friction (see Fig. 9.22, for example).

When a drop of liquid is deposited on a hot solid (in a gravitational field and normal atmosphere), with temperature around the boiling point of the liquid in that atmosphere, the drop boils and quickly vanishes. However, if the solid temperature is much higher than the boiling point, the drop is no longer in contact with the solid, but levitates above its own vapor layer, hence evaporates at a slower rate, remaining at an almost constant and uniform temperature equal to the normal boiling point. This is the essence of the Leidenfrost effect. The system is also similar to the situation of a drop moving over a superhydrophobic layer. This effect does not necessarily require high temperatures; it is all about the temperature difference.

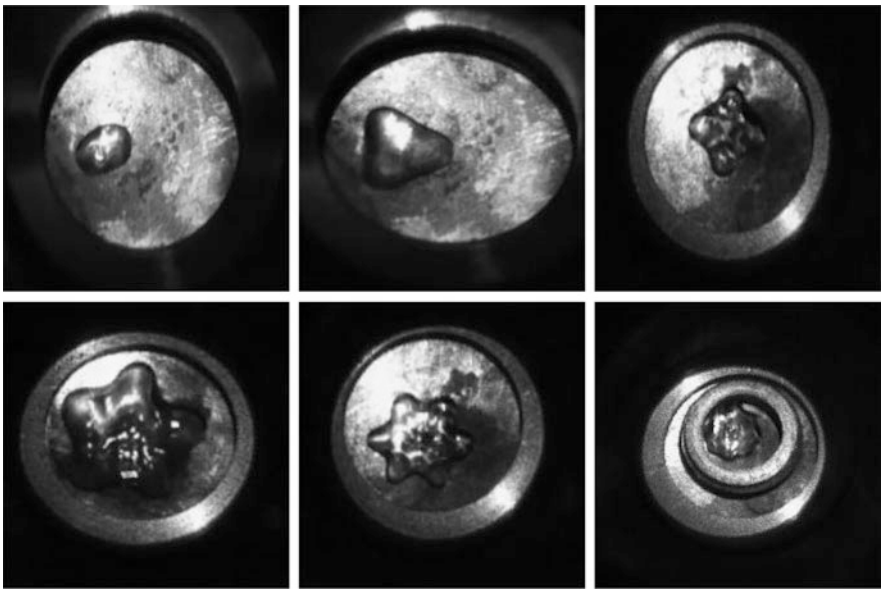


Fig. 9.22 Water Leidenfrost droplets at 280°C, recorded with an AOS Rapid Camera 1000 fps at 1 Mp and a mirror system. *From upper left corner, clockwise*: Dipole, triangle, concave square, concave pentagon, concave hexagon, and concave octopole shapes. Experiments performed in 2014 at the *Wave Motion Lab* at Embry-Riddle Aeronautical University by Ajay Raghavendra

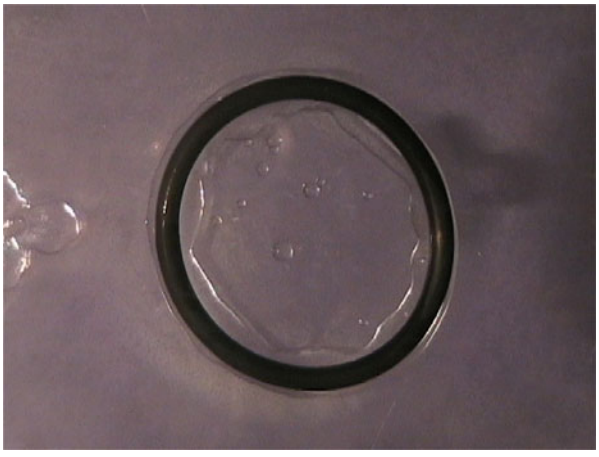


Fig. 9.23 Large pool of liquid nitrogen on a glass plate at room temperature, viewed from above. Several bubbles rise and burst at the upper surface if the liquid surface is large enough. The black toroidal rubber gasket has external diameter 20 mm

Liquid nitrogen generates perfect Leidenfrost drops on a simple glass surface at room temperature (see Figs. 9.22 and 9.23).

Theoretical models involve equations for the balance between surface tension and the Poiseuille flow in the vapor layer. The appearance of star shapes can be explained by the temporal modulation of the eigenfrequency of the drop, due to external forcing, thereby inducing a parametric instability: a first instability leads to a vertical oscillation of the drop, which through a secondary, parametric instability, leads to the formation of oscillating stars. The droplets have a maximum radius beyond which they transform into a torus by the process of a hole nucleation and expansion at their center. Azimuthal oscillating capillary waves on the surface of Leidenfrost drops generate large amplitude star-like undulation. The frequency of oscillations has been shown to be close to the frequency of Rayleigh capillary waves in the droplets. Drops with faceted shapes have also been observed in other drop systems excited by periodic forces with frequency close to the eigenmodes of the drops. Such star shapes arise for drops on vertically vibrated hydrophobic substrates, acoustically levitated drops with low-frequency modulated pressure, liquid metal drops subjected to an oscillating magnetic field, or drops on a pulsating air cushion.

It is shown in [315] that a liquid drop of radius $R < l_c$ placed on a solid surface will be in contact (domain of flattened bottom of the drop) with the solid across a circular domain of diameter

$$\lambda \sim \frac{R^2}{l_c} ,$$

where $l_c = \sqrt{\gamma_{ls}/\rho_l g}$ is the capillary length with γ_{ls} the coefficient of surface tension between the liquid and solid, g the acceleration due to gravity, and ρ_l the density of the liquid. If the drop has radius larger than the capillary length, the liquid forms pools of height about $h \sim 2l_c$ [315]. For example, a Leidenfrost drop of water ($\gamma = 0.059\,\mathrm{N/m}$, $\rho_l = 980\,\mathrm{Kg/m^3}$, $l_c = 2.5\,\mathrm{mm}$) at $t = 150°\mathrm{C}$ (the Leidenfrost temperature for water at which the lifetime is the longest) forms a pool of height $h = 5.0$–$5.1\,\mathrm{mm}$ (see Fig. 9.23).

If the pool is larger, a gas bubble may form at the lower contact surface, close to the center of the pool, and spring out to the surface. It is believed that these bubble bursts are a consequence of a Rayleigh–Taylor instability at the lower surface. Fluctuation and impurities create the germ of a bubble. Then the balance between the surface tension and the buoyancy force generates an unstable equilibrium because the bubble that is trying to exist creates higher local curvature, hence higher pressure. Using a simple energy estimate, it is found that the minimum radius of the pool satisfies a law of the form $\lambda \sim 3.84 l_c \sim 1.92\,h$.

The vapor layer under the drop (if it is large enough to form a pool) lies in the range $e = 10$–$100\,\mu\mathrm{m}$, and interestingly enough, increases with increasing drop size. For an ideal stationary regime, one can estimate the height of the vapor layer under the drop using the Fourier law of heat conduction and the Poiseuille flow law. On the one hand, we can write

$$\Lambda \frac{\mathrm{d}m}{\mathrm{d}t} = \frac{\kappa}{L} \frac{\triangle T}{e} \frac{\pi \lambda^2}{4} ,$$

where Λ is the latent heat of evaporation, $\mathrm{d}m/\mathrm{d}t$ is the rate of evaporation, κ is the thermal conductivity of the vapor, L is the latent heat of evaporation, and $\triangle T$ is the temperature difference between the solid and the boiling temperature. We also have

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \rho_{\mathrm{g}} \frac{2\pi e^3}{3\eta_{\mathrm{g}}} \triangle P \,,$$

where ρ_{g} is the vapor density, η_{g} is the kinematic viscosity of the vapor, and $\triangle P$ is the gas pressure difference between the layer under the drop and the surrounding atmosphere. Combining these last two equations, we obtain an estimate of the drop height:

$$e \sim \left(\frac{3\kappa \triangle T \eta_{\mathrm{g}} R^2}{4L\rho_{\mathrm{v}}\rho_{\mathrm{l}} g l_{\mathrm{c}}} \right)^{1/4} .$$

All the above characteristics can be combined to create devices in which self-propulsion is obtained, using asymmetric textures on hot solid surfaces. Such self-moving non-wetting Leidenfrost drops are very quick, owing to the highly reduced friction. In 2006, Linke et al. [316] discovered such a system for self-propelling drops when asymmetric teeth are present on the supporting solid. The teeth have millimetric lengths and the solid is heated well above the Leidenfrost temperature at which the vapor film builds up, so that the tips of the teeth do not induce boiling. Drops on these hot ratchets accelerate up to a constant velocity. They follow a direction in which they climb steps.

Many effects may be responsible for the propulsion. First, the base of the drop is deformed by the presence of the ratchet below, and this induces a modulation of its curvature and consequent Laplace pressure gradients. Second, a wave propagates from the trailing edge to the leading edge of the drop, allowing transport of matter in the direction of motion. Third, the Marangoni effect, related to temperature differences, might cause a displacement, as seen in Marangoni levitating drops heated asymmetrically using a light source. Fourth, as the drop loses material, this gas flow might cause motion, provided it is made directional (or rectified) by the presence of the teeth. It is crucial to note that all four possibilities above are related to the deformability of the surface of the moving body, whence the importance of the free boundary for this new phenomenon.

Under some special experimental settings, Leidenfrost drops can provide localized solitary wave excitations rotating around the drop boundaries, like for example the "rotons" obtained in liquid nitrogen inside a circular space [121], or the "oscillons" in [317].

9.8 Spinning Polygons

Models for the oscillations, wave patterns, and various motions of the Leidenfrost drops are based on numerical solutions of the Navier–Stokes equations, with corresponding free surface boundary conditions. In general, given the flat aspect ratio of these drops, 2D models are sufficient to describe the main characteristics. Fully 3D models require stronger simplifying hypotheses in order to be able to provide predictions and generate compact solutions. For example, in [318], the authors design a 3D incompressible model based on the Bernoulli potential flow equation. They obtain numerical solutions for a Leidenfrost drop suspended on a cushion of compressed air injected from below.

The reduction of theoretical models to two dimensions is supported by specific experimental configurations, as in the case of [314], where the authors designed a 2D model for water drops sandwiched between two substrates separated by a distance h. Another similar flat configuration is the *lab-on-chip* device, based on digital microfluidics [264], where the authors analyzed 2D patterns and oscillations. In [319], the authors investigate theoretically the behavior of Leidenfrost droplets inserted in a Hele–Shaw cell (between two parallel transparent surfaces whose gap is smaller than the capillary length). In order to have high quality image recordings one needs to keep the droplets as fixed as possible. The drop takes the shape of a flattened saucer-like disc which floats between two vapor layers. These drops are quasi-thermally isolated from the surface by the evaporating vapor layers and they display undulating star-like shapes. For the theoretical model, Pomeau et al. [319] used a 2D 'lubrication approximation' model and numerically obtained shapes for the droplets that matched experiment. They show evidence of capillary azimuthal oscillating modes, and in the hydrodynamic stability limit, they noticed the occurrence of a sudden transition from a flattened disc to an expanding torus.

In the following, we present experiments and a model describing a different type of 2D Leidenfrost system: drops trapped inside a fixed ring and forming a spinning hole at their center (see, for example, Figs. 9.24–9.26, or earlier experiments and discussions in [121]). The ring acts as a rigid, imposed external boundary, while the free liquid surface is inside the drop. The ring, usually a regular toroidal rubber gasket of larger diameter 2 cm and smaller diameter 3 mm, is placed on a horizontal glass plate, and liquid nitrogen is poured inside. The patterns are recorded using a custom-built AOS rapid-photography camera recording 1 000 frames per second from a short distance. When poured, the liquid fills the inside of the ring and some of it spills over the outside, but remains in contact with the outside of the ring thanks to its own capillarity and cohesion forces. Of course, as this liquid is in a Leidenfrost regime, it is not in direct mechanical contact with the ring or the glass, because it is surrounded by a thin vapor layer.

Two independent liquid systems are formed, one outside the ring and one inside. Both these layers oscillate, generating rotational waves and spinning patterns, but at the same time evaporating quickly. Their volumes and sizes continuously decrease and as a consequence all patterns and wave structures have a short lifetime. A given

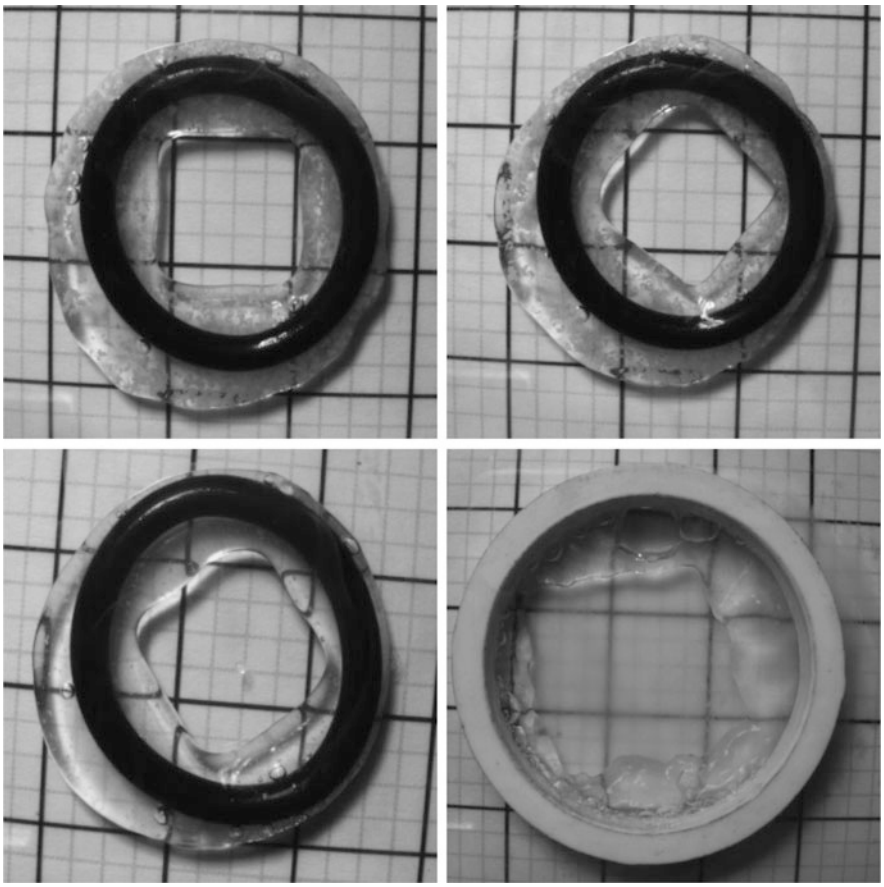


Fig. 9.24 Liquid nitrogen inside a rubber ring placed on a horizontal glass plate. Hollow Leidenfrost drops form and generate spinning squares. We present such squares with rapid photography taken at different moments. All internal diameters are 2 cm and the height of the drops is $h = 3$ mm. A Leidenfrost layer is formed outside of the ring, too, but in these experiments shows chaotic behavior. The squares have edges measuring almost 1 cm, and they always spin counterclockwise with angular velocity $\omega = 220$–360 rad/s

wave or pattern becomes stable at a certain moment, but only for a short while, engaging in regular motion before being replaced by other stable patterns for the new diminished sizes, and so on until full evaporation.

In the beginning, a high frequency wave forms inside the ring, while the flow outside is chaotic and dominated by boiling of undetermined shape (see Figs. 9.23, 9.25 upper right, or 9.26 upper right). In a couple of seconds, the rotating wave inside is replaced by the formation of rigidly rotating polygons: first there are hexagons, then pentagons, then squares, and occasionally triangles (see Figs. 9.24 and 9.25). The transitions between different spinning polygons are

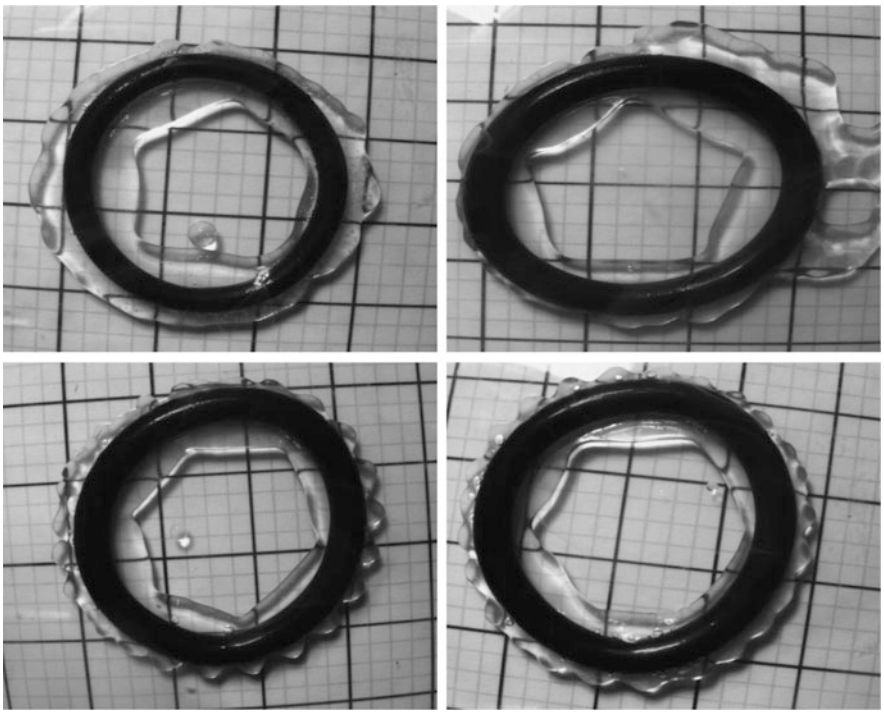


Fig. 9.25 Liquid nitrogen rotating Leidenfrost polygons of order five and six, taken by rapid photography at different moments, inside a toroidal rubber gasket identical with the one in Fig. 9.24. The pentagons/hexagons have edges of length about 1.3 cm, and they spin at $\omega = 320$–400 rad/s. The external liquid displays organized flow, viz., denser rotational waves of small radial amplitude and wavelength, with $\delta r \approx 2$ mm, $\lambda \approx 1.8$–2.5 mm, and $\omega \approx 600$–1 000 rad/s

mediated by short intervals of chaotic flow inside the drop. About the time when the liquid inside the ring ceases to form stable rotational polygons, the outside liquid generally provides a high frequency rotational wave with small amplitude, spinning fast around the exterior of the drop (see Figs. 9.25 bottom row and Fig. 9.26 upper left). After the vanishing of this last pattern, the liquid completely evaporates and water ice from the moisture in the room freezes on the ring and glass. In some configurations the inside polygons are convex, and in some they are concave.

In order to identify the type of motion of the inner patterns, we mixed small fluorescent particles into the liquid nitrogen (e.g., Cosperics commercial powder with beads of diameter between 10 and 200 μm). Rapid photography reveals that both the inside and outside motions only involve geometric waves (energy waves), and no matter waves. For a ring diameter of $R = 2$ cm, the motion of each bead is almost radial, and confined to a range of 3–4 mm about its mean position. Such beads can be observed, for example, in Figs. 9.24 (upper row) or 9.26 (upper right and bottom left). These observations allow one to introduce a suitable theoretical model of shallow-water hydrodynamics with non-slipping boundary conditions,

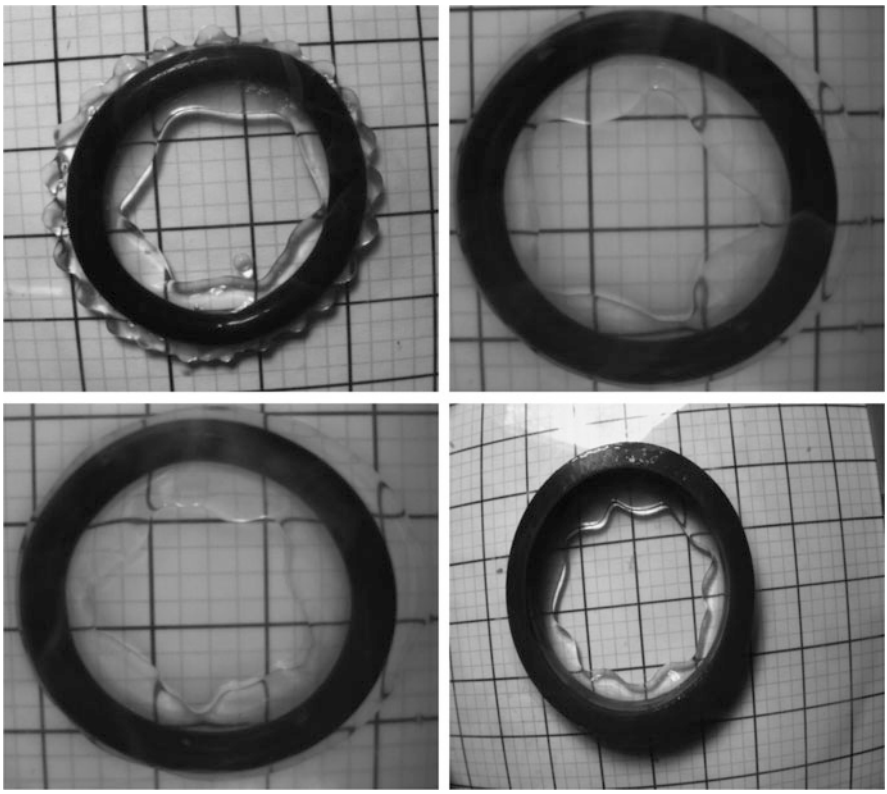


Fig. 9.26 Leidenfrost liquid shells inside a rubber gasket on flat horizontal glass. Concave rotating polygons of order five and six are obtained, together with higher multipolar patterns in the form of stable rotating waves of order up to 9 vertices

rather than a rigid rotator model, as in the case of regular Leidenfrost drops discussed in the literature.

The most stable patterns of hollow rotating polygons obtained experimentally were squares (see Fig. 9.24). Less accurate patterns were obtained for pentagons and hexagons (see Fig. 9.25) and concave polygons of order five and six, or simply stable rotational waves including up to 9 wavelengths along the inner contour (see Fig. 9.26).

Theoretical Model

The theoretical model is 2D, neglects convective vertical motion of the fluid and gravity, and takes into account only rotating waves and spinning patterns with constant angular velocity. This model aims to provide explanations for the formation and stability of such patterns. It assumes all calculations to be performed at a time

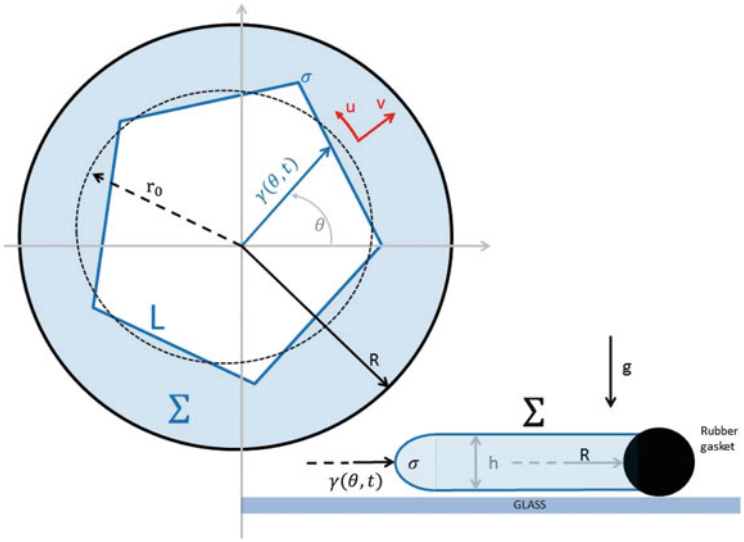


Fig. 9.27 2D model for rotating polygons in Leidenfrost shells. The *left image* shows the drop seen from above, the largest circle R is the gasket, and the *blue line*, of equation $\gamma(\theta, t)$ and perimeter L, is the inner boundary of the Leidenfrost shell. The *dashed circle* at r_0 represents the static equilibrium shape in the absence of motion. The *right image* shows the same shell seen in a vertical cross-section and the shell has an average height h. The *upper* and *lower* horizontal interfaces of the shell are denoted by Σ and the inner vertical surface is denoted by σ. An example of such a σ is presented in Fig. 9.28

t when the liquid shell has mass m and volume V. Because of the fast evaporation, the mass and volume of the liquid shell decrease in time, and the model must use different input parameters to be able to predict the dynamics at later times, for different geometries.

We consider the liquid Leidenfrost shell lying horizontal and flat, and confined by a rigid toroidal gasket of large radius R and small diameter h. The parameter h also represents the height of the liquid shell (see Fig. 9.27). In principle, it is enough to consider h equal to the small diameter of the toroidal ring, but we can in fact evaluate the thickness of the Leidenfrost shell as well as the curvature of the inner contour from optical deformation of a rectangular grid placed under the drops. In the small red circle in Fig. 9.29, we show the optical deformation of the grid lines that allows us to evaluate the true depth and curvature and implement the correct $h = 2.98\,\text{mm}$.

When at rest, the shell has volume $V = h\pi(R^2 - r_0^2)h$ and mass $m = \rho_0 V$. We assume that the model works for a limited amount of time dt centered around t, such that the volume and mass can be considered constant over dt. In polar coordinates,

the inner contour of the shell is given by the function $r = \gamma(\theta, t)$ and instantaneous volume conservation (during dt) requires

$$V = \pi h(R^2 - r_0^2) = h\left[\pi R^2 - \frac{1}{2}\int_0^{2\pi} \gamma^2(\theta, t)\,\mathrm{d}\theta\right] . \tag{9.57}$$

In addition to the area and volume, we also need the length of the inner curve bounding the shell:

$$L = \int_0^{2\pi} \sqrt{\gamma'^2 + \gamma^2}\,\mathrm{d}\theta . \tag{9.58}$$

In cylindrical coordinates, the equation of continuity plus the incompressibility condition for the shell liquid gives the divergence-free condition for the velocity field, i.e., $\nabla \cdot \boldsymbol{V} = 0$:

$$\frac{1}{r}\frac{\partial}{\partial r}(ru_r) + \frac{1}{r}\frac{\partial}{\partial \theta}u_\theta + \frac{\partial}{\partial z}u_z = 0 , \tag{9.59}$$

where the velocity field is given by $\boldsymbol{V}(r, \theta, z, t) = (u_r, u_\theta, u_z)$ in cylindrical coordinates. The Navier–Stokes equations in cylindrical coordinates have the general form

$$\rho_0\left(\frac{\mathrm{D}u_r}{\mathrm{D}t} - \frac{u_\theta^2}{r}\right) = -\frac{\partial P}{\partial r} + f_r + \mu\left(\triangle u_r - \frac{u_r}{r^2} - \frac{2}{r^2}\frac{\partial u_\theta}{\partial \theta}\right) , \tag{9.60}$$

$$\rho\left(\frac{\mathrm{D}u_\theta}{\mathrm{D}t} + \frac{u_\theta u_r}{r}\right) = -\frac{1}{r}\frac{\partial P}{\partial \theta} + f_\theta + \mu\left(\triangle u_\theta - \frac{u_\theta}{r^2} + \frac{2}{r^2}\frac{\partial u_r}{\partial \theta}\right) , \tag{9.61}$$

$$\rho\frac{\mathrm{D}u_z}{\mathrm{D}t} = -\frac{\partial P}{\partial z} + f_z + \mu\triangle u_z , \tag{9.62}$$

where the force density is given by $\boldsymbol{f}(r, \theta, z) = (f_r, f_\theta, f_z)$, the pressure by $P(r, \theta, z)$, $\rho_0 > 0$ is the constant density, and $\mu > 0$ is the dynamic viscosity. Here we have used the Laplace operator in cylindrical coordinates:

$$\triangle = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} + \frac{\partial^2}{\partial z^2} , \tag{9.63}$$

and the total time derivative operator

$$\frac{\mathrm{D}}{\mathrm{D}t} = \frac{\partial}{\partial t} + u_r\frac{\partial}{\partial r} + \frac{u_\theta}{r}\frac{\partial}{\partial \theta} + u_z\frac{\partial}{\partial z} .$$

In order to predict the rotating polygonal holes in the middle of the Leidenfrost shells, it is enough to restrict to 2D motion, as we mentioned above, and we denote the components of the flat Eulerian liquid velocity by $(u_r, u_\theta) = (v, u)$. The degree

of freedom of the Leidenfrost shell associated with the third (vertical) dimension is somehow taken into account on average by the value of h, the shell height. It is assumed that, in the first order of approximation, the shell will take the height enforced by the small diameter h of the rubber ring.

Because we are studying rotating waves, we introduce the coordinates $(r, \theta, t) \rightarrow (r, \xi = \theta - \omega t)$, where the constant angular velocity ω is a parameter of the model, and because we discuss only 'rigidly' rotating patterns inside the shell, volume conservation is automatically assumed. However, when we study the transition from one pattern to another, this process must consider volume conservation.

In this model we assume the flow to be incompressible, irrotational, and inviscid. These simplifications are reasonable for the range of parameters of the liquid used in our experiments (liquid nitrogen), and for the corresponding Reynolds number associated with our experiments ($R \sim 2000$, so the flow is laminar). The equation of continuity and the Navier–Stokes equation along the horizontal components become

$$\frac{\partial (rv)}{\partial r} + \frac{\partial u}{\partial \xi} = 0 \,, \tag{9.64}$$

$$-\omega \frac{\partial v}{\partial \xi} + v \frac{\partial v}{\partial r} + \frac{u}{r} \frac{\partial v}{\partial \xi} - \frac{u^2}{r} = -\frac{1}{\rho_0} \frac{\partial P}{\partial r} \,, \tag{9.65}$$

$$-\omega \frac{\partial u}{\partial \xi} + v \frac{\partial u}{\partial r} + \frac{u}{r} \frac{\partial u}{\partial \xi} + \frac{uv}{r} = -\frac{1}{r\rho_0} \frac{\partial P}{\partial \xi} \,. \tag{9.66}$$

The potential flow enables us to generate the velocity from a velocity potential $\Phi(r, \xi, t)$:

$$v = \frac{\partial \Phi}{\partial r} \,, \qquad u = \frac{1}{r} \frac{\partial \Phi}{\partial \xi} \,, \tag{9.67}$$

while this potential satisfies the Laplace equation

$$\frac{\partial^2 \Phi}{\partial r^2} + \frac{1}{r} \frac{\partial \Phi}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \Phi}{\partial \xi^2} = 0 \,. \tag{9.68}$$

The boundary condition at the external contour provided by the flat rigid rubber ring requires the normal velocity component to cancel, i.e.,

$$v|_{r=R} = 0 \,, \tag{9.69}$$

while the free surface boundary condition, the so-called kinematic free surface condition, at the inner contour is given by

$$S = 0 \,, \qquad \left(\boldsymbol{v} \cdot \nabla S + \frac{\partial S}{\partial t} \right)_{S=0} = 0 \,, \tag{9.70}$$

where $S = 0$ represents the geometric definition of the inner contour (the blue line in Fig. 9.27). In our 2D model, we can write $S = r - \gamma(\xi, t) = 0$ and

consequently (9.70) can be written in the form

$$\left[v + \left(\omega - \frac{u}{r}\right)\frac{\partial \gamma}{\partial \xi}\right]_{r=\gamma(\xi,t)} = 0 \,. \tag{9.71}$$

There is no other boundary constraint because, by definition, Leidenfrost drops are surrounded by a thin layer of vapor and there is no kind of tangential constraint on the velocity, such as no-slip conditions. From the Navier–Stokes equations (9.65) and (9.66) written in the potential and inviscid case, we obtain the Bernoulli equation

$$-\omega\frac{\partial \Phi}{\partial \xi} + \frac{1}{2}(\nabla \Phi)^2 = -\frac{P}{\rho_0} \,. \tag{9.72}$$

Evaluated at the free inner surface and differentiated with respect to ξ, this results in the equation

$$-\omega\frac{\partial(u_\Sigma \gamma)}{\partial \xi} + v_\Sigma\frac{\partial v_\Sigma}{\partial \xi} + \frac{u_\Sigma}{\gamma}\frac{\partial(\gamma u_\Sigma)}{\partial \xi} - \frac{u_\Sigma^2}{\gamma}\frac{\partial \gamma}{\partial \xi} = -\frac{1}{\rho_0}\frac{\partial P_\Sigma}{\partial \xi} \,, \tag{9.73}$$

where the subscript Σ indicates that the quantity carrying it is evaluated at the inner contour. The surface tension at any point of the free liquid surface is given by the well known formula [121]

$$P_\Sigma - P_0 = -2\sigma H \,, \tag{9.74}$$

where σ denotes here the surface tension coefficient and H the mean curvature of the surface (see the definitions in Chap. 4). The two parallel horizontal surfaces have zero mean curvature and the surface around the rubber gasket has constant shape, so these surface elements do not contribute directly to the field of velocity dynamics. However, later on in the development of the model, we consider the potential energy associated with these surfaces, in a process of minimization of the total energy, because the horizontal parts of the surfaces may change through the change in height of the drop and volume conservation.

In order to evaluate the pressure term along the inner contour, we assume that the shape of this surface is a half-tube of small radius $h/2$ following the path of the contour curve $\gamma(\theta, t)$ (see Fig. 9.28). This surface is parameterized by $\boldsymbol{r}(\theta, \phi)$, where $\phi \in [0, \pi]$ represents the variation along the vertical direction and describes the small half-circle. We have [137–139]

$$\begin{aligned}\boldsymbol{r}(\theta,\phi) = \Bigg\{ &\left(1 + \frac{h}{2}\right)\cos\theta\gamma(\theta) - \frac{h}{2}\cos\theta\sin\phi \,, \\ &\left(1 + \frac{h}{2}\right)\sin\theta\gamma(\theta) - \frac{h}{2}\sin\theta\sin\phi \,, \;\; \frac{h}{2} + \frac{h}{2}\cos\phi \Bigg\} \,.\end{aligned} \tag{9.75}$$

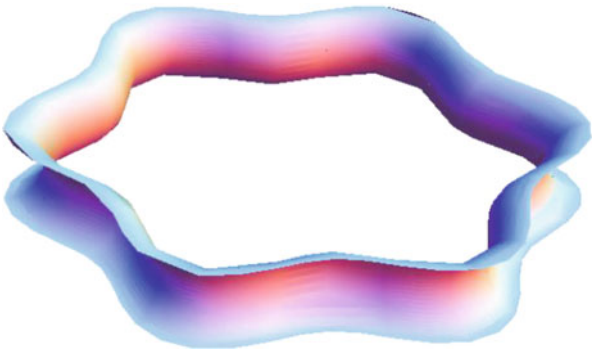


Fig. 9.28 Example of inner surface of a 2D Leidenfrost liquid shell model (see the notation σ in Fig. 9.27). In this example, the inner contour is given by $\gamma = r_0\big[1 + \epsilon \sin(6\theta)\big]$, with $\epsilon = 0.01$. In general, the equation for this surface is given by (9.75), while its mean curvature is given by (9.76)

We need the expression for the mean curvature H from (9.32), calculated from the first and the second fundamental forms of this surface. Performing the calculations and introducing (9.75) into the expressions for the two fundamental forms, the normal and the mean curvature, we obtain the final form for the mean curvature of the inner surface, depending on the contour γ and the height h as follows:

$$
\begin{aligned}
H_{\gamma,h} = -\sqrt{2}\Big\{ & -2(2+h)^3\gamma^3 + 8h(2+h)^2\gamma^2 \sin\phi \\
& +2h\sin\phi\Big[-(2+h)^2\big[-2+\cos(2\phi)\big]\gamma'^2 + h\sin\phi\big[2h\sin\phi + (2+h)\gamma''\big]\Big] \\
& +(2+h)\gamma\Big[-2(2+h)^2\gamma'^2 - 2h\sin\phi\big[5h\sin\phi + (2+h)\gamma''\big]\Big] \\
& \times\Big[-h^2(-2(2+h)^2\gamma^2 + 4h(2+h)\gamma\sin\phi - 2\sin^2\phi(h^2 + (2+h)^2\gamma'^2))\Big]^{-1/2} \\
& \times\Big[2(2+h)^2\gamma^2 - 4h(2+h)\gamma\sin\phi + 2\sin^2\phi\big[h^2 + (2+h)^2\gamma'^2\big]\Big]^{-1}\Big\} \,, \qquad (9.76)
\end{aligned}
$$

where $\gamma' = \mathrm{d}\gamma/\mathrm{d}\xi$, etc. Equation (9.76) describes the mean curvature depending on the vertical coordinate ϕ, which is not actually part of the 2D model. In order to eliminate this dependence, we use the average value of H over the height of the drop, viz.,

$$
H(\theta,\phi) \longrightarrow H(\theta) = \frac{1}{\pi}\int_0^\pi H(\theta,\phi)\,\mathrm{d}\phi \,. \qquad (9.77)
$$

By implementing the mean curvature expression of (9.77) into the surface tension as given by (9.74), and this one into the Bernoulli equation (9.73), we have a system of two partial differential equations (PDE) provided by (9.71) and (9.73). Together with the free surface boundary condition of (9.71), we have altogether three PDEs in the two components of the velocity and the shape of the contour, i.e., three unknown functions.

The typical approach at this stage is to use the Laplace equation, solve it in terms of a series, and implement this result in a perturbative way in the PDEs. Because the shape of the inner contour cannot be too different from the equilibrium value at rest, i.e., $\gamma(\theta, t) \simeq r_0$ (see Fig. 9.27), we can expand the velocity potential in a power series

$$\Phi(r, \theta, t) = \sum_{l \geq 0} (r - r_0)^l f_l(\theta, t) \ , \tag{9.78}$$

or simply

$$\Phi(r, \xi) = \sum_{l \geq 0} (r - r_0)^l f_l(\xi) \ .$$

By introducing this formal series into the Laplace equation (9.68), rewriting the terms r^2 as $[(r - r_0) + r_0]^2$, etc., and identifying the same powers of the radial terms, we obtain the following recursion relations:

$$f_{l+2} = \frac{-r_0(l+1)(2l+1)f_{l+1} - l^2 f_l - f_l''}{r_0^2(l+2)(l+1)} \ , \tag{9.79}$$

for $l \geq 0$. This relation can generate all the coefficient functions f_l in the expression for the velocity potential if we know the expressions for the first two of them, f_0 and f_1.

The boundary condition (9.69) at the rigid ring becomes an infinite series equation for the coefficient functions $f_l(\xi)$:

$$\sum_{l \geq 0} l(R - r_0)^{l-1} f_l = 0 \ . \tag{9.80}$$

This equation cannot be solved simultaneously with (9.79) for a compact solution, so at this point we need to truncate the series in (9.80) and use an approximation up to order f_3 for (9.80):

$$f_1 + 2(R - r_0)f_2 + 3(R - r_0)^2 f_3 \approx 0 \ . \tag{9.81}$$

By combining the last Eq. (9.81) with the recursion equation (9.79), we can determine f_0 as a function of f_1 by integrating the resulting ordinary differential equation:

$$-(R - r_0)^2 f_1'' + (5R^2 + 11r_0^2 - 14Rr_0)f_1 \approx -\frac{6R^2 + 10r_0^2 - 16Rr_0}{r_0} f_0'' \ , \tag{9.82}$$

with the solution

$$f_0(\xi) \approx \frac{r_0(R-r_0)^2}{6R^2+10r_0^2-16Rr_0} f_1(\xi) - r_0 \frac{5R^2+11r_0^2-14Rr_0}{6R^2+10r_0^2-16Rr_0} \int^{\xi} \mathrm{d}x \int^{x} f_1(y)\mathrm{d}y \,, \tag{9.83}$$

and from (9.83) and (9.79), we have all the coefficient functions f_l depending only on the first one f_1.

Equation (9.82) has a special importance. We notice that there is a threshold value for the inner liquid volume given by $r_0 = 3R/5$, when the right-hand side of the differential equation cancels. In this situation, the solutions for $f_1(\xi)$ are real exponential functions of θ and t. In this case, all the coefficient functions in the series describing the velocity potential, and also the velocities, either blow up to infinity, which is a non-realistic solution, or damp to zero, which means that all inner waves, modes, and patterns with

$$r_0 \leq \frac{3}{5}R$$

are unstable (see Fig. 9.27). When the inside of the ring is filled with liquid nitrogen and the liquid evaporates, the flow is chaotic until the inner layer is shallow enough, by the above limit, to allow formation of stable waves and patterns.

At this point in the calculations, the model depends on two independent and arbitrary functions $\gamma(\xi), f_1(\xi)$ and on four parameters ω, h, R, r_0, which are all constrained by the inner contour free surface boundary condition (9.71), and the modified Bernoulli equation and its surface tension expression in (9.73). The problem now consists in a well-posed system of two ordinary differential equations in the periodic variable ξ and must have a unique solution for each set of parameters. We write these two ODEs at the inner contour $r = \gamma(\xi)$ and obtain from the boundary condition (9.71)

$$V + \left(\omega - \frac{U}{\gamma}\right)\gamma' = 0 \,, \tag{9.84}$$

and from (9.73)

$$-\omega(U\gamma)' + VV' + \frac{U}{\gamma}(U\gamma)' - \frac{U^2}{\gamma}\gamma' = -\frac{1}{\rho_0}P' \,, \tag{9.85}$$

where we have denoted the two velocity components evaluated at the inner contour by

$$U(\xi) = u_\Sigma = u(\gamma(\xi), \xi) = \sum_{l\geq 0} \frac{(\gamma - r_0)^l}{\gamma} f_l'$$

and

$$V(\xi) = v_\Sigma = v(\gamma(\xi), \xi) = \sum_{l \geq 1} l(\gamma - r_0)^{l-1} f_l \, .$$

To second order in $(r - r_0)/r_0$, these velocity components can be approximated by

$$V = f_1 + 2(\gamma - r_0) f_2 + \mathcal{O}_2$$

and

$$U = \frac{1}{\gamma} f_0' + \frac{\gamma - r_0}{\gamma} f_1' + \frac{(\gamma - r_0)^2}{\gamma} f_2' + \mathcal{O}_2 \, .$$

We substitute in all the coefficient functions f_l in terms of f_1 according to (9.83). We introduce the notation

$$F(\xi) = \int^{\xi} \mathrm{d}y \int^{y} f_1(x) \mathrm{d}x \, ,$$

and by using once again the recursion relation (9.83) and (9.79), we can write to second order

$$\begin{aligned} V = & \frac{(\gamma - r_0)(5R^2 + 11r_0^2 - 14Rr_0)}{r_0(6R^2 + 10r_0^2 - 16Rr_0)} F \\ & + \left[1 - \frac{\gamma - r_0}{2r_0} - \frac{(\gamma - r_0)(R - r_0)^2}{2r_0(6R^2 + 10r_0^2 - 16Rr_0)} \right] F' + \mathcal{O}_2 \, , \end{aligned}$$

and

$$\begin{aligned} U = & \frac{5R^2 + 11r_0^2 - 14Rr_0}{6R^2 + 10r_0^2 - 16Rr_0} \left[\frac{(\gamma - r_0)^2}{2\gamma r_0} - \frac{r_0}{\gamma} \right] F' \\ & + \left\{ \frac{r_0(R - r_0)^2}{\gamma(6R^2 + 10r_0^2 - 16Rr_0)} + \frac{\gamma - r_0}{\gamma} \right. \\ & \left. - \frac{(\gamma - r_0^2)^2}{2\gamma r_0} \left[1 + \frac{(R - r_0)^2}{6R^2 + 10r_0^2 - 16Rr_0} \right] \right\} F''' + \mathcal{O}_2 \, . \end{aligned}$$

Finally, we implement the above expressions for the velocities in the inner contour boundary condition (9.84) to obtain the differential equation for F, which reads to second order

$$F \approx \omega \gamma' \frac{6R^2 + 10r_0^2 - 16Rr_0}{5R^2 + 11r_0^2 - 14Rr_0} + \mathcal{O}_2 \, . \tag{9.86}$$

This is a typical dependence of the term of order zero in the velocity potential on the expression for the free surface in nonlinear dynamical systems governed by a modified Korteweg-de Vries (mKdV) type of equation. In the following, we use this expression for F from (9.86) and implement it in the modified Bernoulli equation (9.85). In the last step, we succeed in obtaining one nonlinear ordinary differential equation for the contour γ. The expression for this ordinary differential equation is long and tedious and we will not present it in full here. We mention, however, that, in the same second order of approximation, the equation has the generic form

$$\gamma'' \approx C_0 + \gamma C_1 + \gamma^3 C_3 + \mathcal{O}_2 \, .$$

This is similar to the equation for the Jacobi elliptic function $\mathrm{dn}(\xi, \kappa)$, where the parameter κ is its elliptic modulus $\kappa \in [0, 1]$ and depends on the constants $C_{0,1,3}$, which themselves depend on R, r_0, and h. This cnoidal wave solution is doubly periodic with periods $K(\kappa)$ and $K(\sqrt{1-\kappa^2})$, where K is the complete elliptic integral of the first kind. In the limit $\kappa = 0$, the solution $\gamma \sim \mathrm{dn}\,(\xi, \kappa)$ approaches a constant, while in the limit $\kappa = 1$, $\mathrm{dn}(\xi, 1) = \mathrm{sech}(\xi)$, which generates a sech type of soliton as solution. In Fig. 9.29, we present a comparison between the theoretical model and one of the experiments. It is interesting that, in addition to the excellent match, we can model the shape of the inner contour with a four-fold trigonometric function, and the match with experiment is still good. What makes the difference

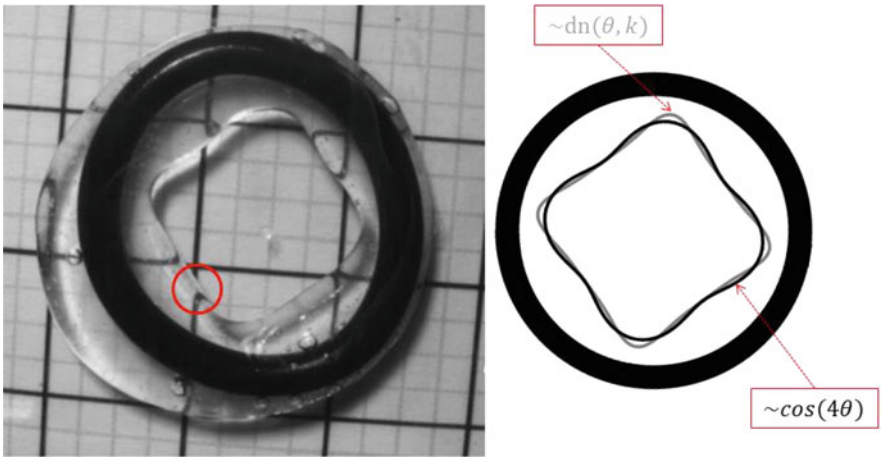


Fig. 9.29 Comparing experiment and theory. The *black contour* in the *right-hand frame* is calculated with linear waves of order 4, while the *gray contour* is a nonlinear cnoidal wave. It is difficult to choose one model in favor of the other from this experiment. The *small red circle* in the *left-hand frame* shows how we evaluated the thickness of the Leidenfrost drop and its curvature towards the inner contour

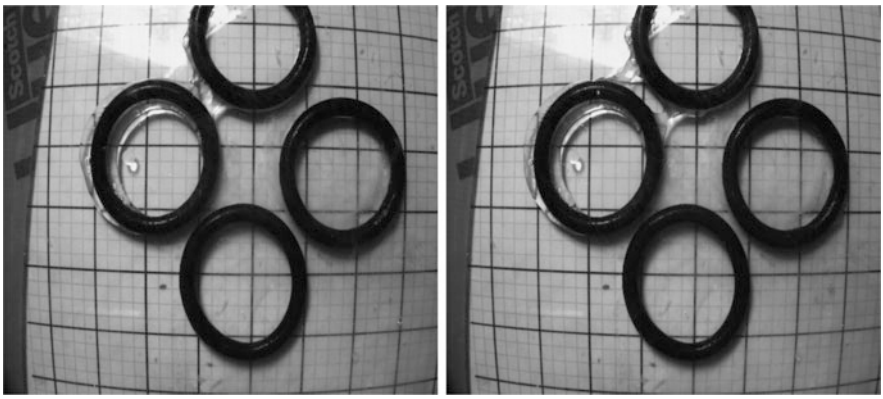


Fig. 9.30 Liquid nitrogen multiple neck formation and oscillations in the convex (outer) zones of the rubber gaskets. The necks bifurcate in double necks, and oscillate for several tens of seconds

and decides which theoretical model is most valid will be the study of the stability of these cnoidal rotating wave solutions.

For different values of the parameters, the cnoidal function can match all the polygons observed experimentally, namely the triangles, pentagons, or hexagons. The higher frequency and angular velocity rotational waves observed experimentally on the outside of the ring can be modeled in a similar way, and the results will be presented elsewhere or in a further edition of this book.

We should mention that, while observing the behavior of the outer liquid, we noticed an interesting type of low frequency oscillation when several rings almost touch, and the liquid tends to join adjacent rings and form 'necks' between the outside contours of the rings. For example, we present the image of such a neck between the two upper left rings in Fig. 9.30: in the left frame, the neck has a narrow configuration and is simply connected, while in the right frame, after several seconds, the neck breaks into two sub-necks and forms a hole in the middle. Later on, the two sub-necks join together and the oscillations repeat for several tens of seconds.

Stability Analysis

We cannot infer from this model what exactly causes the rotation of the inner waves or patterns, or why the angular velocity has a certain value, but we can build a stability diagram for the Hamiltonian associated with this model, and study the continuity and branching of different patterns in this parameter space. We will use the stability analysis based on the Lagrangian of the system, viz., (9.38), which was discussed in Sect. 9.5 for 3D drops.

For a given setting of the parameters R, h, r_0, we choose a polygonal pattern class $\gamma_n(\xi)$ of order n given by

$$\gamma_n(\xi) = r_1 \sum_{k=0}^{n-1} \frac{\cos\dfrac{\pi}{n}}{\cos\left(\dfrac{\pi}{n} - \xi - \dfrac{2k\pi}{n}\right)} \chi_{\left(2k\pi/n,\, 2(k+1)\pi/n\right)}(\xi) \,,$$

where χ is the characteristic function of the interval, i.e., $\chi_{[0,1]}(\xi) = 1$ if $0 \le \xi \le 1$ and zero otherwise. This function describes in polar coordinates a regular n-polygon inscribed in a circle of radius r_1. In addition to this exact polygonal function, we also use as test solutions the cnoidal waves that fit this model polygon, $\gamma = a_1 + a_2 \mathrm{dn}(a_3\xi, \kappa)$, and the n-multiple angle trigonometric functions $\gamma = b_1 + b_2 \cos(n\xi)$ that fit the polygons. For example, in the case $n = 4$, we use $\gamma = 0.67 + 0.29\mathrm{dn}(4.7\xi, 0.99)$, and $\gamma = 0.8 + 0.12\cos(4\xi)$, which are presented in Fig. 9.29. With these functions for the inner contour, we calculate the potential energy of the rotating wave which is actually the total area of the liquid in contact with vapor times the coefficient of surface tension:

$$E_\mathrm{p} = 2\sigma\left[\pi R^2 - \frac{1}{2}\int_0^{2\pi} \gamma^2(\xi)\mathrm{d}\xi\right] + \sigma\pi h \int_0^{2\pi} \sqrt{\gamma^2 + \gamma'^2}\mathrm{d}\xi \,.$$

Here, σ is the surface tension coefficient between nitrogen liquid and vapor. The total Lagrangian functional of (9.38) is the total energy of the liquid, namely,

$$L = E_\mathrm{p}[n, r_1, r_0, h, R] + \frac{\rho_0}{2}\int_0^{2\pi}\int_{\gamma(\xi)}^{R} \left[u^2(r,\xi) + v^2(r,\xi)\right] r\,\mathrm{d}r\,\mathrm{d}\xi \,. \tag{9.87}$$

The velocity components in the kinetic energy term above are obtained from (9.67) with the help of all equations from (9.78) to (9.83), by implementing all the constraints and approximations used. For a fixed experimental configuration characterized by the parameters ρ_0, σ, R, h, and for a given initial volume of the Leidenfrost shell given by the parameter r_0 (and also h), we calculate the Lagrange functional L from (9.87) in terms of the free parameters (n, ω), and apply the stability criteria discussed in Sect. 9.6.

We conclude this section with the observation that the surface flow plays an important role for the polygon states, as can be seen in the top row of Fig. 9.25. Indeed, if fluorescent nanospherical seeding particles are progressively added to the surface of a dry hollow polygon, the corners first straighten out, and the system is finally forced back to the circular state. The tendency of the particles to gather near the center gives them a significant influence on the surface flow near the contact line and through the thin films in the corners of the polygon. By blocking this flow, the polygon is destroyed. A similar result was obtained for rotated liquids in a cylindrical container in [320, 321].

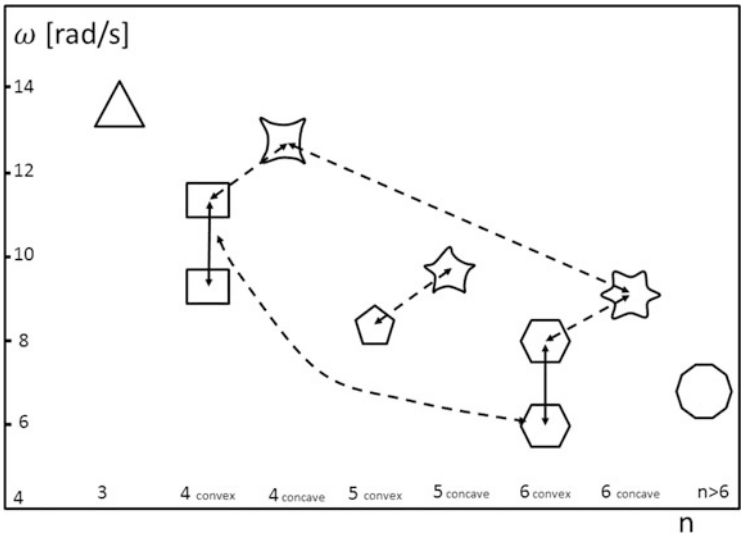


Fig. 9.31 Results from the stability calculations

The results from the stability calculations are summarized in Fig. 9.31. In the plane (n, ω), that is, the number of edges of the polygonal rotating pattern versus angular velocity, we sketch the corresponding polygon for each stability island. The solid arrows show the existence of a continuous domain of stability for the corresponding shape. For example, we determine that squares are stable for $\omega_4 \approx 8.76$–11.33 rad/s and hexagons in the range $\omega_6 \approx 5.70$–8.27 rad/s. The islands of stability for triangles, pentagons, and polygons with more than 6 edges are reduced compared to squares and hexagons, which dominate the stability domain. The dashed arrows represent metastable states where the Leidenfrost shell does not present a stable rotating pattern, but the inner contour oscillates continuously between two such limiting patterns. We have such connections between squares and concave squares, between hexagons and concave hexagons, and directly between squares and hexagons. The concave/convex pentagons are also related by such oscillations, but the Lagrangian functional shows high barrier values between these configurations. From this model, it follows that there is unlikely to be a transition from $n = 3$, to $n = 4$, to $n = 5$, and this result is confirmed experimentally by hundreds of hours of rapid photography recordings.

9.9 Universality in Rotating Fluid Patterns

> *If knowledge be compared to a fruit and the realization of that knowledge to the consumption of the fruit, then a universal statement is to be compared to a hard shell filled with fruit. It is, obviously, of some value, however, not as a shell by itself, but only for its content of fruit. It is of no use to me as long as I do not open it and actually take out a fruit and eat it.*
>
> Hermann Weil,
> *On the New Foundational Crisis in Mathematics* (1998)

The free surface patterns generated in rotating flows contained in compact confinements, either spontaneous as in the previous section, or forced, are not only of fundamental and wide practical interest. Indeed, they also exhibit coherent structures that resemble ones observed in nature or technology [322]. Similar polygonal patterns of rotation occur at very different physical scales. At microscopic scales (10^{-9} m), such structures have been identified in hadron gases and strongly interacting matter, and in Bose–Einstein condensates, while polygonal soliton clusters are generated by necklace-ring beams in dissipative systems [323–325]. At the lab scale, polygons have been observed in liquid nitrogen Leidenfrost drops (10^{-3} m), and water spinning in cylindrical tank experiments (10^{-1} m). At the macroscopic scale, polygonal patterns have been recorded inside hurricane eyes (10^{4} m) and in the six-sided jet stream at Saturn's north pole (10^{8} m).

The common behavior of these processes proves their independence of specific dimensions, it allows them to be measured from a limited set of experimental setups, and then used to predict other processes, with the aid of factorization and perturbative calculations. In a word, we have here a signature of universality. The existence of surface polygons seems to be connected with the fact that the flow is turbulent. In fact, switching transitions are observed between different numbers of edges [320, 321] in similar but smaller systems where the flow irregularly switches between a weakly deformed, rotationally symmetric state and a strongly deformed state with two corners (see also the figures in Sect. 9.8). Transitions between different patterns must be linked with a transition to turbulence, as predicted in [121, Sect. 12.6].

9.9.1 Hollow Polygons on a Rotating Fluid Surface

One traditional way to study such structures is the swirling flow generated by the rotation of one of the end walls in a cylindrical container [320, 326] (see Fig. 9.32). The free surface of the fluid in a circular container with a rotating bottom plate can undergo a surprising instability through which the surface shape spontaneously breaks the rotational symmetry and turns into a rotating polygon (see Fig. 9.33). These shapes were first noticed by Vatistas in 1990 [327] and the polygon rotation

Fig. 9.32 Experiments with a rotated water column contained in a cylinder, viewed from the *top*. *Left*: Squares with Görtler vortices spiraling at the corners. *Right*: Pentagons. Regular polygons with 3–6 edges form spontaneously because of the instability of the inner fluid surface. The number of edges depends on the height, radius, and angular frequency: the shallower the cylinder and the faster it is rotated, the more edges are generated. Courtesy of Phys. Rev. Lett. **96** (2006)

was subsequently analyzed in terms of waves rotating around a vortex core in [328, 329]. The surface polygons are nearly invariant in a frame rotating at a considerably slower rate than the bottom plate and also more slowly than the mean azimuthal flow of the water around the polygon. The instability in such a flow leads to spontaneous symmetry breaking where the vortices formed due to the strong shear flow play the most important role. It is well known that steady patterns of vortices can form in two-dimensional and circular shear flows due to Kelvin–Helmholtz instability [328–331]. The instability in the present system is less well understood since the shear flow in this case is fully three-dimensional and the strength and width of the shear zone are not easily determined as a function of the control parameters.

In a crude model, Mougel and collaborators [332] performed an analysis of the instabilities in rotating free surface flows for an inviscid and incompressible case, neglecting the surface tension, for linearized Euler equations and a linearized free surface boundary condition. In this model, where a cylinder of radius R and height H rotates with constant angular velocity Ω, two types of flow are identified: a solid body rotation (as in Newton's bucket) and a potential flow. The instability of rotational patterns can be parameterized by the Froude number $\mathrm{Fr} = \Omega\sqrt{R/g}$ (see Table 9.1). The equilibrium solutions are found numerically for both cases using a finite-element method. The solutions are then perturbed with a small amplitude term in modal form:

$$\epsilon e^{\mathrm{i}(m\theta - t\omega)} ,$$

where m is the azimuthal wave number of the rotating pattern and ω is the complex frequency.

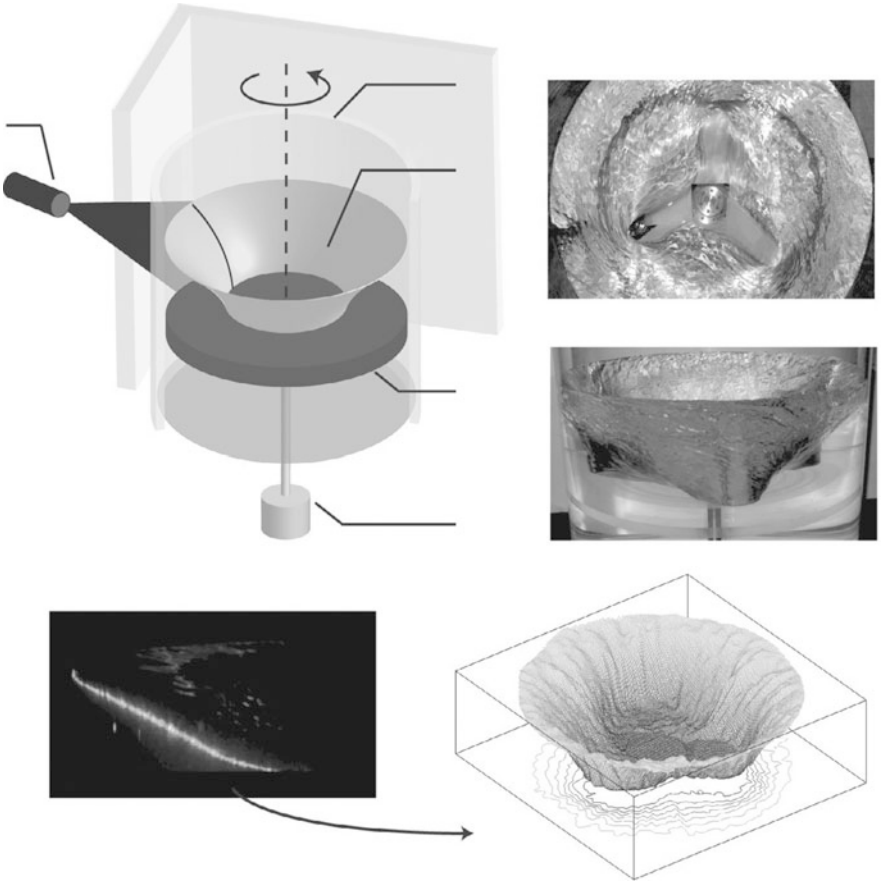


Fig. 9.33 *Clockwise from the top left corner*: Rotating liquid experimental setup for a cylinder of radius 19.4 cm and the PIV system; example of polygon (*triangle*) formation; measurement of the profile of the free surface; reconstruction of the 3D free liquid surface shape. PIV stands for particle image velocimetry, an optical method for reconstructing the velocity field of a microscopically seeded flow, by taking high frequency sequences of laser pictures and analyzing the displacements of the reflections from the seed particles. Courtesy of J. Fluid Mechanics **679**, Cambridge University Press (2011). See the article in [321]

The first configuration, the solid motion of the fluid, occurs if $\mathrm{Fr} < 2\sqrt{H/R}$ and as a qualitative characterization, the bottom of the cylinder remains wet. By implementing the ϵ-exponential perturbation in the linearized Euler equations and in the equation of continuity, the equation for the perturbation ϵ is reduced to a Poincaré equation for the pressure p, in cylindrical coordinates (r, θ, z) :

$$\frac{\partial^2 p}{\partial r^2} + \frac{1}{r}\frac{\partial p}{\partial r} - \frac{m^2 p}{r^2} + \frac{(\omega - m\Omega)^2 - 4\Omega^2}{(\omega - m\Omega)^2}\frac{\partial p}{\partial z} = 0\,. \tag{9.88}$$

Even in the solid motion, we distinguish three possible situations. For $|\omega - m\Omega| > 2\Omega$, (9.88) is elliptic and regular solutions exist. The maximum pressure levels are located along the free surface. These solutions are *gravity waves*, and in the limit of vanishing rotation (Fr $= 0$), their frequencies match with those of the pure sloshing modes in a cylindrical container with flat surface [332]. The stability condition $|\omega - m\Omega| > 2\Omega$ requires the body of liquid to rotate either faster or more slowly than the container.

The most interesting case occurs if $0 \ll |\omega - m\Omega| < 2\Omega$, and consequently (9.88) becomes hyperbolic. In this situation, regular solutions are not expected to correspond to singular inertial modes. Such modes are found to exhibit a striking ray-like structure, similar to the patterns observed in other configurations such as spherical shells, or rotating stars [321, 329–332] (see Fig. 9.33). In a third possible situation of solid motion, when $0 \leq |\omega - m\Omega|$, the perturbations are almost stationary with respect to the rotating frame Ω. In this case, (9.88) becomes degenerate and the Euler equations reduce to the geostrophic equilibrium. In such a quasi-geostrophic context, variations in the height of liquid are known to allow for the possibility of a type of slow wave called Rossby waves.

The second configuration predicted by this model is potential flow, occurring for higher rotation flows, where the fluid leaves dry a central domain at the bottom of the cylinder of radius $\xi < R$. In this configuration, two different kinds of surface waves occur: gravity waves, where the restoring force responsible for them is gravity acting on the external part of the free surface, and centrifugal waves, for which the restoring force is the centrifugal effect acting on the inner part of the free surface where it is nearly vertical.

Experimental results [320] indicate that, while the motion is close to a solid-body rotation in the central region of the container, the outer zone experiences a potential flow, a situation known as a Rankine vortex. The symmetry-breaking transitions observed when switching between different numbers of edges have the characteristics of a low-dimensional linear instability. This situation occurs when a new unstable manifold breaks out from an almost stable rotationally symmetric state. While triangles clearly show dominant vortex structures, for higher order polygons, a point vortex model alone is insufficient to account for the flow or the rotation rate of the polygon.

In conclusion, this type of linear model predicts wave patterns in the inner region where the fluid revolves like a solid body, surrounded by an outer annular sector in which the rotation rate is much less important. The waves can be interpreted in terms of Kelvin normal modes. Kelvin modes are, in general, any normal modes associated with the rotation of the fluid in a stable vortex. They often describe possible small deformations of the vortex. The first configuration, the solid uniform rotation of the fluid, emerging from this linearized model is the simplest Kelvin mode. The waves obtained from (9.88) are hence called Kelvin waves. The Kelvin modes form a basis, so all the deformations of the free surface can be expressed in terms of the Kelvin modes. For the second type of flow configuration of this linearized model, the potential flow or Rankine vortex, the Kelvin modes satisfy similar properties.

In contrast to the linear model presented above, Amaouche and collaborators [329] explain the formation of the dry central shape (hollow core) and corresponding polygonal patterns through a nonlinear formalism, obtaining cnoidal waves as solutions of a nonlinear equation of the Korteweg–de Vries type rather than linear Kelvin modes. Such nonlinear oscillations are observed at the free surface of the fluid in the case of shallow water, at a sufficiently high driven rotational frequency (i.e., a fast spinning disk at the bottom), for large amplitude distortions of the free surface.

In this nonlinear model, the Euler equations are used to describe the inviscid fluid, too, and this option is justified by the high value of the Reynolds number in these experiments ($\sim 10^4$). The model predicts the excitation of inertial waves at the free surface (at the hollow core) as nonlinear disturbances superimposed on a solid body rotation induced by the uniform rotation of the bottom plate. A circular dry region of radius r_0 is considered at the center of the bottom, and the free surface at the bottom has the equation

$$r_{\text{dry}}(\theta, t) = r_0 + \epsilon\xi(\theta, t) \,,$$

where ξ is the deviation from the circular shape. The main assumption made is that the θ component of the fluid velocity v has expression

$$v = \alpha(r)\xi(\theta, t) + \mathcal{O}_2(\epsilon) \,,$$

where α is a function to be determined within the model. By combining the above hypotheses with the Euler equations, the equation of continuity, and the free surface boundary condition, and by considering the nonlinear terms up to order two in a smallness parameter ϵ, and still neglecting the surface tension, the authors in [329] obtain for ξ the equation

$$\frac{\partial\xi}{\partial t} + \left[1 + \beta - B(r_0) + \frac{1}{r_0}\xi\right]\frac{\partial\xi}{\partial\theta} + C(r_0)\frac{\partial^3\xi}{\partial\theta^3} = 0 \,, \tag{9.89}$$

where r_0 is the radius of the dry region at the center of the bottom in the absence of any waves or disturbances,

$$\begin{aligned}
\beta &= -\left[\sqrt{\ln^2(r_0) - \ln(r_0)} + \ln(r_0)\right] \,, \\
B(r_0) &= 2\alpha(r_0) - \beta \\
&\quad + \frac{1}{2}\left[\frac{1}{\beta}\int_1^{r_0} \frac{\alpha^2(r) - 4\alpha(r_0)}{r}\mathrm{d}r - 2\alpha(r_0) + \beta\right]\left[1 + \frac{1}{\beta}\ln(r_0)\right]^{-1} \,,
\end{aligned}$$

and

$$C(r_0) = -\frac{A(r_0)}{2}\left[1 + \frac{1}{\beta}\ln(r_0)\right]^{-1},$$

with

$$A(r) = \frac{1}{r_0}\int_1^{r_0}\frac{\mathrm{d}r}{r}\int_{r_0}^{r}\frac{\mathrm{d}r}{r}\int_1^{r}\alpha(r)\mathrm{d}r\,.$$

Equation (9.89) is known to have traveling periodic cnoidal wave solutions, similar to the spinning polygon solutions obtained for Leidenfrost drops in Sect. 9.8, in the form

$$\xi \sim \ \mathrm{sn}^2(k\theta - \omega t) + \text{const.}\,, \tag{9.90}$$

where the constants are determined from the problem parameters. These solutions are compared with the contour of the observed patterns and with sinusoidal harmonics with the same spatial period, troughs, and crests as the cnoidal solutions. It is interesting to mention that both cnoidal and sinusoidal solutions follow the contours of the observed patterns fairly well, exactly as was noticed in the case of hollow rotating polygons for Leidenfrost drops (see Fig. 9.29). The better agreement of both with the observations for the higher modes is due to the diminution of the wave amplitude. In the opinion of the authors in [329], the good match between the cnoidally shaped nonlinear waves and the linear sine solutions comes from the fact that the modulus of the Jacobi elliptic functions is relatively small. Given the good experimental match of all the models presented above, there remains the question of whether the real physical flow is linear or nonlinear.

Studying the same type of rotational fluid systems, the authors in [330] introduced the term *surface switching*. At low rotation rates, the flow is symmetrical, but the symmetrical shape breaks down at Reynolds numbers around 2000. In this regime, the deformation of the free surface is quite small, but at high rotation rates, the deformation becomes comparable to the scale of the vessel, and the free surface may become polygonal. It is these temporally periodic or non-periodic oscillations of the surface shape between differently deformed surface shapes with polygonal cross-sections that has been called surface switching. The flow transition of the boundary layer plays the role of a trigger for the initiation of turbulent flow, and the coupling between the flow transition and change in surface shape is a key factor in the surface switching mechanism [333].

More insight, and clarification regarding the physical mechanisms triggering the instability in such rotational systems, was brought by Lopez et al. through their Navier–Stokes 3D numerical model and experiments [334]. The authors observed two distinct physical mechanisms responsible for symmetry breaking, depending on the ratio of fluid depth to cylinder radius. For deep systems, the rotating wave results from the instability of the near-wall jet that forms as the boundary layer

on the rotating bottom end wall turns in toward the interior. In this case, the 3D perturbations vanish at the air–water interface. For shallow systems, the fluid at radii less than about half the cylinder radius is in solid rotation, whereas the fluid at larger radii has a strong meridional circulation. The interface between these two regions of flow is unstable to azimuthal disturbances and the resulting rotating wave persists all the way to the air–water interface.

9.9.2 Polygonal Eyewalls in Hurricanes

Although considerable insight into the physics of tropical cyclones has been acquired using axisymmetric theory and models, fundamental questions remain concerning the role of asymmetric processes in the cyclone life cycle. Questions associated with asymmetric potential vorticity redistribution in tropical cyclones can be studied with a hierarchy of dynamical models [335–338]. Formation of polygonal eyewalls, mesovortices, asymmetric eye contraction, and vorticity redistribution are natural phenomena whose dynamics is still far from being fully understood.

In [335], the authors performed numerical simulations in which a ring of elevated vorticity was perturbed with azimuthally broadbanded initial conditions. The simulations shown in this article indicate that the barotropic instability associated with annular regions of relatively high vorticity results in polygonal shapes before an ultimate rearrangement into a nearly circular vortex (see Figs. 9.34 and 9.35). In the eyewall of a hurricane, frictional convergence and moist convection act to concentrate high vorticity there, thus satisfying the necessary condition for

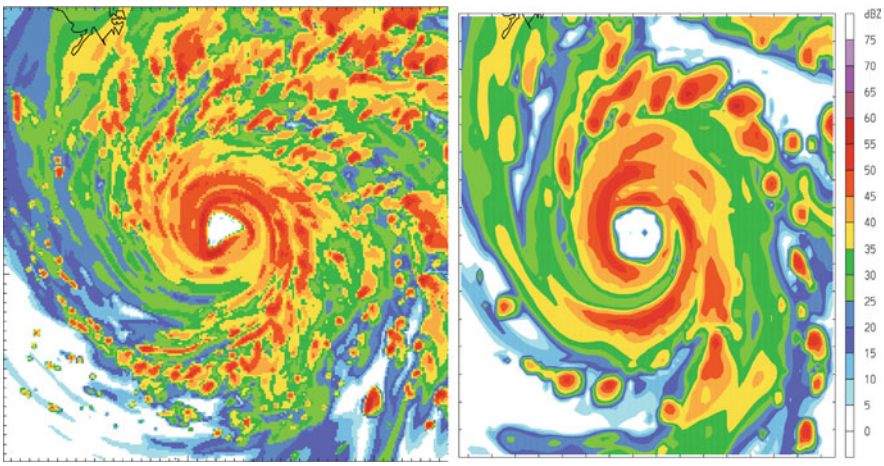


Fig. 9.34 Lower fuselage radar image of triangular and pentagonal patterns in the eyewalls of hurricane Rita. The eyewalls are about 2–4 km in size. Courtesy of NOAA, Hurricane Research Division, 28–29 August 2005

Fig. 9.35 An image of Hurricane Isabel (2003) as seen from the International Space Station, showing a well-defined square eye at the center of the storm. Courtesy NASA

barotropic instability. Natural convective asymmetries near the eyewall provide the perturbations that allow these instabilities to grow. And as they grow, the vorticity pools into a small number of pockets, creating the appearance of a polygon on the inner edge of the original annular region. These pools or pockets are also likely to be responsible for mesovortices.

In [337], the authors used a model with a higher level of complexity regarding the real atmosphere, namely, a shallow-water model. This model is based upon the forced dissipative divergent barotropic (shallow water) equations. The governing equations are Navier–Stokes equations with extra force terms, including the Coriolis force f, the mass sink effect (deep convective heating simulated in the model by a force density proportional to the depth h of the shallow-water model and the rate of mass sink Q), and Rayleigh friction terms (horizontal force of friction proportional to the velocity). The equations are written in terms of the vorticity, and by eliminating the vorticity divergence term from these equations, one obtains

$$\frac{\mathrm{D}P}{\mathrm{D}t} = PQ - \frac{\mu\zeta}{h} + \frac{\nu}{h}(\triangle\zeta - P\triangle h) \,, \tag{9.91}$$

where $\mathrm{D}/\mathrm{D}t$ is the Lagrangian (material) derivative, ν is the kinematic viscosity, μ is the Rayleigh coefficient of friction, ζ is the vorticity of the plane velocity field, and P is the *potential vorticity* defined by $P = (f + \zeta)/h$. Equation (9.91) is solved numerically with initial and boundary conditions.

The results of the simulations provide extensive explanations of the dynamics of the whole shallow-water model with very high chances of realistic interpretations. For example, the results show that *axisymmetric heating* makes a dominant contribution in the heating region. This happens because the mass sink Q produces more potential vorticity and concentrates vorticity, whence the circulation is increased in this region. An imbalance occurs for the vorticity between the inner and outer regions of the mass sink. As the potential vorticity ring breaks down during barotropic instability, potential vorticity is transferred asymmetrically from the eyewall to the eye, and a polygonal contour appears.

Another example of polygonal eyewall structure was provided by Hurricane Dolly (2008) before its landfall on the Texas–Mexico border. This hurricane exhibited dynamical inner-core structural variability, with an eyewall of highly asymmetric form, azimuthal wave number $m = 4$–7 measured by radar reflectivity, and prominent mesovortex and polygonal signatures [336]. When the diameter of the eyewall is approximately 45–50 km, the eyewall shape starts to vacillate between wave number patterns 4–6. On many occasions, the inner edge of the eyewall shows straight-line segments and polygonal shape.

To confirm that the origin of the Dolly asymmetries was dynamic instability of the eyewall, numerical simulations were conducted with a shallow-water model in [336]. The most likely cause of the high wave number asymmetries was a convectively modified form of dynamic instability in a thin potential vorticity ring.

In 2003, Hurricane Isabel revealed similar elaborate patterns within the clouds of its eye [339] (see Fig. 9.35). Through satellite images, it was a surprise for the tropical cyclone science community to observe a pentagonal pattern resembling a starfish in Isabel's eye due to the presence of six distinct mesovortices, one in the center and five others arranged symmetrically around it, a situation that remained fairly steady for a few hours while rotating cyclonically within the eye. From the theoretical standpoint, a two-dimensional barotropic flow model can resemble a vortex sheet supporting barotropic instability at high azimuthal wave numbers and fast growth rates.

9.9.3 From the Lab to Saturn

The 'hexagon' in Saturn's northern hemisphere was first noted in images taken by the Voyager spacecraft in 1988 (see Fig. 9.36 left frame), but it has also been studied more recently during different seasons using the Hubble Space Telescope. The polygon shows a long persistence, lasting several decades, indicating independence of seasonal variations due to solar forcing [340]. The basic ingredients known from rotating fluids or hurricane eyes are present in this case, too: existence of prominent

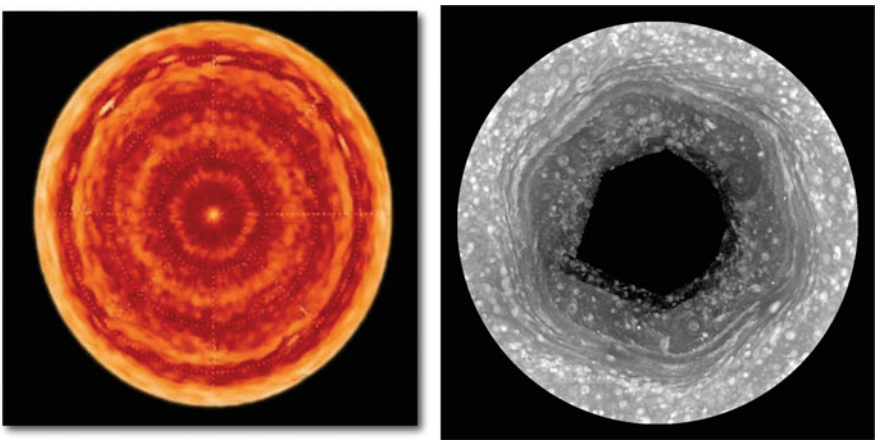


Fig. 9.36 *Left*: Six-sided jet stream at Saturn's north pole known as the hexagon. Courtesy of Cassini imaging, NASA. *Right*: Plasma experiment

vortices at the corners, the rotation from the bottom, and a major change in the rotation rate from a rapidly rotating center to an almost stagnant outer layer.

The latest research results tend to consider the possibility that the hexagon forms as the result of the nonlinear development of a predominantly barotropic instability coming from a strong zonal jet. In [340], the authors conducted a linear stability analysis of the flow, on the barotropic assumption. By solving the eigenvalue problem for the linearized barotropic vorticity equation for an inviscid flow, the authors obtained the growth rate of the radius of the flow as a function of zonal wave number for barotropic instabilities of the north polar jet on Saturn. For an appropriate observed radius of 2500 km, the speed of the eigenmode $m = 6$ is close to that of the zonal wind at the hexagon's latitude, and the position of the peak of the fastest growing mode relative to the jet coincides with the hexagon position relative to the jet, as observed experimentally. Moreover, for the polar jet in the southern hemisphere, this stability analysis predicts a growth rate that does not peak at finite wave numbers, and is weaker, which seems consistent with the absence of a polar hexagon counterpart at the south pole. Analysis of the observations from laboratory experiments (see Fig. 9.36 right frame), where there is instability of quasi-geostrophic barotropic jets and shear layers, supports the hypothesis that the long-lived polygonal structures correspond to wave modes caused by the nonlinear equilibration of barotropically unstable zonal jets.

Every year, new physical systems are discovered and studied at different scales, exhibiting symmetric rotational patterns, self-propulsion, compact liquid–gas structures, or high deformation of shapes as adaptive feedback to specific flow conditions. For example at the nanometer level, when electrons are injected into liquid helium, they force open a cavity free of helium atoms, referred to as an *electron bubble* [341]. The shape of the electron bubble is strongly dependent on the state of the electron and its quantum pressure (the gradient of the electron wave

function), which must balance the surrounding helium surface pressure. Under some conditions the electron bubble develops a neck and breaks up. At the micron level, blood micro-vesicles exhibit exotic shape transitions and aggregation. As the blood flow velocity increases, the fluid vesicle deforms from a biconcave disk shape into parachute and slipper-like shapes [342]. At higher scales, on the millimeter level, new types of levitated Leidenfrost liquid tori with polygonal rotating holes have been experimented [343]. Another example is the *crown splash* produced by the impact of a circular disc on a free liquid surface with generation and breakup of a splash wave created after the impact [344]. Brim waves on Leidenfrost drops and a special resonance at $f = 26$ Hz have been obtained recently. These waves are initiated by evaporation and lubricating flow which induce constant frequency capillary waves beneath the drop, and pressure oscillations under the drop that couple parametrically to the azimuthal star oscillations [345]. Recently, studies have been performed on droplet self-propulsion and spontaneous motion generated by their internal flow [346]. This inner flow changes the distribution of chemical reactivity inside the drop and consequently generates a surface tension gradient which produces the drop motion. The motion changes the drop geometry, and this symmetry breaking enhances the motion in a feedback effect.

9.10 Boundary of Axons and Nerve Pulse Propagation

When a thermodynamic system is forced out of its equilibrium state, the conservation of entropy requires a propagation phenomenon. This observation results from the fact that the nonequilibrium extension of Gibbs thermodynamics equations takes the form of a Fokker-Planck kinetic equation for the local balance of entropy and density. According to this equation, any external supply of entropy must be balanced by a combination of diffusive entropy flux at microscopic scale, and a macroscopic mass flux, the latest describing the local triggering of a system of waves. A thermodynamic system which is not in equilibrium is driven towards equilibrium by thermodynamic forces like chemical reactions, diffusion processes, electrical currents, and phase transitions. In far-from-equilibrium type of processes, Prigogine and Mazur have shown how local equilibrium can still be recovered when all of the relevant degrees of freedom (e.g. positions and velocities) are considered at the same level as the spatial coordinate.

Nerve impulses may provide a frame for such an interesting situation. The most relevant nerve function is the occurrence of the so called action potential as traveling waves generated and propagating along the axonal membrane (the cellular membrane of the elongated part of a nerve cell, the axon). More interestingly, experimental evidence of oscillations of temperature with zero sums of heat exchanges during passing of action potentials indicates an adiabatic process. There have been successful efforts to quantitatively describe the dynamics of the nerve pulses as an electrically driven phenomenon, namely the famous the Hodgkin-Huxley model. There have been also many attempts to describe the nerve pulses

as mechanically and thermodynamic driven, or at least to consider a minimal mechanical-thermal interaction.

The nervous impulse propagation was well established to be a membrane bound phenomenon where, although the cellular machinery and cytoskeleton structure certainly have a role, they are in no way required or necessary for nerve pulse propagation [347]. It was proved as early as 1909 by Einstein that the air bounded liquid interface, and in particular the capillary effect, is thermodynamical decoupled from the bulk, since the application of a Carnot cycle to a free surface requires a nonzero interface heat capacity and water interface entropy. Recent experiments and calculations show as well that a lipid monolayer at air-water interface has its own entropy, and therefore can be considered as an independent thermodynamic system. The hydrodynamic impedance mismatch between the bulk and the interface is another argument in favor of this boundary/bulk thermodynamic decoupling. Both approaches are based on the natural observation that a membrane which is the boundary of an incompressible medium reduces the dynamics of the bulk to the dynamics of the boundary.

A new perspective on the theory of pulse propagation in nerve is represented by a macroscopic thermodynamic model in which the action potential is regarded as an electro-mechanical solitary wave or soliton [79, 347, 348]. There is experimental and theoretical evidence of mechanical/acoustical waves propagating in the biomembrane, and these acoustic waves are reversible adiabatic transformations governed by entropy conservation. Measurements of real nerves in vitro prove that the travel of the action potentials is accompanied by localized mechanical deformation of the axon, namely in the axonal radius, axonal volume, and shortening of the axon at its terminus.

The critical fact though is that there is a liquid-gel phase transition of the biomembrane in the neighborhood of the equilibrium density, temperature and pressure, Fig. 9.37. From thermodynamics we know that the speed of sound (in the membrane, too) depends on the compressibility and density of the medium, the membrane. By including this phase transition in the speed of propagation of acoustic waves in the biomembrane one obtains a nonlinear dynamics similar to the dynamics of Boussinesq or Korteweg-de Vries solitons. From this nonlinear equations it emerges the existence of localized traveling waves (as nerve pulses)

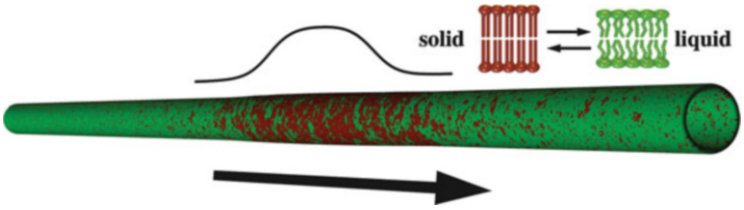


Fig. 9.37 Representation of a solitary wave traveling along the membrane of a nerve fiber. Small local changes in thickness are caused by pressure-induced order transitions in the membrane. *Red* regions correspond to ordered lipids in the otherwise liquid (*green*) lipid membrane. Courtesy of Thomas Heimburg, Niels Bohr Institute 2015

in the geometrical variables, and in density and electric potential. Not only this solitary waves and the corresponding nonlinear membrane model for nerve pulses explain more experimental results, but they also predict the elastic collision of such membrane solitons in a nerve. In addition, the soliton model for nerves is at present the only biophysics model that can explain and provide a quantitative description of the laws of anesthesia [349].

Appendix 1: Second Fundamental Form

Apart from the treatment of the geometry of graphs and networks, this book discusses surfaces embedded in $\mathbb{R}^3$, like the surface of a liquid drop. We thus present here some details of the differential geometry of smooth parameterized 2D surfaces embedded in $\mathbb{R}^3$. Moreover, when we discuss the expressions for the Hamiltonian or Lagrangian of free liquid surfaces, we use the concept of normal variation of a surface. In the following, we present a formal introduction to the variation of smooth parameterized compact surfaces embedded in $\mathbb{R}^3$. The discussion follows the work of Verpoort and Verstraelen [350]. The reader can supplement this section using various sources [136–140, 151, 152, 351, 352], where many more physical examples and abstract constructions can be found.

We consider a deformation of a compact surface $\Sigma \subset \mathbb{R}^3$ (in the following, we consider only smooth 2D parameterized surfaces embedded in $\mathbb{R}^3$) to be a smooth 'deforming' map $\mu : (-\epsilon, \epsilon) \times \Sigma \to \mathbb{R}^3$, $(t, p) \to \mu(t, p)$, $p \in \Sigma$, $\mu(0, p) = p$, and $\Sigma_t = \mu(t, \Sigma)$. Let $\mathcal{X}(\Sigma)$ be the 3D vector bundle over Σ and let $\boldsymbol{Z}_p \in \mathcal{X}(\Sigma)$ defined by

$$\boldsymbol{Z}_p = \left.\frac{\partial}{\partial t}\right|_{t=0} \mu(t, p)$$

be the *deformation vector field*, i.e., the tangent vector describing the trajectories traced out by the point p when the surface begins to be deformed. We can generalize this operator to any geometrical object defined on Σ, and for any deformation field $\boldsymbol{Z}$. So we define the *variation* of a tensor T along the deformation vector field $\boldsymbol{Z}$ by

$$\delta_{\boldsymbol{Z}} T = \left.\frac{\partial}{\partial t}\right|_{t=0} T\big(\mu(t, p)\big) \,.$$

We define the *shape operator* S by its action on vector fields $\boldsymbol{V} \in \mathcal{X}(\Sigma)$:

$$S : \mathcal{X}(\Sigma) \to \mathcal{X}(\Sigma) \,, \qquad S(\boldsymbol{V}) = -\boldsymbol{V}(\boldsymbol{n}) = [\boldsymbol{n}, \boldsymbol{V}] \,,$$

where $\boldsymbol{n}$ is the inner unit normal to Σ, that is, the (negative) derivative of the unit normal in the $\boldsymbol{V}$ direction. Its eigenvalues are the principal curvatures and its eigenvectors are the principal directions in the surface. It is easy to check that

the *tangent* variations of the mean and Gauss curvature are given by their Lie derivatives, i.e., if $\boldsymbol{V} \in T(\Sigma)$, then $\delta_{\boldsymbol{V}} H = \boldsymbol{V}H$ and $\delta_{\boldsymbol{V}} K = \boldsymbol{V}K$. Another interesting relationship in terms of S and H is $\nabla \cdot S = 2\nabla H$.

A deformation prescribed by the formula $\mu(t, p) = p + t\boldsymbol{Z}_p$ is called a linear deformation. Examples are translation of a plane or uniform expansion of a sphere. There is a very nice lemma [350] for the variation of the shape operator. Let $\boldsymbol{X} \in \mathcal{X}(\Sigma)$ and let $\boldsymbol{Z}$ be a linear deformation. Then,

$$\delta_{\boldsymbol{Z}} S(\boldsymbol{W}) = -\boldsymbol{W}(\delta_{\boldsymbol{Z}} \boldsymbol{n}) - S(\boldsymbol{W})\boldsymbol{Z} = [\delta_{\boldsymbol{Z}} \boldsymbol{n}, \boldsymbol{W}] + \left[\boldsymbol{Z}, [\boldsymbol{n}, \boldsymbol{W}]\right] . \tag{9.92}$$

The variation of the mean curvature under a linear deformation is

$$\delta H = -\frac{1}{2}\triangle f + (K - 2H^2)f + \boldsymbol{Z}^t(H) ,$$

where we make the decomposition $\boldsymbol{Z} = f\boldsymbol{n} + \boldsymbol{Z}^t$. As a corollary, we have a fundamental integral formula:

Theorem 14 *Let Σ be a compact surface, $\mathrm{d}\omega$ the volume form, and $\boldsymbol{X}$ a smooth vector field defined on it. Then,*

$$\int_{\Sigma} \mathrm{div}\boldsymbol{X}\, \mathrm{d}\omega = 0 .$$

For a surface Σ parameterized by the function $\boldsymbol{r}(u, v)$, we have the integral formulae of Jellett and Minkowski:

$$|\Sigma| = -\int_{\Sigma} (\boldsymbol{n}, \boldsymbol{r}) H \mathrm{d}\omega ,$$
$$\int_{\Sigma} H \mathrm{d}\omega = -\int_{\Sigma} (\boldsymbol{n}, \boldsymbol{r}) K \mathrm{d}\omega .$$

In [350], the author gathers four different geometric definitions for the second fundamental form on a surface, viz., $\Pi : \mathcal{X}(\Sigma) \times \mathcal{X}(\Sigma) \to C^{\infty}(\Sigma)$. We present these below and exemplify in Fig. 9.38.

Definition 12 Let Σ be a surface, $\Gamma(t)$ a smooth curve parameterized by t lying on it, $\boldsymbol{V}$ the tangent vector to Γ at a point $p \in \Gamma \subset \Sigma$, and $\boldsymbol{W}$ another vector belonging to the tangent plane at p, $\boldsymbol{V}, \boldsymbol{W} \in T_p(\Sigma)$. Then,

$$\Pi(\boldsymbol{V}, \boldsymbol{W}) = -\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{n}(t), W\right) ,$$

where $\boldsymbol{n}(t)$ is the inner unit normal to the surface along the curve Γ parameterized by t. This is illustrated in Fig. 9.38 (upper left).

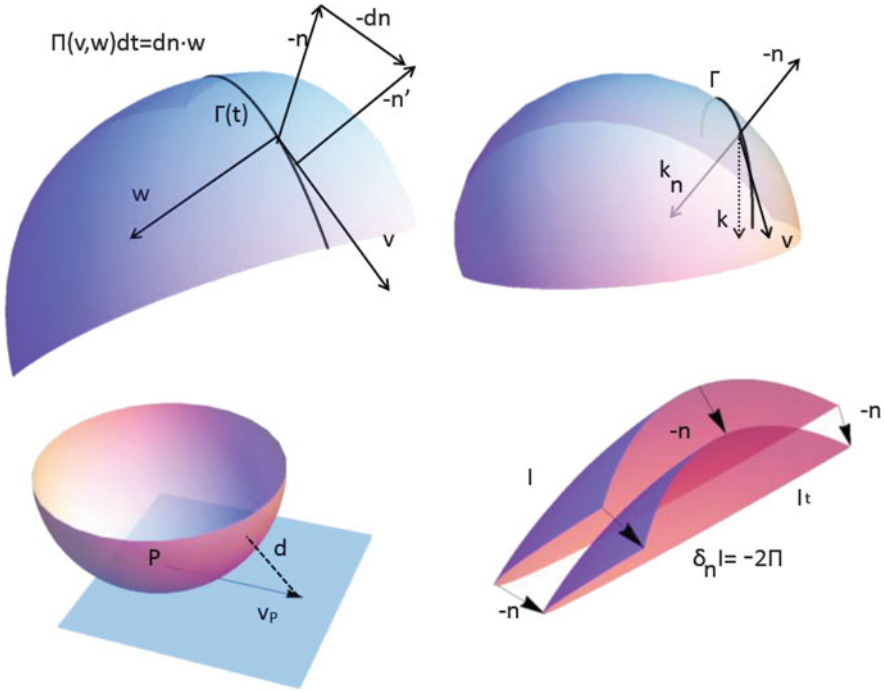


Fig. 9.38 Four definitions of the second fundamental form of a surface. *Upper left*: Variation of $\boldsymbol{n}$ along a curve in dot product with a tangent vector. *Upper right*: Normal curvature κ_n of a curve $\Gamma(t)$ lying on the surface M. The principal curvature of Γ is represented by κ. *Bottom left*: The second fundamental form also measures the distance between a point on the tangent plane and the surface. *Bottom right*: When a surface is deformed along its normal, the second fundamental form measures the rate of change of the area element

Definition 13 Let $\Gamma(t)$ be a parameterized curve lying on Σ, with $\boldsymbol{\tau}$ its tangent vector and κ_n its normal curvature on Σ. Then we have

$$\kappa_n = \Pi(\boldsymbol{\tau}, \boldsymbol{\tau}) \cdot \boldsymbol{n} \ .$$

This is illustrated in Fig. 9.38 (upper right).

Definition 14 Consider a point $p \in \Sigma$ and a tangent vector to the surface at p, $\boldsymbol{V} \in T_p(\Sigma)$. Draw the line $s\boldsymbol{V}$ along this vector. Then,

$$d(s\boldsymbol{V}, \Sigma) = -\frac{s^2}{2}\Pi(\boldsymbol{V}, \boldsymbol{V}) + \mathcal{O}(s^3) \ ,$$

where d is the distance from a point to the surface. This is illustrated in Fig. 9.38 (bottom left).

Definition 15 The most widely used definition is

$$\delta_{\boldsymbol{n}} I = -2\Pi \ .$$

The second fundamental form is thus a measure of the variation of the first fundamental form (surface area element) along a normal deformation (see Fig. 9.38 bottom right).

As one can see, since the first fundamental form measures the element of area on the surface, whence it is an intrinsic characteristic of the surface (independent of its embedding), the second fundamental form is a measurement of how bent and twisted the surface is in some embedding in a higher-dimensional space, whether it be described by the behavior of its normal (Definitions 12 and 13), or by how much deviates from its tangent plane (Definition 14), or finally, by how much its area shrinks (Definition 15).

An infinitesimal deformation is an *infinitesimal congruence* if the displacement of any point of Σ is along the normal at p to first order in the deformation parameter. An infinitesimal deformation is said to be isometric if the lengths of curves on the surface are stationary under the deformation. Then we have the Liebmann theorem which proves that any infinitesimal isometric deformation of the unit sphere is an infinitesimal congruence.

Appendix 2: Calculus of Variations

In the following V is a subset of the real vector space $\mathbb{R}^n$ with its canonical scalar product $\langle\ ,\ \rangle$, and $\lambda \geq 0$ a real non-negative parameter. The set V is *convex* if for any two of its points $x, y \in V$ the whole line segment from one to the other belongs to V, i.e., $\lambda x + (1-\lambda)y \in V$, $\forall \lambda \in [0,1]$. A function $f : V \to \mathbb{R}$ is a (strictly) *convex function* if

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y) \ , \quad \forall \lambda \in [0,1] \ ,$$

where for strict convexity we have strict inequality.

The function f has a global (local) minimum in V at $x_0 \in V$ if $\forall x \in V$ (in V intersected with some neighborhood of x_0), $f(x_0) \leq f(x)$, and we say that x_0 defines a strict global (local) minimum, if the strict inequality holds $\forall x \in V$, $x \neq x_0$. If a strictly convex function on a convex set has a minimum, this minimum is unique on that set. Moreover, for the same function and $\forall x, y \in V$, we have

$$f(y) - f(x) \geq \langle \nabla f, y - x \rangle \ ,$$

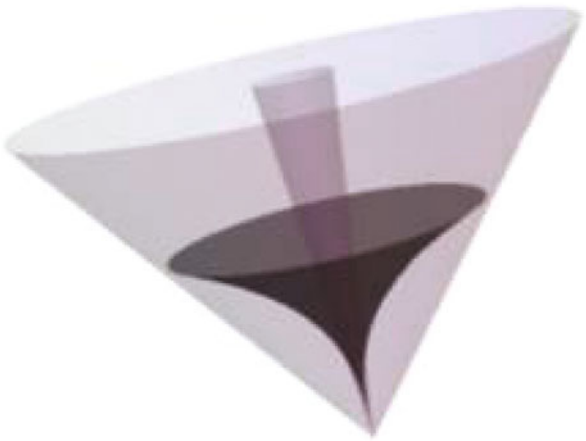


Fig. 9.39 The domain between the two straight cones is the local tangent cone for the parabolic cone presented in-between

and

$$\langle H(f)(x)y, y\rangle \geq 0\ , \quad y \in T(V, y)\ ,$$

that is, the function's Hessian is semi-positive definite at all points belonging to the *local tangent cone* $T(V, y)$ of V at y. By this cone, we understand the following construction (see Fig. 9.39). We define the *cone of normals* to V at y to be the closure of the set

$$N(V, y) = \left\{ z \in \mathbb{R}^n | \lambda z \in \nabla d_V(y) \right\}\ ,$$

for some $\lambda \in (0, \infty)$, where $d_V(y) = \min_{v \in V} |y - v|$ is the distance from the point y to the set V, and the gradient is calculated with respect to y. Then the local tangent cone $T(V, y)$ of V at y is the orthogonal set to the cone of normals of V at y. Put loosely, the local tangent cone is the set of all half-lines starting from y which intersect the domain whose boundary contains V at least once. If x_0 is a local minimum of f in V, then for any point z in the local tangent cone of V at x_0, we have $\langle \nabla f, z\rangle \geq 0$, for all $z \in T(V, x_0)$. In addition, if $f \in C^2(V)$, we have

$$\nabla f = 0\ , \quad \langle H(f); z, z\rangle \geq 0\ , \quad \text{for all } z \in V\ ,$$

where H is the Hessian matrix of f viewed as a bilinear form. Actually, it can be proved that $\langle H(f); z, z\rangle / \langle z, z\rangle$ is the first (smallest) non-negative eigenvalue of the matrix $H(f)$. Finally, we have the Kuhn–Tucker condition in the following form: if x_0 is a local minimizer of f in V, and V is defined by a system of equations $g_j(x) = 0$,

$j = 1, \ldots, m$ such that $\{\nabla g_j\}_{j=1,\ldots,m}$ is linearly independent, then there exists a set of parameters $0 \leq \lambda_j \in \mathbb{R}$ such that

$$\nabla f + \sum_{j=1}^{m} \lambda_j \nabla g_j = 0 \,, \qquad \sum_{j=1}^{m} \lambda_j g_j = 0 \,,$$

and $L = f + \sum_{j=1}^{m} \lambda_j g_j$ is called the Lagrange function.

In the following, we move from finite dimensions to variation of functionals. Let $u(x) \in B \subset C^1[a, b]$ be functions, $\|\cdot\|_B$ a certain norm on $C^1[a, b]$, and $E[u] : B \to \mathbb{R}$ a functional. The functional E is Gâteaux differentiable at u in direction h if there is a bounded linear operator $E'[u] : B \to \mathbb{R}$ defined by $E'[u] = \Phi'_{u,h}(0)$, where $\Phi_{u,h}(t) = E[u + th]$. We define the first and second Gâteaux variations of E at u in direction $h \in B$ by

$$\delta E[u](h) = \frac{\mathrm{d}}{\mathrm{d}t} E[u + th]\bigg|_{t=0} \,, \quad \delta^2 E[u](h) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} E[u + th]\bigg|_{t=0} \,. \tag{9.93}$$

If E is Gâteaux differentiable, then we have

$$\delta E[u](h) = \langle E'[u], h \rangle \,, \tag{9.94}$$

$$\delta^2 E[u](h) = \delta \langle E'[u], h \rangle = \langle E''[u], h \rangle + \langle E'[u], h' \rangle \,. \tag{9.95}$$

Furthermore, we say that $E[u]$ is Fréchet differentiable at $u \in B$ if there exists a bounded linear functional $DE : B \to \mathbb{R}$ such that

$$E[u + h] = E[u] + DE[h] + \mathcal{O}(\|h\|_B) \,, \quad \text{as } \|h\|_B \to 0 \,.$$

If the functional is Fréchet differentiable then it is also Gâteaux differentiable, and $E'[u] = DE[u]$.

If $u \in B$ is a local minimizer for E in B, then $\langle E''[u]h, h \rangle \geq 0$ for all h in B. We now present the two main theorems of variational calculus.

Theorem 15 *If* $u \in \{u \in B / u(a) = u_1, u(b) = u_2, u_{1,2} \in \mathbb{R}$ *fixed*$\}$ *is a local minimizer of the functional*

$$E = \int_a^b f(x, u, u') \, \mathrm{d}x$$

in B, then u satisfies the Euler differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x} f_{u'} = f_u \,.$$

The theorem works equally for vector functions, i.e., $u : [a, b] \to \mathbb{R}^n$, whence the Euler equation becomes a system

$$\frac{\mathrm{d}}{\mathrm{d}x} f_{u'_j} = f_{u_j} \ , \quad j = 1, \ldots, m \ .$$

Theorem 16 *If u is a weak (strong) solution of the Euler equation (or system of equations), and if $\langle E''[u]h, h\rangle \geq 0$ for all $h \in B$, then u is a weak (strong) local minimizer of E in B.*

By 'weak' or 'strong' in the theorem above, we mean satisfying the equation or the inequation in the weak sense, that is, in the $C^1[a, b]$ norm, or pointwise.

Appendix 3: *n*-Dimensional Rotating Drops

The stability of a rotating incompressible liquid drop, unaffected by gravity and cohered by surface tension, was the focus of an astonishing series of experimental and theoretical investigations starting with Joseph Plateau, the father of soap film studies, and pursued in the study of nuclear fission (the conjecture of Bohr and Wheeler in the model of heavy atomic nuclei), all the way to n-dimensional differential geometry generalizations.

In the following, we present a differential geometry model based on variational techniques and the implicit function theorem and which can be used to derive upper limits for the angular velocity as well as the existence, regularity, and stability of an energy minimizing family of rotating liquid drops in a neighborhood of the closed unit ball [353, 354].

We work in the n-dimensional Euclidean space $\mathbb{R}^n$ with coordinates $x = (x^i)$, and we define an incompressible liquid drop model as the class of compact and connected subsets of $E \subset \mathbb{R}^n$, $n \geq 2$, with finite prescribed volume $|E| = \Omega_n$ in the Hausdorff measure. Initially, the drop may have a unit sphere shape, that is $|x|_E(t = 0) = 1$. We recall that the volume of the n-dimensional ball (disk) $D^n \subset \mathbb{R}^n$ in this measure is

$$|E| = \Omega_n = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} \ . \tag{9.96}$$

In addition, we require the drop to have a class C^3 boundary $\partial E = M$ and to have its center of mass (or barycenter) placed at the origin and at rest:

$$\int_E x^i \, \mathrm{d}x = 0 \ , \quad \forall i = 1, \ldots, n \ . \tag{9.97}$$

For any positive angular speed Ω, we define the energy functional of the rotating drop model as the sum of the surface potential energy and the negative centrifugal

kinetic energy:

$$\mathcal{F}_\Omega(E) = \int_M \mathrm{d}A_M - \frac{\Omega^2}{2}\int_E |\pi_{\mathbb{R}^{n-1}}(x)|^2 \mathrm{d}x\,, \tag{9.98}$$

where $\mathrm{d}A$ is the area form and $\pi_{\mathbb{R}^{n-1}}(x)$ is the orthogonal projection of the position vector centered at the center of mass onto the hyperplane $x_n = 0$ perpendicular to the rotation direction. For star-shaped drops, the set E is a star-shaped region (that is, any straight line joining any point x in E to the origin belongs to E) and we chose a class $C^3(S^{n-1})$ map from the unit sphere parameterized by s to the boundary of E, viz., $X(s) : S^{n-1} \to N \subset \mathbb{R}^n$. With this notation and $|X| = r(s)$, the drop surface is a compact Riemannian manifold with metric induced by the standard Euclidean metric on the sphere $g^0_{ij}(s)$:

$$g_{ij}(s) = r^2 g^0_{ij}(s) + \nabla_i r \nabla_j r\,. \tag{9.99}$$

The mean curvature of the drop surface is

$$H(r) = \frac{n-1-|\nabla^N r|^2 - r\triangle^N r}{r\sqrt{1-|\nabla^N r|^2}}\,, \tag{9.100}$$

where ∇^N and $\triangle^N$ are the tangent to the N surface gradient and the Laplace–Beltrami operator, respectively [136–138, 353, 354].

By calculating the critical points of the first variation of the energy under the constraints of volume preservation and constant position of the center of mass, we obtain the corresponding Euler–Lagrange equations for the rotating drop:

$$\begin{aligned} H(r) - \frac{\Omega^2}{2}|\pi_{\mathbb{R}^{n-1}}(x)|^2 = X^j(s)\int_E \left[H(r) - \frac{\Omega^2}{2}|\pi_{\mathbb{R}^{n-1}}(x)|^2\right] X^i \mathrm{d}x \\ \times \int_N X^i(s)X^j(s)\mathrm{d}A\,, \end{aligned} \tag{9.101}$$

where $r = r(s)$, $x \in E$ on the left-hand side, and the two integrals are taken in the sense of $L^2(E)$ and $L^2(N)$, respectively. The left-hand side represents the variation of the energy and the right-hand side the Lagrange multipliers of the two constraints. In [353, 354], it is proved that, for any $\epsilon > 0$, (9.101) is solvable and has a smooth solution $r_\Omega(s)$ bounded by a closed neighborhood $\bar{B}_\epsilon$ of 0 for $\Omega < \epsilon$.

If $r(s)$ is a solution for the rotating drop with angular velocity Ω, and if we find $\epsilon > 0$ such that $\max_{s\in S^{n-1}} |r(s) - 1| \le \epsilon$, there is a bounded function $f(n)$ such that the solution is stable if

$$\Omega < \frac{n+1}{2}\big[1 - \epsilon f(n)\big]\,. \tag{9.102}$$

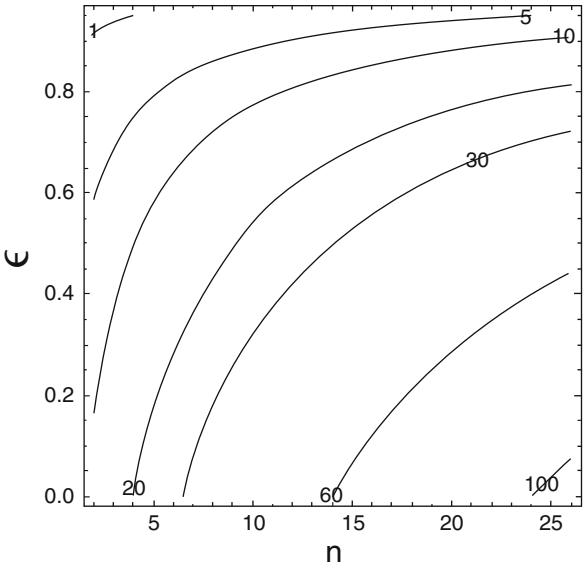


Fig. 9.40 Contour plot of the maximum angular speed of rotation $\Omega_{\max}(n, \epsilon)$ (in arbitrary units) for stable smooth solutions for the rotating liquid drop versus the number of dimensions of the space n, and the maximum deviation from the unit sphere of the drop shape. Smaller space dimensions and a more highly deformed shape decrease stability and the upper rotational limit

The stability is understood in the sense of the solutions corresponding to stable energy minimizers of the energy (9.98) of the rotating drop. This theorem is proved using a Rayleigh quotient operator weakly identical to the second variation of the energy functional in the Fréchet derivative sense. A sketch of the dependence of the upper bound of the angular speed for stable solutions is presented in Fig. 9.40.

There is an interpretation of rotations in more than three dimensions if we consider the phase space of a system containing a large number of particles, like a statistical or thermodynamic system, where n approaches the value of Avogadro's number. Consider a distribution function with compact and connected support in the phase space of a Hamiltonian dynamical system. According to the Liouville theorem, the motion of the density function through the phase space is a fluid flow of system points with zero convective derivative, that is, an incompressible flow [355]. In addition, the constant position constraint on the center of mass is satisfied simultaneously in both configuration and momentum space, so a system with its center of mass at rest will be represented by a phase space distribution whose center will also be at rest.

In principle, we can create the equivalent of a liquid drop in the phase space satisfying the constraints (9.96) and (9.97). We can induce a phase space rotation of the drop if this distribution is a Wigner distribution of a Gaussian wave packet in a quadratic trapping potential, like a harmonic oscillator. It is well known that the flow of the Wigner distribution follows circular paths in the phase space. It is also known

that a straight motion in the phase space is equivalent to a Galilean transformation in the configuration space. This means that the states of a Galilean invariant many-body system constrained to have cyclic dynamics will perform liquid drop rotations in the phase space, with a property of inertia (tendency to conserve straight uniform translations in the phase space). In this context, we can assimilate the fluid motion in the phase space as the rotation of a liquid drop with constant volume, constant center of mass position, and a centrifugal field of forces.

Such a virtual motion can be examined formally by the above theoretical approach for n-dimensional drops in rotation. If we accept this model, we can infer from the stability criterion for the rotating drop shape presented above that an increase in the number of particles will allow rotations with higher angular speeds to remain stable. The more particles in a statistical system forced to oscillate, the higher the eigenfrequencies of its stable oscillations.

This observation may work like a universality principle for large amplitude high frequency oscillations in many-body systems: more particles and hence more collective interactions allow higher frequencies of stable oscillations. This observation is in agreement with many current models, theories, and experiments, like the Langumir waves (plasma oscillations in which the plasma frequency increases as the square root of the electron density), the oscillation frequencies of the collective excitations of a trapped Bose–Einstein condensate (the leading corrections to the frequencies are proportional to the number of atoms in the condensate to the power 1/5 [356]), the frequency of the collective oscillations of a trapped Fermi gas (the frequency is proportional to the square root of the electron concentration [357]), the self-maintained coherence effect in collective neutrino oscillations (the frequency increases with the neutrino gas density [358]), etc.

Chapter 10
Conclusions

The perspective in this book arises from several decades of experience studying the importance of boundaries for various dynamical or complex systems: drops, clusters, exotic nuclei, swimming cells, etc. We have tried to present a coherent view of the common features of the boundary in the arts and sciences, mainly using a mathematical language. The list of topics demanding a serious study of the dynamics of boundaries is much longer than we have presented here. Among many intriguing new domains of research not mentioned here, we could consider fractal boundaries [119, 359], the dependence of the shape of the boundary of a collectivity as a function of the strength of the interaction between its members, the importance of the axon membrane in memory processes, biological systems with free boundaries, patterns in the universe, tumor growth, coating flows, boundaries of pseudo-manifolds and non-manifolds (for example, systems that are partially 3D and partially 2D, or even 1D), biological morphogenesis, dynamics of the benthic boundary layer modeled as a framed bordism, and so on.

The evolution of societies unfolds in space, time, and basic needs (the principles of Glaser's choice theory). The interrelations between these directions develop specific metrics adapted for normal human scales: tools, means of existence, storage. Time was historically measured in terms of the 'human scale': natural cycles, life/death, etc. Social relations were initially structured topologically, locally connected and compact (family, tribe). Later on, societies tried to extend the scales of time, space, and human activity: the struggle for large spatial scales (travel, discovery, astronomy, cosmology), understanding life and trying to prolong it, and improving living and working conditions through buildings, cities, hospitals, schools, and empires. With success in these essential struggles, other parallel struggles grew in importance for societies. The struggle toward smaller scales in medicine and pharmacy, clocks, miniature art, atoms, molecular biology, and elementary particles. These struggles in space and time have caused societies to move faster (rapid photography, relativity), or more slowly (cosmological time, quality of life). In the social communication direction, scaling was initially

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5_10

determined by biological criteria like the human senses (distance for talking, range for signaling, smelling, tasting), but later, these scales were expanded by exploiting electromagnetism, telephone networks, radio networks, TV, and the Internet.

While the history of human creation has always recorded races towards the expansion of various scales, it is true that many scales or directions have been shown to be bounded by some fundamental limit: the speed of light, the mass of the universe, the Planck scale, and so on. Somehow, mathematics tells us that this spontaneous occurrence of boundaries should not come as a surprise. We know from topology that any topological space, even if it is unbounded, open, and not compact, can be embedded in a compact space. Topologists call this *compactification*. There are quite a few methods for compactifying a space. One of the most common, the Alexandroff compactification, works by adding one extra point called ∞ to the non-compact space. This is exactly how Einstein and Minkowski brought in the speed of light in vacuum, and changed the unbounded Galilean transformations into the compact Lorentz pseudo-rotations. In general, one can add a new point 'at infinity' in any direction where an infinite space tends to escape. In real life, in the observable universe, we may be able to prove that any type of expansion of natural or social scales can be embedded in a sort of compactification. Compactness, or the presence of boundaries and frames, are not always easy to establish. Deleuze and Guattari wrote [360, p. 189]: "It has been said that sound has no frame". At first sight this makes sense: sounds, music, and language seem unbounded. Further on, however, they note that "music [...] embodies the frame even more powerfully." It seems the boundaries and frames are not always visible, but may become visible when one changes the system in which one embeds them.

In order to predict the future directions of expansion of social scales, we may need to comprehend what other new dimensions need to be introduced in addition to space, time, behavioral needs, and communications. More importantly, we need to see what coordinates can describe such new dimensions, and what scales will correspond to those associated with the traditional human struggles for 'larger', 'smaller', 'faster', 'slower', etc. Maybe some of these new dimensions will contain questions relating to the largest intelligent memory, maximum computer storage and transfer speed (no matter what medium is used for this), the most general theory, Virilio's information bomb or his accident theory, etc. Maybe the gaming part of the brain will become more commonly used in everyday life, or maybe the rapid communication environment and global knowledge will change our perspective on boundaries.

From our present experience we can infer that in the evolution of any system there is a correlation between its degree of complexity, the nonlinearity of its response, and the use of its boundaries. Whether this correlation is subjected to some sort of uncertainty principle is not quite clear. For example, physicists have found that humans can discriminate a sound's frequency and timing more than 10 times better than the limit imposed by the Fourier uncertainty principle, a result which rules out the majority of auditory processing algorithms that have been proposed for the brain, since only a few models can match this impressive human performance [361]. Maybe, of all the senses, olfaction will prove to be the sense most firmly connected

to the existence of external boundaries. While vision and hearing have a strong spectral dependence and what is a boundary for one wavelength may not be for another, olfaction is strongly related to the material support of odorous substances.

There is experimental evidence that some quantitative descriptor of human evolution and civilization dynamics tends to follow certain analytic laws, such as Moore's law conjecturing that, during the history of computer hardware, the number of transistors in a dense integrated circuit will double approximately every two years. There are other similar conjectures on the exponential dynamics of human products, such as Rock's law which predicts that the capital cost of a semiconductor fabrication plant will increase exponentially over time. Furthermore, there are conjectures concerning the dynamics of hard disk drive area density, network capacity, cost of pixels, DNA sequencing technologies, the Malthusian growth model for populations, and library expansion, or the very recent self-replicating colloidal cluster model [362].

All these conjectures have in common an exponential or power law time behavior, accelerating change, and no predictions of the existence of a limitation or plateau. It is very likely that models for the dynamics of these parameters of human civilization will not follow an analytic shape or indeed be generated by a time-dependent dynamic differential equation. At the same time, the existence of fundamental limits for these time variable parameters seems to be a realistic hypothesis. The occurrence of such limiting asymptotic behavior and the levelling out of the time evolution must be related to the existence of a fixed point (such as the speed of light in vacuum for the Lorentz transformations), which at the same time must be related to the compactness of the system and the contraction property of the evolution operators.

Another possibility would be to hypothesize that there may be time-local dynamical laws governing the evolution, but laws that themselves change. For example, the order of differentiation may change with time in a differential equation. More interestingly, we may conjecture that the order of differentiation of the dynamical prediction itself changes with time, according to some other law. This conjecture would require the use of fractional derivatives and fractional calculus. A lot of progress has been made in this field and there are already existence and uniqueness theorems for the solutions of ordinary differential equations with fractional derivatives [363]. Consequently, it may not be too exotic to consider the evolution of such quantities as governed by differential equations in which the order of differentiation is itself an independent variable, together with the time.

As mentioned at the beginning of Chap. 2, the peripatetic principle is able to relate human sensorial perceptions to mathematical cognition. There is a large body of evidence from experimental and observational studies that mathematical knowledge arises from rudimentary knowledge acquired by perception. It is known [364] that both foundations of mathematics, that is, arithmetic and geometry, have developed brain structures by extensive ontogenetic and phylogenetic perception experiences related to hearing and motion on the one hand, and sight and touch on the other. There is a fundamental relationship between visual perception and geometric abstractions, since simple objects from everyday life can be mentally

decomposed into points, lines, angles, surfaces, and more general abstract shapes [83]. Öğmen and Herzog found that geometry is the appropriate branch of mathematics for understanding how information is represented and processed in the visual system. Conversely, the visual part of our brain is the part responsible for allowing us to discover and use geometry. There is also recent evidence that both humans and primates have the capacity to predict and represent auditory sequences using a mechanism that is similar to, and co-located in the brain with the one that deals with arithmetic abilities [365].

Further support for the existence of a primordial type of mathematics in our brain can be found in studies by Bressloff and coworkers (see, e.g., [366]), where they prove that the geometrical structure of the primary visual cortex is responsible for the entoptic shapes of 'form constants' which appear in cave art, or are seen when a subject is suffering from sensory deprivation, exposed to rhythmic music or flickering lights, or has ingested hallucinogenic drugs, for example. Such geometric visual hallucinations are not images of external visual objects, but rather patterns of neural firing which constitute a new state of equilibrium which the brain triggers by reacting to the above-mentioned stress agents.

In this natural bijection (sight+touch $\leftrightarrow$ geometric intuition, and hearing $\leftrightarrow$ algebraic and spoken language intuition), there is a dilemma raised by dance and ballet, by rhythmic motion of the hands when humming, or singing or listening to music, or by conducting an orchestra, because we have a mixture of visual and auditive information. However, on a closer analysis, we note that the mechanics of hand motion and even whole body motions is represented by a finite-dimensional space, and by a finite-dimensional number of degrees of freedom, namely 230. At the same time, visual information is mediated by the electromagnetic field, which is a physical system with an infinite number of degrees of freedom, through the visual system, but also counts 75–150 million rods in the retina, and 1.1–1.2 million nerve fibers in the optic nerve. So even visually perceived, acts of ballet, dance and conducting, are rather connected to hearing and music, hence to the algebraic proto-structures in the brain, rather than to geometry and the visual system. The body motions are purely mechanical motions of a deformable solid system, and these motions can be faithfully represented by a time series for the evolution of a vector in some finite-dimensional space. Even the dance figures of a ballet company would be describable through a time-dependent vector with a finite number of components, that is, just the motion of one point in an n-dimensional space. In contrast, image and painting arrive in the brain instantaneously as a huge field of data. Vision dynamics is a field theory type of dynamics, with infinite-dimensional degrees of freedom.

The question about the importance of frame in a general context can be addressed to every type of art. Forty million olfactory receptor neurons are responsible for the sense of smell, but there is no formal development for the art of olfaction. In a 2014 article in *Style* about Chandler Burr, the curator for New York's Museum of Art and Design, Afsun Qureshi writes: "By any reasonable and rational definition that applies to an art medium, scent is one." At the same time, our long olfactory experience must have developed some special structures and abilities in the brain, similar to the way mathematical abilities (arithmetic and geometry) have developed

by hearing and sight. For humans, the latter are the most highly developed senses, but for other animals, olfaction is in many cases the most important sense [367]. Through olfaction, we would expect animals to have developed other cognitive abilities relating to a 'different' type of mathematics, and not 'our mathematics', which is based on knowledge of quantities and their algebraic and geometrical relations. Cognitive ethology and behavioral ecology have developed a growing body of evidence that members of many species can judge proportions, amounts, sounds, time intervals, smells, and so on.

It may be that the field of ethological and cognitive science investigates the comparison between animal and human abilities only to the extent of reasoning in terms of numbers or paths and shortcuts, but not beyond. It still lacks an adequate language for a comparative analysis through other types of reasoning based on olfaction [368]. If there is a certain type of olfactory cognition and reasoning, it would be interesting to understand what might be the representation of topology and boundaries in this language.

Another unusual example of the importance of frames is seen in culinary art, where objects are always bounded by tables, platters, or plates. In the Western tradition, a meal typically contains one piece of meat, a salad, a side of vegetables, and maybe a tray with spices, all bounded by an imaginary horizontal frame delimiting the consumer space at the table. In Japanese cuisine, sushi confers a different topological structure. We can define a sushi plate (SP) by starting from a Western plate (WP) followed by a compression $\mathbf{C}$, and then a linear combination of translated (T) versions of this compression, pretty much as in the wavelet decomposition:

$$SP = \sum_{k=1}^{n} c_k \mathbf{T}_k \mathbf{C} \, (WP) \; .$$

This formula shows us that the frame around an image can generate unique effects in the perception and processing of an image, with consequences for the cognition, recognition, and decisions of an observer. To quote Cavanagh [49], "the artist can take shortcuts, presenting cues more economically ... to suit the message of the piece rather than the requirements of the physical world."

References

1. M. Walter, E.B. Ebbesen, A.R. Zeiss, J. Pers. Soc. Psychol. **21**(2), 204–218 (1972)
2. W. Wolf, W. Bernhart, *Framing Borders in Literature and Other Media* (Editions Rodopi B.V., Amsterdam, 2006)
3. J.D. Beckenstein, Phys. Rev. D **23**(2), 287–298 (1981)
4. G.E. Moore, IEEE Solid-State Circ. Newslett. **3**(20), 33–35 (2006)
5. S. Lloyd, Nature **4006**, 1047–1054 (2000)
6. C.R. Shalizi, *Methods and techniques of complex systems science: an overview*. Preprint `arxiv.nlin:0307015` (2007)
7. C.R. Shalizi, K.L. Shalizi, R. Haslinger, Phys. Rev. Lett. **93**, 1118701 (2004)
8. M.S. Livingstone, Science **290**, 1299 (2000); M. Livingstone, D. Hubel, Science **240**, 740–749 (1988)
9. P. Mamassian, Vision Res. **48**, 2143–2153 (2008)
10. S.C. Pont, H.T. Nefs, A.J. van Doorn, M.W.A. Wijntjes, S.F. te Pas, H. de Ridder, J.J. Koenderink, Seeing Perceiving **25**(3–4), 339–340 (2012)
11. R.N. Haber, Am. Sci. **68**(4), 370–380 (1980)
12. P.H. Nidditch, J. Locke, *An Essay Concerning Human Understanding* (Oxford University Press, Oxford, 1979); H.H. Grelland, *Space-Time, Phenomenology, and the Picture Theory of Language*, vol. 167 (2010), pp. 281–290
13. Ph. Blanchard, J.R. Darwin, D. Volchenkov, Eur. Phys. J. Spec. Top. **184**, 1–82 (2010)
14. A.W. Toga, J.C. Mazziotta, *Brain Mapping: The Methods* (Academic, London, 2002)
15. K. Gröchenig, *Foundations of Time–Frequency Analysis* (Birkhäuser, Boston, 2001)
16. G. Azzopardi, N. Petkov, Biol. Cybern. **106**(3), 177–189 (2012)
17. R.E.B. Mruczek, I.S. von Loga, S. Kastner, J. Neurophysiol. **109**(12), 2883–2896 (2013)
18. B. Kamai et al., Am. Astron. Soc. Meet. Abs. **221** (2013); M. Moyer, Sci. Am. **306**, 30–37 (2012)
19. R. Hayman, M.A. Verriotis, A. Jovalekic, A.A. Fenton, K.J. Jeffery, Nat. Neurosci. **14**, 1182–1188 (2011)
20. F. Savelli, J.J. Knierim, Nat. Neurosci. **14**, 1102–1103 (2011)
21. A.D. Ekstrom et al., Nature **425**, 184–188 (2003)
22. A.D. Ekstrom, Curr. Biol. **24**(4), R167–R168 (2014)
23. M.A. Goodale, Proc. R. Soc. B Biol. Sci. **281**(1785), 20140337 (2014)
24. J. Maldacena, Science **344**(6186), 806–807 (2014); J. Nishimura, Prog. Theor. Exp. Phys. **01A101** (2012); C.J. Hogan, Phys. Rev. D **85**(6), 064007 (2012); M. Hanada, Y. Hyakutake, G. Ishiki, J. Nishimura, Science **344**(6186), 882–885 (2014)
25. N. Ulanovsky, Curr. Biol. **21**(21), R886–R887 (2011)

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5

26. P. Duro (ed.), *The Rhetoric of the Frame* (Cambridge University Press, Cambridge, 1996)
27. S.M. Ebenholtz, G.W. Glaser, Percept. Psychophys. **32**(2), 134–140 (1982)
28. J.A. Saunders, B.T. Backus, J. Vis. **6**(9), 933–954 (2006)
29. S. Coren, Psychol. Rev. **79**(4), 359–367 (1972)
30. M.A. Georgeson, G.D. Sullivan, J. Physiol. **252**, 627–656 (1975)
31. W.S. Giesler, J.S. Perry, J. Najemnil, J. Vis. **6**, 858–873 (2006)
32. T. Knapen, M. Rolfs, M. Wexler, P. Cavanagh, J. Vis. **10**(1, 8), 1–13 (2010)
33. F.C. Fortenbaugh, S. Sanghvi, M.A. Silver, L.C. Robertson, J. Vis. **12**(2, 19), 1–18 (2012)
34. A.M. Herbert, G.K. Humphrey, P. Jolicoeur, Can. J. Exp. Psychol. **48**(1), 140–148 (1994)
35. K. Ferrara, S. Park, J. Vis. **13**, 9 (2013)
36. M. Ozawa, Phys. Rev. A **67**(4), 042105–042110 (2003)
37. R. Paley, N. Wiener, *Fourier Transforms in the Complex Domain*, vol. XIX (AMS Colloquium Publications, New York, 1934)
38. J.M. Chen, J. Smith, J. Wolfe, Science **319**, 726 (2008)
39. A. Haar, Math. Ann. **69**, 331–371 (1910)
40. I. Daubechies, Commun. Pure Appl. Math. **41**(7), 909–996 (1988)
41. J. Morlet, *Issues in Acoustic Signal-Image Processing and Recognition* (Springer, Berlin, 1983), pp. 233–261
42. S.G. Mallat, *A Wavelet Tour of Signal Processing* (Academic, London, 1999)
43. S.E. Kelly, M.A. Kon, L.A. Raphael, J. Funct. Anal. **126**(1), 102–138 (1994)
44. J.G. Daugman, J. Opt. Soc. Am. A **2**, 1160–1169 (1985)
45. L. Bonnar, F. Gosselin, P.G. Schyns, Perception **31**, 683–691 (2002)
46. J. Intriligator, P. Cavanagh, Cogn. Psychol. **43**, 171–216 (2001)
47. M. Bochner, *Solar System & Rest Rooms: Writings and Interviews* (1965–2007), pp. 96–101
48. A. Kranjec, J. Cogn. Neurosci. **25**(12), 2015–2024 (2013)
49. P. Cavanagh, Nature **434**, 301–307 (2005); Spat. Vis. **21**, 261–270 (2008)
50. C. Pinhanez, M. Podlaseck, To frame or not to frame: the role and design of frameless displays in ubiquitous applications, in *UbiComp*, ed. by M. Beigl et al. Lecture Notes in Computer Science, vol. 3660 (Springer, Berlin, 2005), pp. 340–357
51. B. Sayim, P. Cavanagh, Iperception **2**(7), 679 (2011)
52. W. IJsselsteijn, H. de Ridder, R. Hamberg, D. Bouwhuis, J. Freeman, Displays **18**, 207–214 (1998)
53. R. Allen, Representation, illusion and the cinema, in *The Visual Turn: Classical Film Theory and Art History*, ed. by A.D. Vacche (Rutgers University Press, New Brunswick, 2003)
54. T.L. Hubbard, J.L. Hutchinson, J.R. Courteny, Q. J. Exp. Psychol. **63**(8), 1467–1494 (2010); S.L. Mullally, H. Intraub, E.A. Maguire, Curr. Biol. **22**, 261–268 (2012); P. Chapman, D. Ropar, P. Mitchell, K. Ackroyd, Vis. Cogn. **12**(7), 1265–1290 (2005)
55. S. Lowenstam, Trans. Am. Philos. Assoc. **127**, 21–76 (1997)
56. D. Von Bothmer, Metropol. Museum Art Bull. **31**(1), 3–9 (1972)
57. J. Bazant, Letras Clássicas **8**, 11–26 (2004)
58. T. Grieder, Am. Antiq. **29**(4), 442–448 (1964)
59. G. Ferrari, Class. Antiq. **22**, 37–54 (2003)
60. R.T. Neer, *Style and Politics in Athenian Vase-Painting* (Cambridge University Press, Cambridge, 2002)
61. J.L. Reinish, Stud. Int. **186**(958), 63 (1973)
62. M. Lyons, J. Budynek, S. Akamatsu, Classifying images of facial expression using a Gabor wavelet representation, in *Proceedings of 2nd International Conference on Cognitive Science*, Tokyo (1999), pp. 113–118
63. A.M. Herbert, G.K. Humphrey, P. Jolicoeur, Can. J. Exp. Psychol. **48**(1), 140–149 (1994)
64. P.A. Williams, J.T. Enns, Perception **25**, 921–926 (1996)
65. M. de Montalembert, P. Mamassian, Neuropsychologia **48**, 3245–3251 (2010)
66. S. Kennett, M. Taylor-Clarke, P. Haggard, Curr. Biol. **11**(15), 1188–1191 (2001)
67. S. Grossberg, Percept. Psychophys. **55**(1), 48–120 (1994)
68. P. Bak, C. Tang, K. Wiesenfeld, Phys. Rev. Lett. **59**(4), 381–384 (1987)

69. T. Luft, C. Colditz, O. Deussen, ACM Trans. Graph. **25**(3), 1206–1213 (2006)
70. E.R. Kandel, J.H. Schwartz, T.M. Jessell, *Principles of Neural Science* (McGraw-Hill, New York, 2000), pp. 515–520; T. Heiberg, B. Kriener, T. Tetzlaff, A. Casti, G.T. Einevoll, H.E. Plesser, J. Comput. Neurosci. **35**, 359–375 (2013); G. Azzopardi, N. Petkov, Biol. Cybern. **106**(3), 177–189 (2012)
71. B.S. Manjunath, R. Chellappa, IEEE Trans. Neural Netw. **4**(1), 96–108 (1993)
72. H.K. Hartline, Am. J. Physiol. **121**, 400–415 (1938)
73. P. Cavanagh, S. Anstis, Vis. Res. **91**, 8–20 (2013)
74. J. Schramme, Vis. Res. **32**(11), 2129–2134 (1992)
75. E.D. Grunfeld, H. Spitzer, Vis. Res. **35**(2), 275–283 (1995)
76. C. von Campenhausen and J. Schramme: Perception **24** (1995) 695–717
77. C.E. Benham, Nature **51**, 200 (1894)
78. C.F. Stromeyer III, R.J.W. Mansfield, Percept. Psychophys. **7**(2), 108–114 (1970)
79. E.V. Vargas, A. Ludu, R. Hustert, P. Gumrich, A.D. Jackson, T. Heimburg, Biophys. Chem. **153**(2–3), 159–167 (2011)
80. T. Hartley, C. Lever, Neuron **82**(1), 1–3 (2014)
81. T. Hartley, C. Lever, N. Burgess, J. O'Keefe, Philos. Trans. R. Soc. B **369**(1635), 20120510 (2014)
82. T.L. Bjerknes, E.I. Moser, M.-B. Moder, Neuron **82**(1), 71–78 (2014)
83. H. Öğmen, M.H. Herzog, Proc. IEEE **98**(3), 479–492 (2010)
84. E. Cartan, C. R. Acad. Sci. Paris **196**, 582–586 (1993)
85. J.D. Mollon, Vis. Neurosci. **23**, 297–309 (2006)
86. A. Gilchrist, S. Delman, A. Jacobsen, Percept. Psychophys. **33**(5), 425–436 (1983)
87. R. Shapley, M.J. Hawken, Vis. Res. **51**, 701–717 (2011)
88. J.F. Iaccino, *Left Brain–Right Brain Differences: Inquiries, Evidence, and New Approaches* (Psychology Press, East Sussex, 2014)
89. N.N. Nikolaenko, Acta Neuropsychol. **1**(1), 144–158 (2003)
90. N.N. Nikolaenko, M. Brener, J. Evol. Biochem. Physiol. **39**(4), 491–501 (2003)
91. N.N. Nikolaenko, A.V. Egorov, E.A. Freiman, Behav. Neurol. **10**, 49–59 (1997)
92. P. Almer, Framing the real: frames and processes of framing in René Magritte's Œuvre, in *Framing Borders in Literature and Other Media*, ed. by W. Wolf, W. Bernhart (Rodopi B.V., Amsterdam, 2006)
93. M. Lamont, V. Molnár, Annu. Rev. Soc. **28**, 167–195 (2002)
94. M.A. Pachucki, S. Pendergrass, M. Lamont, Poetics **35**(6), 331–351 (2007)
95. F. Barth (ed.), *Ethnic Groups and Boundaries* (Allen and Unwin, London, 1969)
96. A. Donaldson, D. Wood, Environ. Plan. D **22**, 373–392 (2004)
97. E.A. Phelps, K.J. O'Connor, W.J. Cunningham, E.S. Funayama, J.C. Gatenby, J.C. Gore, M.R. Banaji, J. Cogn. Neurosci. **12**(5), 729–738 (2000)
98. M. Emirbayer, Am. J. Sociol. **103**(2), 281–318 (1997)
99. D. Wood, S. Graham, Permeable boundaries in the software-sorted society: surveillance and the differentiation of mobility, in *Mobile Technologies of the City*, ed. by M. Sheller, J. Urry (Routledge, London, 2006), pp. 177–191
100. M. Castells, Contemp. Sociol. **29**(5), 693–699 (2000)
101. A.K. Watkins, *Pixelated Urbanism*. Thesis, Department of Architecture, University of Washington, 2011
102. B.N. Vis, Establishing boundaries: a conceptualization for the comparative social study of built environment configurations, in *Spaces and Flows: An International Conference on Urban and Extra-Urban Studies* vol. 2 (Common Ground Public LLC, Champaign, 2013), p. 4
103. B. Latour, *Reassembling the Social: An Introduction to Actor–Network Theory* (Oxford University Press, Oxford, 2005); M. Callon, Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay, in *Power, Action and Belief: A New Sociology of Knowledge*, ed. by J. Law (Routledge & Kegan Paul, London, 1986)
104. P. Sloterdijk, *Regeln für den Menschenpark* (Suhrkamp, Frankfurt am Main, 1999)

105. H. van Houtum, R. Pijpers, Antipode **39**(2), 291–309 (2007)
106. S. Havlin, D.Y. Kenett, E. Ben-Jacob, A. Bunde, R. Cohen, H. Hermann, J.W. Kantelhardt, J. Kertész, S. Kirkpatrick, J. Kurths, J. Portugali, S. Solomon, Eur. Phys. J. Spec. Top. **214**, 273–293 (2012)
107. N. Gilbert, *Agent-Based Models*. Quantitative Applications in the Social Sciences, Series, vol. 153 (Sage, Los Angeles) (2008)
108. T.A.B. Snijders, Annu. Rev. Sociol. **37**, 131–153 (2011)
109. N. Pinter-Wollman, E.A. Hobson, J.E. Smith, A.J. Edelman, D. Shizuka, S. de Silva, J.S. Waters, S.D. Prager, T. Sasaki, G. Wittemyer, J. Fewell, D.B. McDonald, Behav. Ecol. **25**, 242–255 (2013)
110. A.D. Henry, P. Pralat, C.-Q. Zhang, Proc. Natl. Acad. Sci. **108**(21), 8605–8610 (2011)
111. E.S. Bogardus, Sociol. Soc. Res. **17**, 265–271 (1933)
112. P. Karampelas, Lect. Notes Soc. Netw. **3**, 127–136 (2013)
113. P.D. Hoff, A.E. Raftery, M.S. Handcock, *Latent Space Approaches to Social Network Analysis*. Technical Report no. 399, University of Washington, Seattle (2001)
114. Y. Yamakawa, R. Kanai, M. Matsumura, E. Naito, PLOS One **4**(2), e4360 (2009)
115. P. Expert, T.S. Evans, V.D. Blondel, R. Lambiotte, Proc. Natl. Acad. Sci. **108**(19), 7663–7668 (2011)
116. M. Szell, R. Lambiotte, S. Thurner, Proc. Natl. Acad. Sci. **107**(31), 13636–13641 (2010)
117. L.C. Freeman, Am. J. Sociol. **98**, 152–66 (1992); J.E. Holly, Am. Math. Mon. 721–728 (2001)
118. H.S. Wio, C. Escudero, J.A. Revelli, R.R. Deza, M.S. de la Lama, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **369**(1935), 396–411 (2011)
119. S. Bleher, C. Grebogi, E. Ott, R. Brown, Phys. Rev. Lett. **38**(2), 930–938 (1988)
120. C.P. Roca, J.A. Cuesta, A. Sánchez, Phys. Rev. E **80**(4), 046106 (2009); N. Bellomo, H. Berestycki, F. Brezzi, J.P. Nadal, Math. Models Methods Appl. Sci. **20**(Suppl.), 1391–1395 (2010)
121. A. Ludu, *Nonlinear Waves and Solitons on Contours and Closed Surfaces* (Springer, Berlin, 2012)
122. R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, Data Min. Knowl. Disc. **11**(1), 5–33 (2005)
123. K.A. Arwini, C.T.J. Dodson, *Information Geometry: Near Randomness and Near Independence* (Springer, Berlin, 2008)
124. A. Landherr, B. Friedl, J. Heidemann, Bus. Inf. Syst. Eng. **6**, 371–385 (2010)
125. R. Cooper, *The Breaking of Nations: Order and Chaos in the Twenty-First Century* (Atlantic Monthly Press, New York, 2004)
126. J.P. Van Bendegem, Real-life mathematics versus ideal mathematics: the ugly truth, in *Empirical Logic and Public Debate*, vol. 3, ed. by E.C.W. Krabbe, R.J. Dalitz, P.A. Smit (Rodopi, Amsterdam, 1993)
127. R. Hersh, *What Is Mathematics, Really?* (Oxford University Press, Oxford, 1997)
128. R. Iemhoff, Intuitionism in the philosophy of mathematics, in *The Stanford Encyclopedia of Philosophy* (2008), http://plato.stanford.edu/entries/intuitionism/
129. L. Lovász, Discrete and continuous: two sides of the same? in *Visions in Mathematics* (Birkhäuser, Basel, 2010)
130. Y.I. Manin, *Mathematics and Physics* (Birkhäuser, Basel, 1981)
131. C. Nash, S. Sen, *Topology and Geometry for Physicists* (Academic, San Diego, 1983)
132. L. Schwartz, *Analyse*, 2nd edn. (Hermann, Paris, 1970)
133. A. Kolmogorov, S. Fomine, *Eléments de la théorie des fonctions et de l'analyse fonctionelle* (Editions MIR, Moscow, 1974)
134. J. Dieudonné, *Fondements de l'analyse moderne*, vol. 1 (Gauthier-Villars, Paris, 1965)
135. M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions*, vol. 55 (National Bureau of Standards Applied Mathematics Series, Washington, DC, 1964). Available online at http://www.math.sfu.ca/~cbm/aands/intro.htm
136. M.P. do Carmo, *Differential Geometry of Curves and Surfaces* (Prentice-Hall, Englewood Cliffs, 1976)

137. M. Spivak, *A Comprehensive Introduction to Differential Geometry* (Publish or Perish Inc., Boston, 1970)
138. S. Kobayashi, K. Nomizu, *Foundations of Differential Geometry* (Interscience Publishers, London, 1963)
139. C.E. Weatherburn, *Differential Geometry of Three Dimensions* (Cambridge University Press, Cambridge, 1927)
140. N. Steenrod, *The Topology of Fibre Bundles* (Princeton University Press, Princeton, 1951)
141. R. Courant, D. Hilbert, *Methods of Mathematical Physics* (Interscience Publishers, New York, 1966)
142. R. Abraham, J.E. Marsden, *Foundations of Mechanics* (W.A. Benjamin, New York, 1967)
143. A.J. Chorin, J.E. Marsden, *A Mathematical Introduction to Fluid Mechanics* (Springer, New York, 1992)
144. H. Lamb, *Hydrodynamics* (Dover, New York, 1932)
145. R. Aris, *Vectors, Tensors, and the Basic Equations of Fluid Mechanics* (Dover, New York, 1989)
146. V.I. Arnold, B.A. Khesin, *Topological Methods in Hydrodynamics* (Springer, New York, 1998)
147. A. Friedman, Free boundary problems in science and technology. Not. Am. Math. Soc. **47**(8), 854–861 (2000)
148. G.N. Nicolis, *Foundations of Complex Systems* (World Scientific, New Jersey, 2007)
149. P. Alexandroff, *Elementary Concepts of Topology* (Dover, New York, 1961)
150. R. Uijlenhoet, M. Steiner, J.A. Smith, J. Hydrometeorol. **4**(1), 43–61 (2003)
151. A.A. Kosinski, *Differential Manifolds* (Dover, Mineola, 1993)
152. J.W. Milnor, *Topology from the Differential Viewpoint* (Princeton University Press, Princeton, 1997)
153. D. Lovelock, H. Rund, *Tensors, Differential Forms, and Variational Principles* (Dover, New York, 1989)
154. S.G. Rajeev, *Geometry of the Motion of Ideal Fluids and Rigid Bodies*. Preprint `arXiv:0906.0184` (2009)
155. A. Ludu, J. Nonlinear Math. Phys. **15**(2), 151–170 (2008)
156. H. Luo, T.R. Bewley, J. Comput. Phys. **199**, 353–375 (2004)
157. L.E. Payne, G. Pólya, H.F. Weinberger, J. Math. Phys. **35**, 289–298 (1956); M.S. Ashbaugh, Proc. Indian Acad. Sci. Math. Sci. **112**, 3–30 (2002)
158. I. Chavel, *Eigenvalues in Riemannian Geometry* (Academic, New York, 1984)
159. M.S. Ashbaugh, Isoperimetric and universal inequalities for eigenvalues, in *Spectral Theory and Geometry*, ed. by E.B. Davies, Yu. Safarov. London Mathematical Society Lecture Series, vol. 273 (Cambridge University Press, Cambridge, 1999), pp. 95–139
160. M.A. Khabou, L. Hermi, M.B.H. Rhouma, Pattern Recogn. **40**(1), 141–153 (2007)
161. M. Kac, Am. Math. Mon. **73**, 1–23 (1966)
162. D. Raviv, R. Kimmel, Int. J. Comput. Vis. **111**(1), 1–11 (2014); F. Dornaika, A. Assoum, Y. Ruichek, Graph optimized Laplacian eigenmaps for face recognition, in *Proceedings of the SPIE 9406*. Intelligent Robots and Computer Vision XXXII: Algorithms and Techniques, 94060E (2015)
163. H.C.M. Smith, C. Pearce, D.L. Millar, Renew. Energy **40**(1), 51–64 (2012)
164. D. Henry, *Perturbation of the Boundary in Boundary-Value Problems of Partial Differential Equations*. London Mathematical Society Lecture Notes, vol. 318 (Cambridge University Press, Cambridge, 2005)
165. R. Courant, D. Hilbert, *Methods of Mathematical Physics* (Interscience, London, 1953); G. Polya, G. Szëgo, *Isoperimetric Inequalities in Mathematical Physics*. Annals of Mathematics Studies, vol. 27 (Princeton University Press, Princeton, 1951)
166. A. Saeid, B. Azarnavid, M.S. Alhuthali, J. Comput. Appl. Math. **279**, 293–305 (2015)
167. D. Lewis, J. Marsden, R. Montgomery, T. Ratiu, Phys. D **18**, 391–404 (1986)
168. T.J. Bridges, Math. Proc. Camb. Philos. Soc. **121**, 147–190 (1997)
169. J.E. Marsden, S. Pekarsky, S. Shkoller, M. West, J. Geom. Phys. **38**(3–4), 253–284 (2001)

170. M. Reuter, F.-E. Wolter, M. Shenton, M. Niethammer, Comput. Aided Des. **41**(10), 739–755 (2009)
171. M. Mortara, M. Spagnuolo, Comput. Graph. **25**, 1 (2001)
172. R. Diestel, *Graph Theory* (Springer, Heidelberg, 2012)
173. A.E. Brouwer, W.H. Haemers, *Spectra of Graphics* (Springer, New York, 2010)
174. D. Cvetković, P. Rowlinson, S. Simić, *An Introduction to the Theory of Graph Spectra* (Cambridge University Press, Cambridge, 2010)
175. F.R.K. Chung, *Spectral Graph Theory* (American Mathematical Society, Providence, 1997)
176. P.V. Mieghem, *Graph Spectra for Complex Networks* (Cambridge University Press, Cambridge, 2011)
177. R. Albert, H. Jeong, A.-L. Barabási, Nature **401**, 130–131 (1999)
178. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalaon, R. Stata, A. Tomkins, J. Wiener, Comput. Netw. **33**(1–6), 309–320 (2000)
179. W.N. Anderson, T.D. Morley, Linear Multilinear Algebra **18**, 141–145 (1985)
180. B. Bollobás, O. Riordan, Combinatorica **24**(1), 5–34 (2004)
181. B. Bollobás, Trans. Am. Math. Soc. **267**, 41–52 (1981)
182. Ju.D. Burtin, Dokl. Akad. Nauk SSSR **209**, 765–768 (1973); I. Teor. Verojatnost. i Primenen. **19**, 740–754 (1974)
183. B. Bollobás, W.F. de la Vega, Combinatorica **2**, 125–134 (1982)
184. V. Klee, D.G. Larman, Can. J. Math. **33**, 618–640 (1981); B. Bollobás, V. Klee, Combinatorica **4**, 7–19 (1984); B. Bollobás, F.R.K. Chung, SIAM J. Discrete Math. **1**, 328–333 (1988)
185. D.J. Watts, S.H. Strogatz, Nature **393**, 440–442 (1998); M.E.J. Newman, S.H. Strogatz, D.J. Watts, Phys. Rev. E **64**, 026118 (2001)
186. G. Besson, B. Colbois, G. Courtois, Trans. Am. Math. Soc. **350**(1), 331–345 (1998); B. Helffer, M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, M.P. Owen, Commun. Math. Phys. **202**(3), 629–649 (1999); Y. Colin de Verdierè, J. Comb. Theory. Ser. B **74**(2), 121–146 (1998)
187. H.V.D. Holst, Electron. J. Linear Algebra **20**, 574–585 (2010)
188. S. Hoory, N. Linial, A. Wigderson, Bull. Am. Math. Soc. **43**(4), 439–561 (2006)
189. J.F. Davis, P. Kirk, *Lecture Notes in Algebraic Topology*, vol. 35 (American Mathematical Society, Providence, 2001)
190. S.K. Donaldson, P.B. Kronheimer, *The Geometry of Four-Manifolds*. Oxford Mathematical Monographs (Clarendon Press, Oxford, 1997)
191. J. Milnor, Differential topology forty-six years later. Not. Am. Math. Soc. **58**(6), 804–809 (2011)
192. Y. Matsumoto, *An Introduction to Morse Theory* [Translations of Mathematical Monographs] (American Mathematical Society, Providence, 2002)
193. J.H.C. Whitehead, Ann. Math. **41**(4), 809–824 (1940); **73**(1), 154–212 (1961)
194. J.W. Milnor, *Collected Works*, vol. III. Differential Topology (American Mathematical Society, Providence, 2007)
195. H. Whitney, *Geometric Integration Theory* (Princeton University Press, Princeton, 1957)
196. J. Dieudonné, *A History of Algebraic and Differential Topology* (Birkhäuser, Boston, 1989)
197. R.C. Kirby, L.C. Siebenmann, *Foundational Essays on Topological Manifolds, Smoothings, and Triangulations* (Princeton University Press, Princeton, 1977)
198. A. Hatcher, *Algebraic Topology* (Cambridge University Press, Cambridge, 2002)
199. Ph. Blanchard, J.R. Dawin, D. Volchenkov, Eur. Phys. J. Spec. Top. **184**, 1–82 (2010)
200. N. Cartwright, *The Dappled World: A Study of the Boundaries of Science* (Cambridge University Press, Cambridge, 1999)
201. D. Lehoux, Spontaneous Gener. J. Hist. Philos. Sci. **3**(1), 28–34 (2009)
202. S. Gil-Riaño, V. Hamilton, Spontaneous Gener. J. Hist. Philos. Sci. **3**(1), 1–8 (2009)
203. J. Harambam, S. Aupers, Publ. Underst. Sci. **24**(4), 466–480 (2014)
204. A.P. Spee, P. Jarzabkowski, Strateg. Organ. **7**(2), 223–232 (2009)
205. P.T. Clemson, A. Stefanovska, Phys. Rep. **542**(4), 297–368 (2014)

206. R.M. Sainsbury, *Departing from Frege: Essays in the Philosophy of Language* (Routledge, London, 2003)
207. R. Keefe, P. Smith (eds.), *Vagueness: A Reader* (MIT Press, Cambridge, 1996), pp. 251–264
208. R.M. Sainsbury, *Concepts Without Boundaries, from Vagueness: A Reader*, ed. by R. Keefe, P. Smith (MIT Press, Cambridge, 1996), pp. 251–264 [Inaugural Lecture Delivered at King's College, London, 1990]
209. M. Callon, An essay on framing and overflowing: economic externalities revisited by sociology, in *The Laws of the Markets*, ed. by M. Callon (Blackwell, Oxford, 1998)
210. G. Caldarelli, R. Marchetti, L. Pietronero, Europhys. Lett. **52**(4), 386–391 (2000)
211. G. Tanaka, X. Hou, X. Ma, G.D. Edgecombe, N.J. Strausfeld, Nature **502**, 364–367 (2013)
212. S.J. Mojzsis, C.D. Coath, J.P. Greenwood, K.D. McKeegan, T.M. Harrison, Geochimica et Cosmochimica Acta **67**(9), 1635–1658 (2003)
213. J. Kleinberg, The small-world phenomenon: an algorithmic perspective, in *Proceedings of the 32nd ACM Symposium on the Theory of Computing* (2000)
214. A. Ganesh, F. Xue, *On the connectivity and diameter of small-world networks*, MSR Technical Report (2007)
215. D.J. Watts, *Small Worlds* (Princeton University Press, Princeton, 1999)
216. D.R. White, M. Houseman, Complexity **8**(1), 72–81 (2002)
217. N. Mathias, V. Gopal, Phys. Rev. E **63**, 021117 (2011)
218. O. Sporns, C.J. Honey, Proc. Natl. Acad. Sci. **103**(51), 19219–19220 (2006); D.S. Bassett, A. Meyer-Lindenberg, S. Achard, T. Duke, E. Bullmore, Proc. Natl. Acad. Sci. **103**, 19518–19523 (2006)
219. S. Wasserman, K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994)
220. J. Bruggeman, *Social Networks. An Introduction* (Routledge, Abingdon, 2008)
221. E.W. Zegura, K.L. Calvert, M.J. Donahoo, IEEE/ACM Trans. Netw. **5**(6), 770–783 (1997)
222. D. Liu, D. Antolos, A. Ludu, Burglary crime analysis using logistic regression, in *56th Human Factors and Ergonomics Society Annual Meeting*, Westin, Boston, 22–26 October 2012
223. J. Ugander, B. Karrer, L. Backstrom, and C. Marlow: *The Anatomy of the Facebook Social Graph*, `arXiv:1111.4503` (2011)
224. P. Erdös, A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17–61 (1960)
225. E.K. Çetinkaya, M.J.F. Alenazi, J.P. Rohrer, J.P.G. Sterbenz, Topology connectivity analysis of Internet infrastructure using graph spectra, in *4th International Workshop on Relaible Design and Modeling in ICUMT* (IEEE, New York, 2012), pp. 752–758
226. S.-H. Yook, H. Jeong, A.-L. Barabási, Proc. Natl. Acad. Sci. **99**(21), 13382–13386 (2002)
227. S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, Proc. Natl. Acad. Sci. **104**(27), 11150–11154 (2007)
228. A. Keltenbrunner, S. Scellato, Y. Volkovich, D. Laniado, D. Currie, E.J. Jutemar, C. Mascolo, *WOSN'12*, Helsinki (2012)
229. Y. Hu, Y. Wang, Z. Di, `arXiv:0802.0047v2 (physics.soc-ph)` (2013); Z. Yuan, C. Zhao, Z. Di, W.-X. Wang, Y.-C. Lai, Nat. Commun. **4**, 2447 (2013)
230. A.L. Barabási, R. Albert, H. Jeong, Phys. A **281**, 69–77 (2000)
231. A.D.I. Kramer, J.E. Guillory, J.T. Hancock, Proc. Natl. Acad. Sci. **111**(24), 8788–8790 (2014)
232. N. Ellison, C. Steinfield, C. Lampe, Spatially bounded online social networks and social capital: the role of Facebook, in *Annual Conference of the International Communication Association*, Dresden (2006)
233. R. Conte et al., Eur. Phys. J. Spec. Top. **214**, 325–346 (2012)
234. V. Marx, Nature **498**, 255–260 (2013)
235. D. Bollier, *The Promise and Peril of Big Data* (Aspen Institute, Washington, DC, 2010)
236. F. Camastra, Pattern Recogn. **36**(12), 2945–2954 (2003)
237. R. Ghrist, Bull. Am. Math. Soc. **45**(1), 61–75 (2008)
238. G. Carlsson, Bull. Am. Math. Soc. **46**(2), 255–308 (2009)
239. V. Robins, Computational topology for point data: Betti numbers of α-shapes, in *Morphology of Condensed Matter* (Springer, Berlin, 2002), pp. 261–274

240. P. Niyogi, S. Smale, S. Weinberger, Discrete Comput. Geom. **39**(1–3), MR2383768 (2008)
241. E. Carlsson, G. Carlsson, V. de Silva, Int. J. Comput. Geom. Appl. **16**(4), 291–314 (2006)
242. M.J. Egenhofer, E. Clementini, P. Di Felice, Int. J. Geogr. Inf. Syst. **8**(2), 129–144 (1994)
243. O.R. Enriquez, I.R. Peters, S. Gekle, J. Fluid Mech. **701**, 40–58 (2012)
244. G.R. Lawlor, J. Differ. Equ. **24**(1), 190–204 (2014); J. Harrison, J. Geom. Anal. **14**(1), 47–61 (2004)
245. V.A. Lubarda, Acta Mech. **224**(7), 1365–1382 (2013)
246. N. Takagaki, S. Komori, Int. J. Multiph. Flow **60**, 30–39 (2014)
247. L. Shanhong, G. Guanqing, L. Caiting, Tangqi, Commun. Comput. Inf. Sci. **405**, 265–275 (2014)
248. G. Liang, Y. Guo, S. Shen, Exp. Thermal Fluid Sci. **53**, 244–250 (2014)
249. Z. Wang, Y. Nagao, Electrochim. Acta **129**, 343–347 (2014)
250. M.B. Said, M. Selzer, B. Nestler, D. Braun, C. Greiner, H. Garcke, Langmuir **30**(14), 4033–4039 (2014)
251. M. Jalaal, K. Mehravaran, Phys. Fluids **26**, 012101 (2014)
252. D.R. Richardson, A.J. Brinker, M.D. Polanka, A. Lynch, D.L. Blunck, *AIAA SciTech 52nd Aerospace Sciences Meeting* (National Harbor, Maryland, 2014)
253. F. Peña-Polo, A. Blanco, L. Di G. Sigalotti, *Environmental Science and Engineering* (Springer, Switzerland, 2014), pp. 307–314
254. C. Song, J.K. Moon, K. Lee, K. Kim, H.K. Pak, Soft Matter **10**, 2679–2684 (2014)
255. K. Ueno, S. Hamasaki, E.J. Wanless, Y. Nakamura, S. Fujii, Langmuir **30**(11), 3051–3059 (2014)
256. P. Wägli, Y.-C. Chang, K. Hans, A. Homsy, L. Hvozdara, H.P. Herzig, M. Sigrist, N.F. de Rooij, Anal. Chem. **85**(3), 1924–1924 (2014)
257. K. Waldrona, W.D. Wua, Z. Wua, W. Liua, C. Selomulyaa, D. Zhaoa, X.D. Chen, J. Colloid Interface Sci. **48**, 225–233 (2014)
258. G. Stoitcheva, A. Ludu, J.P. Draayer, Math. Comput. Simul. **55**(4–6), 621–625 (2001)
259. A. Ludu, A. Sandulescu, W. Greiner, K.M. Kallman, M. Brenner, T. Lonnroth, P. Manngard, J. Phys. G Nucl. Part. Phys. **21**(6), L41 (1995)
260. V. Makhalov, K. Martiyanov, A. Turlapov, Phys. Rev. Lett. **112**, 045301 (2014)
261. L.F. Palhares, E.S. Fraga, Phys. Rev. D **82**, 125018 (2010); J.J. Bjerrum-Bohr, I.N. Mishustin, T. Døssing, Nucl. Phys. A **882**, 90–106 (2012); C. Herold, M. Nahrgang, I. Mishustin, M. Bleicher, Nucl. Phys. A **925**, 14–24 (2014)
262. J.I. Kapusta, A.P. Vischer, R. Venugopalan, Phys. Rev. C **51**(2), 901–910 (1995)
263. G. Lu, H. Hu, Y. Duan, Y. Sun, Appl. Phys. Lett. **103**(25), 253104 (2013)
264. S.-Y. Teh, R. Lin, L.-H. Hung, A.P. Lee, Lab Chip **8**, 198–220 (2008)
265. A.P. Berke, L. Turner, H.C. Berg, E. Lauga, Phys. Rev. Lett. **101**, 038102 (2008)
266. M.A. Fernández, C. Albor, M. Ingelmo-Torres, S.J. Nixon, C. Ferguson, T. Kurzchalia, F. Tebar, C. Enrich, R.G. Parton, A. Pol, Science **313**(5793), 1628–1632 (2006)
267. G.P. Ayers, T.V. Larson, J. Atmos. Chem. **11**(1–2), 143–167 (1990)
268. S. Abadie et al., *National Tsunami Hazard Mitigation Program*, NOAA Special Report (2011)
269. P. Möller, J.R. Nix, K.-L. Kratz, At. Data Nucl. Data Tables **66**(2), 131–343 (1997)
270. H. Heiselberg, C.J. Pethick, E.F. Staubo, Phys. Rev. Lett. **70**, 1355 (1993)
271. P.A. Hwang, W.J. Teague, J. Atmos. Oceanic. Technol. **17**, 847–853 (2000)
272. M. Hutchings, F. Morgan, M. Ritoré, A. Ros, Ann. Math. **155**, 459–489 (2002)
273. A. Tufaile, J.C. Sartorelli, Phys. Rev. E **66**, 056204 (2002)
274. M.C. Álvarez, J. Corneli, G. Walsh, S. Beheshti, Exp. Math. **12**(1), 79–89 (2003)
275. J. Corneli, I. Corwin, S. Hurder, V. Sesum, Y. Xu, E. Adams, D. Davis, M. Lee, R. Visocchi, N. Hoffman, Houst. J. Math. **34**(1), 181–204 (2008)
276. B. Scheid, S. Dorbolo, L. Arriga, E. Rio, Phys. Rev. Lett. **109**, 264502 (2012)
277. J. Zou, C. Ji, BG. Yuen, X.D. Ruan, X. Fu, Phys. Rev. E **87**, 061002(R) (2013)
278. X. Ren, J. Wei, *A double bubble assembly as a new phase of a ternary inhibitory system*. Preprint (2014)
279. R. Salmon, Ann. Rev. Fluid Mech. **20**, 225–256 (1988)

280. M. Holst, E. Lunasin, G. Tsogtgerel, J. Nonlinear Sci. **20**(5), 523–567 (2010); Preprint arXiv:0901.4412 (2009)
281. T.L. Story, *Introduction to Differential Geometry with Applications to Navier–Stokes Dynamics* (iUniverse, Inc., Bloomington, 2005)
282. J.B. Etnyre, R. Ghrist, Trans. Am. Math. Soc. **352**(12), 5781–5794 (2000)
283. J.B. Etnyre, K. Honda, Ann. Math. **153**, 749–766 (2001); J.B. Etnyre, *Introductory Lectures on Contact Geometry*. Preprint arXiv:math/0111118 (2002)
284. K. Modin, M. Perlmutter, S. Marsland, R. McLachlan, Res. Lett. Inf. Math. Sci. **14**, 79–106 (2010)
285. D.G. Ebin, J. Marsden, Ann. Math. **92**(1), 102–163 (1970)
286. B. Khesin, Not. Am. Math. Soc. **52**, 9–19 (2005)
287. R. Ghrist, On the contact topology and geometry of ideal fluids, in *Handbook of Mathematical Fluid Dynamics*, vol. IV, ed. by S. Friedlander, D. Serre (Elsevier, Amsterdam, 2007), pp. 1–38
288. E. Becker, W.J. Hiller, T.A. Kowalewski, J. Fluid Mech. **258**, 191–216 (1994)
289. U. Brosa, Z. Naturforsch. A **43**, 1141 (1986)
290. J.A.F. Plateau, *Annual Report of the Board of Regents of the Smithsionian Institution*, Washington, DC (1863), pp. 270–285; H. Poincaré, Acta Math. **7**, 259–302 (1885); Lord Rayleigh Philos. Mag. **28**, 161 (1914)
291. J. Harrison, J. Geom. Anal. **24**(1), 271–297 (2014)
292. V.I. Arnold, Ann. Inst. Fourier (Grenoble) **16**(1), 319–361 (1966); V.I Arnold, B.A. Khesin, Ann. Rev. Fluid Mech. **24**, 145–166 (1992)
293. V.E. Zacharova, E.A. Kuznetsov, Phys. Usp. **40**(11), 1087–1116 (1997)
294. D. Lewis, J. Marsden, T. Ratiu, J. Math. Phys. **28**(10), 2508–2515 (1987)
295. P.J. Morrison, Rev. Mod. Phys. **70**(2), 467–521 (1998)
296. A.M. Vinogradov, B.A. Kupershmidt, Russ. Math. Surv. **32**(4), 245 (1977)
297. S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos* (Springer, New York, 2003)
298. R. Tagg, L. Cammack, A. Croonquist, T.G. Wang, *Rotating liquid drops: Plateau's experiment revisited*, JPL Report 80–66 (1980)
299. C.-J. Heine, IMA J. Numer. Anal. **26**(4), 723–751 (2006)
300. N. Kapouleas, Commun. Math. Phys. **129**, 139–159 (1990)
301. S. Chandrasekhar, Proc. R. Soc. Lond. A Math. Phys. Sci. **286**(1404), 1–26 (1965)
302. D.K. Ross, Aust. J. Phys. **21**, 823–835 (1968)
303. R.A. Brown, L.E. Scriven, Proc. R. Soc. Lond. A Math. Phys. Sci. **371**(1746), 331–357 (1980)
304. L.H. Ungar, R.A. Brown, Philos. Trans. R. Soc. Lond. A **306**(1493), 347–370 (1982)
305. V. Cardoso, Physics **1**, 38 (2008)
306. R.J.A. Hill, L. Eaves, Phys. Rev. Lett. **101**, 234501 (2008)
307. D. Chae, J. Differ. Equ. **249**, 571–577 (2010)
308. D.B. Khismatullin, Y. Renardy, Phys. Fluids **15**(5), 1351–1354 (2003)
309. J. Ratzkin, *An end-to-end gluing construction for surfaces of constant mean curvature*. Ph.D. Thesis, University of Washington, 2001
310. S.-Y. Teh, R. Lin, L.-H. Hung, A.P. Lee, Lab Chip **8**, 198–220 (2008)
311. W. Bouwhuis, K.G. Winkels, I.R. Peters, P. Brunet, D. van der Meer, J.H. Snoeijer, Phys. Rev. E **88**, 023017 (2013)
312. J.G. Leidenfrost, *De Aquae Communis Nonnullis Qualitatibus Tractatus*, Duisburg (1756)
313. D. Quéré, Annu. Rev. Fluid Mech. **45**, 197–215 (2013)
314. D. Mampallil, H.B. Eral, A. Staicu, F. Mugele, D. van den Ende, Phys. Rev. E **88**, 053015 (2013)
315. A.-L. Biance, C. Clanet, D. Quéré, Phys. Fluids **15**(6), 1632–1637 (2003)
316. H. Linke, B.J. Alemán, L.D. Melling, M.J. Taormina, M.J. Francis, C.C. Dow-Hygelund, V. Narayanan, R.P. Taylor, A. Stout, Phys. Rev. Lett. **96**(15), 154502 (2006)
317. A. Snezhko, E.B. Jacob, I.S. Aranson, New J. Phys. **10**, 043034 (2008)

318. W. Bouwhuis, K.G. Winkels, I.R. Peters, P. Brunet, D. van der Meer, J.H. Snoeijer, Phys. Rev. E **88**, 023017 (2013)
319. Y. Pomeau, M. Le Berre, F. Celestini, T. Frisch, C. R. Méc. **340**(11), 867–881 (2013); F. Celestini, T. Frisch, A. Cohen, C. Raufaste, L. Duchemin, Y. Pomeau, Phys. Fluids **26**, 032103 (2014)
320. T.R.N. Jansson, M.P. Haspang, K.H. Jensen, P. Hersen, T. Bohr, Phys. Rev. Lett. **96**, 174502 (2006)
321. R. Bergmann, L. Tophøj, A.M. Homan, P. Hersen, A. Andersen, T. Bohr, J. Fluid Mech. **679**, 415–431 (2011)
322. H.A. Abderrahmane, K. Siddiqui, G.H. Vatistas, Exp. Fluid **50**, 677–688 (2011)
323. Y. He, D. Mihalache, B.A. Malomed, Y. Qiu, Z. Chen, Y. Li, *Generations of polygonal soliton clusters and fundamental solitons by radially-azimuthally phase-modulated necklace-ring beams in dissipative systems*. Preprint `arXiv:1206.1900` (2012)
324. Y. He, B. Mihalache, B.A. Malomed, Y. Qiu, Z. Chen, Y. Li, Generations of polygonal soliton clusters and fundamental solitons by radially-azimuthally phase-modulated necklace-ring beams in dissipative systems. Phys. Rev. E **85**, 066206 (2012)
325. L.M. Satarov, M.N. Dmitriev, I.N. Mishustin, *Equation of state of hadron resonance gas and the phase diagram of strongly interacting matter*. `arXiv:0901.1430` (2009)
326. K. Iga, S. Yokota, S. Watanabe, T. Ikeda, H. Niino, N. Misawa, Fluid Dyn. Res. **46**, 031409 (2014)
327. G.H. Vatistas, J. Fluid Mech. **217**, 241–248 (1990)
328. G.H. Vatistas, H.A. Abderrahmane, M.H.K. Siddiqui, Phys. Rev. Lett. **100**, 174503 (2008)
329. M. Amaouche, H.A. Abderrahmane, G.H. Vatistas, Eur. J. Mech. B Fluids **41**, 133–137 (2013)
330. Y. Tasaka, M. Iima, J. Fluid Mech. **636**, 475–484 (2009)
331. H.A. Abderrahmane, M. Fayed, H.D. Ng, G.H. Vatistas, J. Fluid Mech. **724**, 695–703 (2013)
332. J. Mougel, D. Fabre, L. Lacaze, Waves and instabilities in rotating free surface flows, in *21ième Congrés Francais de Mécanique*, CFM (2013)
333. T. Suzuki, M. Iima, Y. Hayase, Phys. Fluids **18**, 101701 (2006)
334. J.M. Lopez, F. Marques, A.H. Hirsa, R. Miraghaie, J. Fluid Mech. **502**, 99–126 (2004)
335. E.A. Hendricks, W.A. Schubert, Y.-H. Chen, H.-C. Kuo, M.S. Peng, J. Atmos. Sci. **71**, 1623–1643 (2014)
336. E.A. Hendricks, B.D. McNoldy, W.H. Schubert, Mon. Weather Rev. **140**, 4066–4077 (2012)
337. E.A. Hendricks, W.H. Schubert, R.K. Taft, J. Atmos. Sci. **66**, 705–722 (2009)
338. W.H. Schubert, M.T. Montgomery, R.K. Taft, T.A. Guinn, S.R. Fulton, J.P. Kossin, J.P. Edwards, J. Atmos. Sci. **56**, 1197–1223 (1999)
339. J.P. Kossin, B.D. McNoldy, W.H. Schubert, Mon. Weather Rev. **130**, 3144–3149 (2002); J.P. Kossin, W.H. Schubert, *BAMS* (2004), pp. 151–153
340. A.C.B. Aguiar, P.L. Read, R.D. Wordsworth, T. Salter, Y.H. Yamasaki, Icarus **206**, 755–763 (2010)
341. D. Konstantinov, H.J. Maris, Phys. Rev. Lett. **90**(2), 025302 (2003); H.J. Maris, J. Low Temp. Phys. **132**(1–2), 77–95 (2003)
342. H. Noguchi, G. Gompper, Proc. Natl. Acad. Sci. **102**(40), 14159–14164 (2005)
343. S. Perrard, Y. Couder, E. Fort, L. Limat, Europhys. Lett. **100**, 5406 (2012)
344. I.R. Peters, D. van der Meer, J.M. Gordillo, J. Fluid. Mech. **724**, 553–580 (2013)
345. J.C. Burton, P.Y. Lu, S.R. Nagel, Phys. Rev. Lett. **111**(18), 188001 (2013)
346. N. Yoshinaga, Phys. Rev. E **89**(1), 012913 (2014)
347. S. Shrivastava, K.H. Kang, M.F. Schneider, Phys. Rev. E **91**(1), 012715A (2015); A.E. Hady, B.B. Machta, Nat. Commun. **6**, 6697 (2015)
348. A. Gonzalez-Perez, R. Budvytyte, L.D. Mosgaard, S. Nissen, T. Heimburg, Phys. Rev. X **4** 031047 (2014); A. Gonzalez-Perez, L.D. Mosgaard, R. Budvytyte, E. Villagran-Vargas, A.D. Jackson, T. Heimburg, Preprint arXiv:1502.07166 (2015)
349. T. Heimburg, A.D. Jackson, Biophys. Rev. Lett. **2**, 57–78 (2007); K. Graesboll, H. Sasse-Middelhoff, T. Heimburg, Biophys. J. **106**, 2143–2156 (2014)

350. S. Verpoort, *The Geometry of the Second Fundamental Form: Curvature Properties and Variational Aspects*. Ph.D. thesis, Katholieke Universiteit Leuven, 2008
351. J.M. Lee, *Manifolds and Differential Geometry* (American Mathematical Society, Providence, 2009)
352. J.M. Lee, *Differential and Physical Geometry* (notes posted online)
353. N. Wilkin-Smith, J. Math. Anal. Appl. **332**, 577–606 (2007)
354. S. Albano, E.H.A. Gonzales, Indiana Univ. Math. J. **32**(5), 687–702 (1983)
355. F. Schwabl, *Statistical Mechanics* (Springer, Berlin, 2000)
356. E. Braaten, J. Pearson, Phys. Rev. Lett. **82**(2), 255–258 (1999)
357. A. Bulgac, G. Bertsch, Phys. Rev. Lett. **94**, 070401 (2005)
358. B. Dasgupta, A. Dighe, A. Mirizzi, G. Raffelt, Phys. Rev. D **78**, 033014 (2008)
359. B. Sapoval, T. Gobron, A. Margolina, Phys. Rev. Lett. **67**(21), 2974–2977 (1991)
360. G. Deleuze, F. Guattari, *What Is Philosophy?* (Columbia University Press, New York, 1994)
361. J.N. Oppenheim, M.O. Magnasco, Phys. Rev. Lett. **110**, 044301 (2013)
362. Z. Zeravcic, M.P. Brenner, Proc. Natl. Acad. Sci. **111**(5), 1748–1753 (2014)
363. A.S. Balankin, B.E. Elizarraraz, Phys. Rev. E **85**, 056314 (2012)
364. K. Wynn, Math. Cogn. **1**(1), 35–60 (1995); P. Kitcher, *The Nature of Mathematical Knowledge* (Oxford University Press, Oxford, 1984); W. Fias, M.H. Fischer, Spatial representation of numbers, in *Handbook of Mathematical Cognition*, ed. by J.I.D. Campbell (Psychology Press, New York, 2005)
365. L. Uhrig, S. Dehaene, B. Jarraya, J. Neurosci. **34**(4), 1127–1132 (2014)
366. P.C. Bressloff, J.D. Cowan, M. Golubitsky, P.J. Thomas, M.C. Wiener, Philos. Trans. R. Soc. B Biol. Sci. **356**(1407), 299–330 (2001)
367. L. Richardson, Sniffing and smelling. Philos. Stud. **162**, 401 (2011)
368. Z. Reznikova, B. Ryabko, Behaviour **148**(4), 405–434 (2011)

Index

A. Ludu, *Boundaries of a Complex World*, Springer Series in Synergetics,
DOI 10.1007/978-3-662-49078-5