DISCOVERY *of* COSMIC FRACTALS

YURIJ BARYSHEV & PEKKA TEERIKORPI

With a foreword by Benoit Mandelbrot

DISCOVERY *of* COSMIC FRACTALS

DISCOVERY *of* COSMIC FRACTALS

YURIJ BARYSHEV
Astronomical Institute
St. Petersburg University
Russia

PEKKA TEERIKORPI
Tuorla Observatory
University of Turku
Finland

World Scientific
New Jersey • London • Singapore • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.
P O Box 128, Farrer Road, Singapore 912805
USA office: Suite 202, 1060 Main Street, River Edge, NJ 07661
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

DISCOVERY OF COSMIC FRACTALS

ISBN 981-02-4871-7
ISBN 981-02-4872-5 (pbk)

Printed by Fulsland Offset Printing (S) Pte Ltd, Singapore

To our families
whose love and patience have made this cosmic journey possible

Contents

Foreword by Benoit Mandelbrot

A contribution to the history of the conditional cosmological principle

The authors have asked me to present to the reader this most attractive and wide-ranging book of theirs. But this book hardly needs being presented! It is so well-informed of the history of its subject that it stands by itself and deserves to be appreciated from at least two viewpoints. Firstly, as the second word in the title suggests, it introduces to a very wide potential readership many facts and theories of cosmology. The style is precise but highly personal, a relief after too many introductions beholden to a stylistic formula. Secondly, as the remainder of the title suggests, this book has the unusual distinction of being a notable contribution to the history of ideas.

The authors also asked me to add to their work by describing the history of the fractal model of the cosmos, as I lived it during a long period when this model was not even criticized but simply dismissed. Specifically, they want to hear the story of how I came to improve upon the work of past giants by arguing that the cosmological principle should be restated in a conditional form. It is a pleasure to oblige.

The contrast between the homogeneous and fractal models of the universe will have to be resolved on its own terms. But conceptually this contrast is not isolated. Quite to the contrary: as I see it today, it is best understood and appreciated as a major facet of a long-standing, though also long-subdued, dialectic opposition between thoughts directed towards smoothness or roughness.

Instead of “roughness”, I used to speak or write of “irregularity”, but I now view that word as not only pedantic but clearly inappropriate. Indeed, it means “non-regularity” and somehow implies that regularity came first and roughness only later. In historical fact, the precise opposite was the case. Indeed, before the emergence of science and engineering, nearly-plane surfaces used to be exceptional in Man’s experience, a rare example being a quiet body of water and an even rarer one being a crystal. Circles were suggested by the full moon, a pebble’s effect on a flat body of water, or an eye’s pupil and iris. Spheres were suggested by some seeds.

Is it despite or because of their being exceptional, that those examples proved so extraordinarily attractive (or repellent?) to Mankind? Indeed, many cultures took idealized forms of smooth shapes as foundations of more or less developed but, in effect, universal forms of pre-classical geometry. Classical geometry followed when Euclid collected and organized all that was known and went far beyond by introducing the axiomatic method. That geometry has been developing ever since. It is not only the foundation of the overwhelming bulk of the study of nature, but also of the study of many aspects of culture – a short word I like to use to denote all of Man’s works. The old technology – as exemplified by highways, tables, and knives – had no choice but to tolerate roughness and only then as second best, behind an ideal represented by perfect flatness or smoothness.

For roughness, to the contrary, no comparable theoretical developments can be cited. No notion of “perfect roughness” was defined and made the focus of systematic study.

Does the preceding thumbnail history imply that, among shapes that matter to humanity, the development of technology and science witnessed a thorough “victory” of the flat over the rough? In the 1960s, there was no explicit awareness of past battles between rough “natives” and a smooth “conqueror”. That is, the word “victory” would not have come to mind. Quite appropriately, everyone identified geometry, science, and technology as centered on the flat or smooth, allowing as exceptions the smooth endowed with a few odd corners, or the smooth perturbed by the downy skin of a peach.

However, a development that was destined to become the seed of a major exception was planted around 1900. Within the official history of ideas I learned as a student, this seed-to-be consisted in esoterica that everyone called “mathematical monsters”. They were described as having no past and no precedent, as having been "invented" wholly armed from the brow of Jupiter and specifically intended to have no conceivable use in the sciences. Considerate teachers kept the monsters away from impressionable young minds.

Many of the original monsters happen to be self-similar, that is, made of parts deduced from the whole by a linear reduction. But the development of the mathematical esoterica immediately generalized over this property. Generality is praised by mathematicians for its own sake. Therefore, in order to use those esoterica, the first thing I had to do was to reestablish self-similarity and to lean on it heavily as a principle of invariance. But I am getting ahead of the story.

What follows is necessarily autobiographical. Due to an education which events perturbed to an extreme degree, I combined three features that seemed around 1950 to be mutually exclusive: a close acquaintance with the monsters in question, fluency in probabilistic esoterica, and (more surprisingly) a passionate wish to find some regularity in parts of both nature and culture that science had not previously touched. Those parts of culture go beyond highways, tables and knives and include the financial markets and other large but mostly uncontrolled designs such as the internet. They exemplify a high level of perceived "messiness" that I hoped to tame into mere complexity. Early on, I began to combine freely all those high and low caste concerns together, and in due time I conceived around them a new geometry that I had the privilege to name. In 1975, I coined for it the term, "fractal".

What is fractal geometry and what do I hope for it? Down to earth, it is the first organized step towards something that did not exist: a theory of roughness that could, to some extent, complement the great and diversified theory of smoothness.

How does this ambitious program concern clustering, therefore affect cosmology? The path that led me to fractal geometry began in a context altogether different from cosmology, but one that, in due time, made galaxy clusters come to mind unavoidably. The first step was taken in the early 1960s when I studied the clustering of errors in telephone channels and (metaphorically) found that what seemed like a small nut could only be opened by an intellectual sledge-hammer. That is, I had to devise tools that seemed unnecessarily powerful. Then, in the mid 1960s, chance reading made me turn to clustering of matter in the universe, and those unnecessarily powerful tools became handy and suggested the now-familiar "conditional" form of the classical "cosmological principle". Let me retell those events more slowly.

To describe the background, a model of galaxy clustering had been proposed in the 1950s by Jerzy Neyman and Elizabeth Scott. They postulated a compound Poisson process, constructing deliberately a randomized form of hierarchy. Since the Poisson process of constant density yields only a shadow of clustering, they took it as a first approximation and proceeded to improve it recursively, as follows. First, they injected clusters by allowing the Poisson density to vary according to a master process. Next, they injected superclusters by varying the master process density according to a supermaster process. The Neyman-Scott procedure could be extended as far as fancy wished, but it was a truly "Ptolemaic" throwback that had few admirers. Nearly everyone I respected dismissed it as an arbitrary exercise in curve-fitting. It was true that many desired features were present, but only for the reason that they had been very deliberately put in. That is, the model was far from being parsimonious. Nevertheless, Neyman enjoyed such great authority that in 1962 the engineers concerned with clusters of errors on telephone channels invoked the same Ptolemaic compound Poisson process.

The very different tack I took began with a bit of folklore. The engineers with whom I was working told me that, somehow, error clustering was the same at all scales. For example, subdivide a sample of duration T into equal subsamples and "mark" all the subsamples that include at least one error. Folklore asserted that the marked subsamples follow the same cluster pattern irrespectively of the value of T. This represented a property of invariance by reduction or dilation that, soon afterwards, I called self-similarity.

Ten years later, self-similarity also entered statistical physics as being a form of exact renormalizability. But – notwithstanding recently coined anecdotes to the contrary – statistical physics had no influence whatsoever on my scaling/fractal approach to clustering and conditional cosmographic principle.

Back to the story. I knew a restricted and limited form of clustering that is designed into a "monster" set defined in 1883 by Cantor (to run ahead of the story, the 1907 model of galactic clustering of Fournier d'Albe was – consciously or not – a three-dimensional Cantor set!) Unfortunately, practically every physicist had accepted the mathematicians' proclamations since 1900 and believed that the Cantor set could have no role in the modeling of nature. Moreover, there was no point of fighting those prejudices because it was true that the Cantor set could serve only if modified very deeply. Its hierarchy is relentless and deliberately "designed in"; its self-similarity applies only to certain values of T and n. Whether planned or not, it is another "Plotemaic" construction carried to the infinitely small and large: epicycles down to zero and "subcycles" forever. Worse, it has a privileged center, a lethal defect that can only be avoided by being replaced by an infinite sequence of increasingly "global" centers that asymptotically oscillates between plus and minus infinity.

My second good fortune was to know that self-similarity could be preserved and the lethal defects of the Cantor set avoided by replacing it by a certain random construction. I later called it a "Levy dust" and described the method of generating it as a one-directional "Levy flight" (to run ahead of the story again, a three-dimensional variant generates my fractal model of the "seeded universe" of galaxies.)

Moreover, I had discovered that the Levy dust is magnificently "creative". The construction itself involves a single parameter that is not a spatial scale but a fractal dimension. It is absolutely not Ptolemaic, not hierarchical. The true novelty of this work resided in two features.

Firstly, humans invariably perceive a sample of Levy dust as involving an infinite clustering hierarchy. To restate this, clustering and superclustering are not present due to being willfully injected, but because they follow necessarily from the basic scaling invariance, that is, from fractality.

Furthermore, the Levy dust has no privileged center. More precisely, every center Ω that belongs to the Levy dust L yields exactly the same statistical distribution for the other points in L. But a point Ω chosen at random is very different: it falls (with probability 1) into a great void; that is, one whose duration is (with probability 1) infinite. In probabilistic terms, my model required serious reexamination of essential concepts: it was not stationary, but only satisfies a weaker property I called conditional stationarity.

A brouhaha ensued. In my model the probability of errors is zero, but engineers know that the actual probability may be very small but is surely not zero. Hence, my model was bound to break down for large enough error-free time intervals, and results in a well-defined unconditioned distribution being available for every origin Ω. I was told that this made it unnecessary to condition Ω to belong to L. I agreed in principle but pointed out that, in practice, the said unconditional distribution is the product of two factors.

The first was a power-law with a solid empirical foundation. The second was a multiplier equal to the probability of an arbitrary Ω falling in L, which is at best a highly uncertain quantity. Therefore, it was best to avoid relying on the multiplier and observe that everything of interest reduced to the power-law factor.

By the mid 1960s, the above described structure of clustering on the line was essentially completed. At this point, a 1954 article by G. Gamow in "Scientific American" came to my attention and made me aware of Charlier's publication on the clustering of galaxies. Instantly, my mind extended all that preceeds from 1 to 3 dimensions, transposing it into new terms relative to cosmology. I read Charlier and followed up his credit to Fournier d'Albe, who had been thoroughly forgotten.

My work was an essential improvement on those predecessors. Firstly, clustering and the appearance of hierarchy were no longer specifically inserted but followed as necessary consequences of scale invariance, that is, fractality. Secondly, the sledge-hammer of fractality was no longer used to crack a nut but a major issue. Conditional stationarity instantly achieved a greatly increased "status" by being translated into a conditional form of the cosmological principle. Thirdly, the question of why the universe should be expected to be clustered was thoroughly modifed and became easier to handle. Once it has been transformed into the question of why it should be fractal, another bit of high mathematical esoterica came back to my mind, namely, the so-called "Frostman's potential theory" that relates fractals to attraction proportional to an inverse power of distance.

Having described how I avoided the deficiencies of the Fournier d'Albe–Charlier approach, I hasten to report how I experienced another great surprise that is of broad character, very different from this book's scientific focus but closely relevant to its historical form. Evidence started accumulating that the study of roughness (more pedantically, of extreme irregularity) was not starting with the proverbial empty table, "tabula rasa". No study interpretable in terms of roughness had been carried on to a technical level until I "tamed" the "monsters". However, relevant general thoughts were very much part of the historical record. More generally, my books "Les objets fractals" and "The Fractal Geometry of Nature" were widely read and commented upon, even before the former actually appeared and increasingly so after the publication of the latter. As a result, evidence came forth and continues to accumulate that the mainstream of mathematics, science, and engineering was the only context (if I dare use "only" for such a purpose) in which the flat overwhelmingly prevailed over the rough. The rough that was not relegated to mathematical esoterica had survived in many odd corners all over human experience. As usual, history was written by those who had prevailed, and my teachers' perception of the history of ideas was altogether biased.

This brings us back at long last to the book by Baryshev and Teerikorpi. It is original and personal (a form of praise under my pen) and worthy and useful on its own terms of cosmology. But it is also an important contribution to a much broader task triggered when I conceived and organized fractal geometry, used it in many fields, and only later realized that fractals had a past beyond 50 to 100 years of mathematical esoterica.

Searching and documenting all this past far exceeds any single individual's qualifications, but motivates several recent books. A book by Eglash showed that forms identified after the fact as fractal have long characterized African art and design. A forthcoming book by Jackson will show that proto-fractals also permeate religious iconography in Europe and many parts of Asia. A great contemporary composer, G. Ligeti, pointed out that music also turns out to have fractal aspects. Some old theories had clearly outlined them, but only dimly. Altogether, an intuitive understanding of fractality can be traced to the dawn of humanity.

Let us now return to the place organized sciences take among Man's eternal questions. Most are younger than art, but astronomy/cosmology is arguably an exception. Illustrating quite literally the concept of an endless frontier, it is also both one of the youngest and most active. As such, it is also (most unfortunately) one of the most oblivious of its rich past. It is therefore perfectly proper that the first systematic examination of the prehistory of fractals in a domain of science should concern this field.

Before concluding, one must comment on this admirable scholarship's final goal. Now that science is organized and books are plentiful, each scientist has many teachers and, if successful, many disciples. A subject's history is documented and can be followed step by step, in particular as it migrates from one favorite topic to another. But this book deals with times before science became organized, when scientists were few and isolated from one another in time and space, and interacted little. The situation that used to be the rule had a late but illustrative example in a topic of mathematics that started in 1872 with Weierstrass. (While it does not matter here, let it be said that it concerned "continuous but non-differentiable functions".)

Fifty years later, it was revealed that Weierstrass had been anticipated by fifty years by Bernard Bolzano. However, the latter did not publish and no evidence is known that he exerted an indirect influence.

To make my next point, recall that the French language distinguishes between "grande et petite histoire". In general, "petite histoire" is irrelevant gossip and the like. But in the history of ideas and of sciences it may help make a very important distinction. The "grande" history of that topic in mathematics did not begin until Weierstrass. The work of Bolzano only stars in "petty" history. It is illustrative and extraordinarily attractive, and I take pride in having moved it into wide public awareness. But a distinction must be drawn clearly between the streams of history which reach the "sea" and those which become lost in the sands. How does this distinction relate to this book? Beyond the "protofractal" thoughts of Kant, Charlier, and others already quoted, Baryshev and Teerikorpi have literally unearthed many other authors whose work remained isolated, undeveloped, and with little influence. They belong to the grand history of ideas, but only to the petty history of the fractal model.

The remarkable book I now have the privilege and pleasure of recommending to the reader is a joy to read and taught me a great deal. I developed a strong admiration for the authors' expository skills and I am wowed by the historical and geographical breadth of their scholarship. I wish them all the best.

Benoit Mandelbrot

Sterling Professor of Mathematical Sciences, Yale University, New Haven, USA

January 2002

Preface

Our book discusses issues at the heart of the modern science of the Universe. Astronomical discoveries of recent years have inspired us to relate how the understanding of the structures in the Universe has evolved, looking at this story from the middle of the events, as if caught in a storm.

Astronomy is an endless enterprise, held in motion by the promise of unexpected phenomena. The discovery of megafractals is such a surprise for cosmology. And new ideas spring from the fractals, questioning our basic beliefs about the cosmos. Writing this book we were delighted to find how history once and again confirms the golden rules: ***cast doubt on the obvious, expect the unexpected, keep watch for new paths.***

We divide our book to four parts to help the reader make the excursion from the birth of cosmological thinking in Part I, through the cosmological physics of the 20th century in Part II and encountering the deep enigmas of modern cosmology in Part III. Finally in Part IV the reader is taken aboard for advanced studies at the frontiers of research into the fractal architecture of the Universe.

We have endeavoured to explore in a simple manner the key questions in our science, to allow the reader to obtain a feeling for the events taking place at the frontline of cosmological research. In the footnotes and the Appendix the interested reader may find brief, more technical details and definitions. This book may also be regarded as illuminating our view on the philosophy of cosmology, complementing our forthcoming scientific monograph "Fundamental Questions of Practical Cosmology".

We would like to note that our citations come from texts available to us in different, not always original, languages. Sometimes we use existing translations, sometimes our own ones from the Russian, Finnish, German, French and Swedish. We will mostly use the unit system of Gauss, i.e. cm, g, sec, which is normal in astronomical literature. Also, when convenient, we use other units, e.g. give distances in astronomical units and parsecs. If the value of the Hubble constant is not mentioned, we usually adopt $H = 60$ km $\mathrm{s}^{-1}/\mathrm{Mpc}$. We also note that the Index contains the persons and the main concepts appearing in our story, and refers to the corresponding sections (for example, 10.5 means the section 5 in the chapter 10).

Acknowledgements

We are indebted to many astronomers and physicists, with whom we have had stimulating and pleasant contacts which have formed the necessary background for this book. Especially it is a pleasure to thank Chris Flynn for his careful reading of the manuscript and innumerable useful remarks and suggestions. We also would like to thank Luciano Pietronero and Francesco Sylos Labini, who greatly helped us in the preparation of the book.

We are very grateful to Benoit Mandelbrot for his warm support and helpful advice throughout the writing of this book.

Vitalij Gorbatskij, Georges Paturel, Laurent Nottale, Eric Lerner, Kondrad Rudnicki, Leena Tähtinen, Arthur Chernin, Andrej Berdyugin, Alexander Mylläri, Daniel Pfenniger, Françoise Combes, and Raimo Lehti have read parts of the manuscript and given many valuable suggestions. We thank Mikko Hanski and Kari Nilsson for help with some Figures and Robin Golser and Christoph Leubner for some information on Franz Selety.

Knut Lundmark bequeathed his collection of old astronomical literature to the University of Turku. It has been a rich treasury for us during the preparation of this book. The librarian of Tuorla Observatory Mats-Olof Snåre has kindly helped us to reach invaluable old titles from other libraries. We remember with warmth the late Toivo Jaakkola whose critical and practical approach to cosmology has greatly influenced us. Much of this book was written during the visits of one of us (Yu.B.) to Tuorla Observatory. He is grateful for the hospitality of the observatory staff, in particular to Sirpa Reinikainen. Now and then we refer to our own work on the subjects which we discuss. A formidable part of that science has been supported by the Academy of Finland, through the research project "Cosmology from the local to the deep galaxy universe" headed by P.T.

Credits for Figures

Usually we give the credits for the Figures in the captions. Some figures are from the Lundmark Archive of Tuorla Observatory. Here we thank for the permissions to use the following figures:

Fig.1.1 – ©The British Museum; Fig.5.1 – Photographie Benjamin Couprie, Institut International de Physique Solvay, courtesy AIP Emilio Segrè Visual Archives; Fig.7.6 – Halton Arp; Fig.7.7 – The Astronomical Association Ursa; Fig.8.1 – from "Alexander A. Friedmann" by E. Tropp, V. Frenkel, A. Chernin (Nauka, 1988; transl. Cambridge, 1993) with the authors' permission; Fig.9.1 – International Communication Agency; United States Information Service, courtesy AIP Emilio Segrè Visual Archives, Fermi Film and Lande Collections; Fig.9.3 – Sky & Telescope, Oct.1996;©Sky Publishing Corp; Figs.10.3, 15.7, 18.9 – Space Telescope Science Institute; Fig.12.3 – the Boomerang collaboration; Fig.16.9 – Anthony Fairall and John Wiley & Sons Publishing House; Figs.15.3, 17.3 – Benoit Mandelbrot; Fig.17.2 – Springer Verlag, from "The World of Galaxies", eds. B. Tully and L. Bottinelli; Figs.17.5, 17.11 – Francesco Sylos Labini; Fig.18.2 – from Fred Hoyle: "Home is Where the Wind Blows", University Science Books.

We especially thank Georges Paturel (Figs. 10.4, 16.2, 17.4), Arthur Chernin (Figs. 8.4, 18.3), Alexander Shneivais (Fig. 18.1), and Francesco Sylos Labini (Figs. 17.7, 18.7) for private photos and cartoons. Figures 15.6 and 16.6 utilize 100μ data from the IRAS infrared satellite (Wheelock et al.: IRAS Sky Survey Atlas: Explanatory Supplement).

* * *

We warmly thank our families. Yurij – his wife Nadezhda, his daughter Natalia, and his son Alexej and daughter-in-law Margarita and grandson Andrej. Pekka – his wife Elena, and his daughters Anastasia and Pauliina. They have asked us to explain what our science is and why it is important. Yurij is grateful to his father and mother whose encouragement has been so important. And we thank *babushka* of Pekka's Nastya and Paulushka for so much hot tea and care when we worked together in St. Petersburg.

Yurij Baryshev & Pekka Teerikorpi

at Tuorla Observatory, February 2002

Prologue

The 20th century was a great time for the science of the Universe – it was an era of true voyages of discovery, when both the covert shadows of the microworld and the breath-taking depths of the galaxy universe were penetrated. In the new realm of galaxies three totally unexpected cosmic phenomena were found: the cosmological redshift, the cosmic background radiation, and the large scale structures in the universe. These are cornerstones of modern cosmology.

This book is especially devoted to the development of ideas and observations related to the largest structures in the universe. The last decade has dramatically changed the possibility to study the distribution of galaxies in space and time. In 1951, George Gamow in his *The Creation of the Universe* regarded the idea of galaxy clustering as quite fascinating and lamented that it cannot be tested by observations. But what was not in reach fifty years ago, became a practical effort towards the end of the 1990's, when a new epoch of 3-D astronomy began. And the first results turned out to be amazing.

Three dimensional astronomy has arrived

Extensive measurements of redshifts for distant galaxies, using new technology telescopes, have made it possible to see their real positions in space. Before that astronomers could study only the projected appearances on the celestial vault and all conclusions about the space arrangement of galaxies were argued indirectly. Intriguing structures have been brought to light: the galaxy universe contains giant filamentary structures and huge voids, instead of dull uniformity. It exhibits *fractal* properties like many natural phenomena on and beyond the Earth. But now these self-similar structures, megafractals, have sizes approaching the limits of the largest 3-D maps so far observed.

Upheavals in the way of understanding the Universe

The cosmic architecture is the great mystery of the universe. It has always attracted the minds of deep thinkers, and at some bright moments of life it touches every human being. Scientific views of the structure of the world have passed a long path from the ancient harmony of heavenly spheres to the modern beauty of cosmic fractals. Within the period of quiescent evolution of science, a dramatic interplay between observations, theories, and principles arises from time to time, leading to a deeper understanding of the cosmic order.

The success of the modern Big Bang model may make one feel that an ultimate cosmology theory is already in hand. It again seems that we know almost everything and small details only of the grand picture remain to be finished. This admiration of the beauty and power of modern cosmology is shared, and with good reason, by a great majority of astronomers and physicists. But now, intriguing titles have started to appear on the pages of science magazines: *Is Cosmology in Crisis?*, *Revolution in Cosmology*, and *Fractal Universe: Everything we know about the cosmos might be wrong...*

Such conflicting views may only be appropriate for the difficult science of cosmology. What is called the "standard cosmological model" has actually undergone dramatic changes in the last decades, from a universe filled by ordinary matter to a world of unknown dark substance. A key question behind such headlines is the nature and organization of cosmic matter, both visible and invisible. Dark substance in all its forms has suddenly conquered the cosmological scene, becoming a star actor. The visible matter, less than one percent of all, has become a bit of spice in the tasteless substrate of dark matter.

Cosmology has always been and will probably always be surrounded by paradoxes and enigmas. Sometimes a paradox tells us that something important is missing, indicating a way towards a deeper view of Nature. And it may be that some enigmas will never be solved – reality may be too complicated and our observational and intellectual powers too limited. And often when we think that we know, the Universe strikes back...

The story of cosmic fractals

This book takes the reader on a voyage from the treasures of the past to the frontiers of the modern science of the universe. The profound transformation of the view on the cosmic structures has stimulated us to tell how cosmic fractals were discovered. The book reflects our personal experience in the study of the Milky Way, galaxies, and cosmology.

We highlight persons who have played remarkable roles in the attempts to explore the cosmic fabric. For instance, Fournier d'Albe wrote a book on a hierarchic world, which was the first attempt to make a simple fractal universe. Long before, Emanuel Swedenborg adopted self-similarity as a guide for understanding Nature. Knut Lundmark who measured the distance of the Andromeda galaxy, was, in effect, the first to ponder what later would be termed the fractal dimension of the local universe. Edwin Carpenter found an intriguing relation between the density and size of galaxy clusters, which three decades later inspired Gérard de Vaucouleurs to look seriously at hierarchic cosmology. Benoit Mandelbrot, the discoverer of fractals, was the first to consider a genuine fractal model for the galaxy universe. And we were impressed to see, while writing this book, how Henri Poincaré emerges in many connections, including fractals, relativity, and gravitation.

In recent times two of our colleagues have made decisive contributions to the understanding of the galaxy universe and to the discovery of cosmic megafractals. Georges Paturel made possible a comprehensive view of the local universe by creating the marvelous Lyon Extragalactic Database. Luciano Pietronero could see a novel general picture in the apparent confusion of diverse observations and introduced new concepts and methods of analysis.

Our spiritual guide in appreciating the value of observation in cosmology has been Allan Sandage, the founder of modern practical cosmology. These three persons have over the years taught us precious things in our science, some of which appear in our story.

PART I

THE SCIENCE OF COSMIC ORDER

Chapter 1

The birth of cosmological principles

History is a patient teacher, even though the pupils are often arrogant. It is true that nowadays we possess a tremendous amount of knowledge about the universe. But it is also true that an infinitely large store of information is still unknown for us. To see the birth, life, and decay of myths, guesses, and "ultimate theories" about the world is prone to soften human proudness. With the partial and limited knowledge our understanding of the cosmos is always an elegant simplification and an unintentional distortion of reality.

Of course, many ideas which have turned out to be fundamentally in error, deserve to be remembered with sympathy and respect, as the ancestors of our science, which gave birth to new generations of thought. But not all guesses have gone to ashes. Some old ideas appear now as fresh and brilliant as thousands of years ago. We begin our story of cosmic order with a journey through Antiquity, along a road covered by enigmas and paradoxes, adorned by triumphs, and after inevitable crises, now and again lighted up by creative imagination and the joy of discovery.

* * *

1.1 The seeds were sown – the myth explains the world

What we now speak about as cosmological, has deep roots in old times when our ancestors gathered around the fire and wondered about the glimmering stars. Their words left few traces before writing was invented, but what has been preserved offers precious glimpses on how inquiries on the universe have evolved, how old answers have given room for new questions.

The origins and structure of the world are described in old myths from peoples all around the Earth. The myths may differ, yet they share the attempt to explain how the universe was started and why it works as it does. They satisfied the need to understand, that restless spark within. For us the myths sound like strange, child-like tales. For the ancients they were real life poetry which extended by the means of language the experience towards unreachable places and times. The myths agreed with the animistic view of things. The Sun, Moon, and stars were living and possessed souls, as did the closer neighbors: Man himself and all animals and flowers on Earth. This view helped one to grasp why heavenly lights repeatedly appeared in the sky, formed constellations, and moved in an orderly fashion, though some of them also went their own, more capricious routes.

A story about the stars, told by the aborigines on the Malayan Peninsula, illustrates a myth. The Moon and the Sun are women. Only the Moon has children: these are the stars. In good old times the Sun had as many children. The ladies were afraid that mankind could not bear so much light and heat and agreed that each would devour her children. But the Moon hid her kiddies from the Sun's sight, who then believed that the Moon had really eaten them and ate up her own brood. The Moon brought her family out of their hiding-place. The Sun was filled with rage and chased the Moon to kill her. This has lasted ever since. Sometimes the Sun even comes near enough to bite the Moon, causing an eclipse. The Moon hides her children all the day, and only brings them out at night when her enemy is away.

This sad story tries to answer a burning question. Why is the sky not full of bright stars, pouring out intolerable heat and light? There seems to be an imbalance between day and night. The moon and her multitude of children can not light up the night anywhere to what the sun makes alone in the day, despite the fact that the two women are similar in form, size and color and both wander regularly across the sky. Apparently the sun could have children, but if she had, then it might be impossible to live on earth. An answer to this ancient riddle of dark night was given by the myth.

The listener was filled with an awe before celestial phenomena, but also with a feeling of comprehension. Knowing what happened in the sky, he could submit this knowledge to his children. A good myth was convincing and free from contradictions, like a good modern theory. That the myth includes things "fantastic" for us, like living celestial bodies, was no problem, but rather a basic assumption or "principle" of those times.

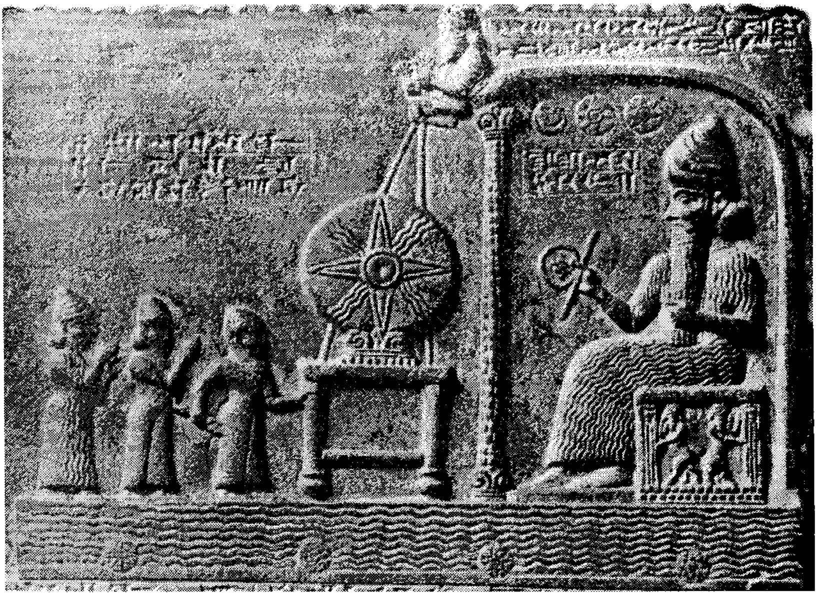


Fig. 1.1 A Babylonian stone tablet (c. 850 BC) showing the Sun God Shamash sitting. Close to his head one may see the symbols of the Moon, the Sun, and Venus. Two gods above the pillar hold the Solar disc by two ropes, ready to move it across the sky. At this time the concept of the celestial sphere was not yet known.

1.2 Celestial writing on the Babylonian sky

When the spherical Earth was not yet familiar, "up" and "down" were two equally mysterious directions. The solid Earth formed an impenetrable obstacle to inquiries about "down", while the sky, though equally unreachable, at least offered ample food for thought with its lights, colors, and movements. In ancient Babylon the sky was thought to be made of three layers. The lowest part was of blue jasper-stone and contained the stars. The second layer was made of saggilmut-stone. There lived the great god Marduk. The highest layer, of luludanitu-stone, was ruled by the supreme god called Anu. The gods, counted in hundreds if not thousands, could visit all the floors of this celestial building by going up or down like in a lift. As to "down", it was imagined that below the crust of the Earth were a cosmic reservoir of drinking water and the abodes of death.

With all our education we "see through" the sky vault, know that stars reside in the immensity of space. Nothing of the sort was known for the

Babylonian who tried to give the heaven some structure and meaning.

Writing had been invented around 3000 B.C. by Sumerians at the time populating this same fertile region of the rivers Tigris and Euphrates. Originally useful for bookkeeping in temples which were large economic centers, the art of cuneiform writing gradually found application also in other fields than business. Classification of the multitude of gods certainly benefitted from the ability to make notes, instead of relying on memory only. Around 2000 B.C. the first known astronomical notes recorded the planet Venus.

The Babylonian priest was interested in seeing signs from the gods who controlled things on the Earth. The stars, their constellations, and wandering planets were "celestial writing". It was natural to try to interpret it or put in correspondence with mundane incidents. This led to careful observations of the stars. The tradition of watching the sky for purposes so remote from the modern ones, was converted in the hands of the Greeks to astronomy, the driving force for the science of cosmology.

1.3 The Ionian Revolution

Something peculiar happened in the sixth century B.C. among the Ionian Greeks living in their flourishing colonies on the western coast of Asia Minor. Some of them started to look at things in a radically different way. Aristotle, 250 years later, tells us that they began to search for *principles.*

What is a principle? A basic, deep property of the world ties together apparently different things, allowing one to understand the diversity appearing around us. Also, using one principle one may hope to predict the occurrence of phenomena that previously were under the control of the plethora of capricious gods. Says Aristotle: *That of which all things that are consist, the first from which they come to be, the last into which they are resolved..., this, they say, is the element and this the principle of things, and therefore they think nothing is either generated or destroyed, since this sort of entity is always conserved... Yet, they do not all agree as to the number and the nature of these principles. Thales, the founder of this type of philosophy, says the principle is water.*

Thales of Miletus c.624-547

For modern man, fluently speaking about "the laws of Nature", all this sounds familiar. Such a mental revolution likely had predecessors in the shadows of history. But in Ionia the new thinking about the structure of the world burst into life with such a vigor that even though the first philosophers did not leave any writings, their influence was strongly felt and acknowledged by the later scholars of Antiquity.

The Ionian Greeks, having left the mainland in a search for better life, lived in the focal point of trade and exchange of ideas between the various nations. In this "situation of compulsory originality" (wrote M. I. Finley in his fine *The Ancient Greeks*) new modes of thinking had to arise.

To take one example, one is tempted to note that money was invented about 700 B.C. The first coins were made in Lydia, in Asia Minor. The Greeks adopted the new invention quickly. Money implied that quite different things (like an ox and a plough), have something in common which may be compared – their value or price. But there may be also something else in common, "from which all other things come to be, it being conserved". This was a cosmological claim about the whole world. The element is conserved in the changes that occur in this decaying world.

1.4 Anaximander solves the paradox of unfalling Earth

In the new atmosphere of reasoning Thales's water soon found competitors. Anaximander, a friend of Thales, said that the element which makes everything is something that cannot be named from among the known forms of matter. As to the structure of all that, it was not yet known that the Earth is spherical. Thales visioned it as flat and floating on water.

Anaximander c.611-546

Anaximander, however, was one of the first to suggest that the Earth did not need anything to float on. Clearly, if the Earth needs a support in order to avoid falling, say water, then what supports the water? If the water is in an immense bowl, what keeps the bowl steady?

Anaximander solved this riddle by proposing that the Earth rested freely in the center of the cosmos. He explained: as the Earth is equally far from every part of the cosmos surrounding it, it has no reason to move in any special direction, so it remains at rest. The Gordian knot of an endless chain of supporters was cut by assuming that the universe is symmetrical.*

This reasoning called forth Aristotle's dry humor: It was as if a hungry man surrounded by a circle of food and wine were starving, because he

---

*We also recall a passage in the Old Testament, where an unknown prophet and poet characterizes the power of God who "stretcheth out the north over the empty place, and hangeth the earth upon nothing" (Job 26:7). Could this be another early expression of the idea which rejected the mythical three whales as supporters of the Earth?

cannot decide from which direction to pick his meal...A Medieval soul mate of the poor man was the famous ass of Buridan, suffering between two huge haystacks.

Anaximenes c.585-526

Anaximenes, also from Miletus, was perhaps the first to view the sky as a sphere (or at least a hemisphere, like a cap), where the stars are fixed. This was a wonderful invention. In fact, a sphere is not the first impression, when one looks at the sky. The clouds make it look more like the lid of a saucepan. Anaximenes's flash of imagination is a good example of the Ionian way of looking at things: One revolving globe is enough to explain the daily rotation of thousands of stars. When at Babylon each star was a god, their parade on the flat sky depended critically on the discipline and care of each, no slipping from the row could be accepted.

1.5 The sky becomes a sphere

Pythagoras 6th century

While most philosophers searched for the basic element among some substance (water, air, or more exotic), Pythagoras of Samos, an obscure but so influential figure in history, is said to have made the far-reaching claim that the principle is *number*. As the cosmos (meaning 'ordered universe') is ruled by mathematics, it is possible for a thinking human being to learn of its structure, even without visiting its every corner. Geometrical forms and numbers became a part of cosmological thinking, which heralded the mathematization of the universe, now so basic in science: descriptions or *models* of phenomena are given via the language of mathematics.

The Pythagoreans taught that the Earth is spherical, as is the starry sky. [†] Planets, including the Sun and the Moon, are each attached to their own spheres which revolve around the Earth. The radii of the spheres were in "harmonic" ratios. This concept came from the first physics experiments where the results had been given in terms of numbers: the integer ratios of the strings of the lyre had been found, giving rise to harmonies pleasant to the human ear. And the first cosmologists, who called music and astronomy sister sciences, hurried to apply these "local" physical laws to the heavens, as their modern colleagues also eagerly do.

[†] There were hints suggesting that we live on a globe. For example, seafarers had noticed that a ship sailing away disappeared over the horizon, starting from the hull and finishing with the masthead. And travellers noticed that the sky changes when they go from north to south, but remains unchanged during a trip in the east-west direction.

The Pythagoreans thought that everything in Nature can be measured by integer numbers, the only type of number conceived of at the time. A shock came when one of them showed, using the Master's famous theorem, that the ratio of the diagonal and the side of any square cannot be expressed in terms of integers ($=\sqrt{2}$). They had imagined that lines are formed by large numbers of points, like atoms side by side, and the ratio of any two line segments would always be rational. This view had to be abandoned. In its place came two types of numbers, the old ones ("rational") and the new ones "irrational". The way was paved for irrational numbers and the mathematical continuum of modern mathematics.

1.6 Atomists see a glimpse of the microcosm

Leukippos of Miletus 5th century

As another candidate for the basic principle atomists taught that it is not continuous matter, but is formed of very small and hard particles, called atoms (meaning indivisible). This theory, founded by Leukippos, pictured the world as a void where the atoms fly without aim or purpose. Infinite space and endless time guarantee that sooner or later the atoms that sometimes collide, form whole worlds. Ours is only one example. [‡]

As the atoms underlie everything, the atomists could offer explanations of visible things, starting from invisible ones. A practical example is how wet clothes hung on a rope grow dry in the sun. We cannot see the moisture leaving them, because it is split up into minute parts which the eye cannot detect. Also, we cannot see the air, though it quite clearly is a material substance as any stormy day testifies. Invisible particles move swiftly and when they all together bombard the sails of a boat, a visible effect results.

For modern man the concept of the atom is natural as mother's milk, and of all the Greek scientific ideas he feels best at home with the atoms roaming in the infinite universe. However, the atomists in all their originality and feverish search for truth could not advance their science beyond interesting speculations. The thread of thought leading to modern cosmology, rather started from the contemplation of regular heavenly movements.

[‡] The vacuum of the atomists resembled the space of Newton: it was infinite and did not offer any resistance to the atoms speeding through it. Instead of blind collisions, Newton added gravitational interaction to the bodies, the "Love" of Empedocles, which in his universe collected matter together to form stars. Empedocles, (cir. 494-434) known for his four elements (Fire, Air, Water, Earth), wrote poems about the forces in Nature. He called the attractive force "Love", the repulsive force was "Hate".

1.7 Plato's mathematical heaven

Plato 427-347

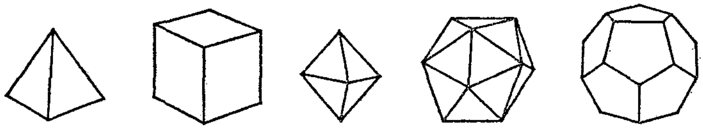
On the path of science, an important step was made on the initiative of a man who held the radical view that one may learn of the true essence of the world by *pure thinking.* The philosopher Plato taught that what we see is only an incomplete and muddy image of the real world, but a thinker should be able to traverse this misty curtain and find the true mathematical laws. Though no longer as confident of the power of pure thinking, the modern scientist does, like Plato, assume that the universe and its phenomena can be given a mathematical description. For Plato, this was possible only for the orderly Heavens, while all disorder is present on the Earth.

Eudoxos of Cnidos c.408-355

In his Academy, situated in a peaceful Athenian park named after the mythical hero Akademos, Plato had a group of talented students. It is told that he gave them a task: can the complicated movements of the stars and planets, sometimes so confusing for the eye, be understood by simple, uniform motions, pleasant for the mind and essential for the true World? Though Plato's actual role is not so clear, Eudoxos presented along these lines his famous *theory of homocentric spheres.* It was the first mathematical model made for understanding in some detail the motions of the sky. The model needed 26 spheres rotating with different but uniform speeds around their axes. These axes connected an inner sphere to the next one and were inclined with fixed angles relative to each other. The model could explain why some planets occasionally make a loop in the sky.

In his old age Plato wrote the impressive dialog *Timaeus* where he described the universe as he saw it. The backbone of Plato's philosophy is the existence of the real, directly unobservable, and eternal world which can be approached by reason only. It is characterized by beauty, harmony, order, similarity, and symmetry, i.e. by concepts which Plato (and modern man) associate with mathematics. Plato proposes that the observable universe was created by a benevolent Demiurge, the artificer of the world, as a unique "copy" of the true world. Incidentally, he wisely emphasizes that the story of how the universe was created is only a probable one. The structure of the universe can also be only approximately known. But it is well worth trying!

Like a good architect, the Demiurge wanted to make his construction of the universe as good as possible, the best one in fact. He wished to impress the mathematical pattern of the true world on the featureless primordial matter. In order to reflect completely its ideal, the resulting material world



Tetrahedron Cube Octahedron Icosahedron Dodecahedron

Fig. 1.2 The five regular "Platonic solids" representing (from left to right) fire, earth, air, water, and celestial matter. Thus regular geometry not only belonged to the realm of celestial spheres, but could also be found amongst the things down here, albeit those with plane surfaces and sharp edges.

should have such a form which contains all existing forms, including itself. This is what is now known as a "fractal". Plato, who had not heard about fractals, regarded that of all the geometrical forms the sphere is the most perfect and most similar to itself. And thus the Demiurge, "who rated similarity thousands of times higher than dissimilarity", chose the form of the universe to be a sphere. And it was put in rotation. Uniform, circular motion where the heavenly body regularly returns to its previous place is closest to "unchangeable", the property of the unobservable eternal world.

Stars were made spherical as the universe itself, and similarly to the universe they also rotate around their axis, besides taking part in the universal rotation together with the spherical sky. In *Timeaus* Plato mentions the "choir dance" of the planets, but does not go into details, which he says would require an "instrument", perhaps referring to Eudoxos's model.

Of course, the Earth is also spherical. But how can the "disorder" down here reflect the order of the eternal world? Plato has an ingenious solution. The Demiurge hid the order and mathematics into the microcosm. There are four elements around us. But one does not easily notice that the elements are made of very small geometric forms as if atoms or crystals. The Demiurge made this material basis as beautiful as possible. The elementary forms are regular solids, each corresponding to an element. The tetrahedron was related to fire, the cube to earth, the octahedron to air, and the icosahedron with its 20 equilateral triangles was assigned to water. There is also a fifth (but no more!) regular solid, the dodecahedron made of 12 pentagons. This has some resemblance to a sphere and Plato is inclined to associate it with the universe as a whole. Later philosophers spoke about the fifth element out of which the heavenly realm is made.

These geometric forms were not just passive. Plato sketched a kind of chemistry which allowed transformation of one element to another. The generation of new combinations leads to the rich material world that we see all around. This was restricted to the three elements with equilateral triangles in their forms (fire, air, water). The equilateral triangles may be divided into six similar rectangular triangles (with angles 90, 60, 30 degrees) and these are the permanent forms which do not vanish. So, the tetrahedron of fire may in a "reaction" dissolve into 4×6 rectangular triangles and the octahedron of air into 8×6 triangles. From one fire element (24 triangles) and two air elements ($2 \times 48 = 96$ triangles) can be made one water element having in its icosahedron $20 \times 6 = 120$ equilateral triangles. [§]

1.8 Aristotle's scientific method

Aristotle
384-322

Aristotle, the brilliant pupil of Plato, defined what scientific knowledge is. According to him, different branches of science rest on their initial axioms, from which one can derive all knowledge by logic. Axioms can be recognized by careful *observation* of Nature, together with deep *intuition.* This is not so far from what a present-day scientist thinks. Though, in Antiquity, active experiments were rare. Aristotle also very firmly demanded that the axioms should represent the highest level of knowledge: they should be "true, primary, immediate, better known than, prior to, and causative of the conclusion". Modern scientist is more modest with his axioms which rather are temporary assumptions or hypotheses.

Clearly, the world around us is an arena of change and motion. Motion is a basic phenomenon that an inquiring mind would like to grasp. Aristotle saw two types of motion. The *natural motion* either strives towards the world center which coincides with the center of the Earth or away from it, or is a circular motion around it. There is also *forced motion.* So, there are no freely moving bodies: any motion either requires a permanent force or is a natural one. Stated in modern terms, Aristotle's physics did not contain inertial motion.

Aristotle took over Eudoxos's planetary model, together with his own physics of motion, as a basis for his geocentric cosmology. The Earth is in the center of a huge clockwork made of no less than 56 globes. The outermost globe carries on its surface the fixed stars and has rotated uniformly

---

[§] The cubic (earth) is not made of such fundamental triangles, hence this most solid of the elements does not participate in the transformations.

forever. It transmits downward rotational motion for the various spheres producing the motions of the planets.

One may describe Aristotle's cosmos as stationary – it had always existed and did not evolve. When one looked at the various chains of phenomena, they seemed to flow through ready-made channels, towards what seemed to be the final causes of the processes. A stone falls, because its goal is its natural place in the center of the universe. Aristotle insisted that a scientific explanation of a process is not complete without an account of its final cause (*telos*) which is like a force coming from the future, influencing what should happen now. For us this seems exotic – modern science rather starts from the past (from "initial conditions"), permits evolution and attempts to understand what would happen in the future.

In medieval Europe Aristotle was enthroned as the highest authority in matters of science. It has been sometimes implied that his influence slowed down the development of science. Such a view seems rather narrow. His and other Greek texts must have offered immensely important impulses in an epoch when there was little native scientific tradition.

Aristotle was a keen observer of Nature, a true scientist, and a prolific writer of seminal texts in many areas of knowledge. His scientific method aimed at finding fundamental principles by *observing* phenomena. He is not to be blamed if his medieval followers preferred to read his books as the final truth, perhaps in the same spirit as university students read textbooks without yet realizing that science is the activity which changes the contents of these same books. [¶]

1.9 The principle of circular motion

Eudoxos's model was a geometric construction reflecting Plato's world of ideas, while Aristotle's system was more intended to represent the observable universe. But there was a problem with Eudoxos's planetary model. When the planets make their loops, they are brighter than in other times, so probably closer to the Earth. In Eudoxos's model, the planets stay always

---

[¶] Aristotle wrote plenty of books. It is an example of how weak our connection with the past is, that none of these were preserved in complete form. Our knowledge of his thoughts is based on "lecture notes" and summaries, and even these were lost for two centuries before they were found in the cellar of a descendant of one of his pupils. Interestingly, only a few years ago archaeologists discovered the place in Athens, in which Aristotle's famous *Lyceum* was situated.

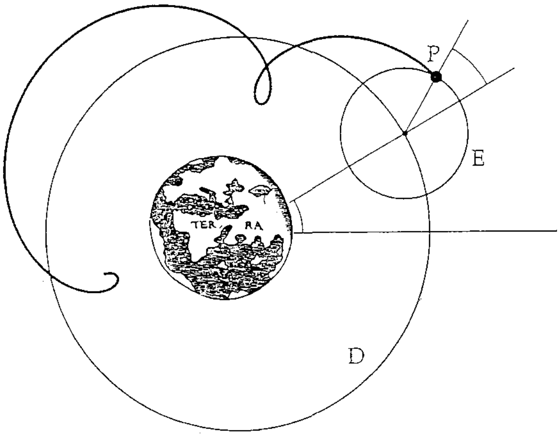


Fig. 1.3 The complex motion of a planet (P) around the Earth was explained by Apollonius as a sum of two circular motions, those around the epicycle (E) whose center goes around the deferent (D). During a loop in the sky, the planet is closest to the Earth.

at the same distance from the Earth. This inconsistency with observations was remedied by Apollonius of Perga. He invented the *epicyclic theory.* In his model, a planet did not stay firmly on its own sphere-carrier; it was destined to wander on a smaller sphere (*epicycle*), the center of which was fixed on the rotating main globe. Such a cosmic merry-go-round explained fairly well the motions of the planets.

Apollonius c.265-176

The epicyclic theory was given its most developed form by Ptolemy of Alexandria in his Great Synthesis which Islamic savants later called the *Almagest.* The Ptolemaic system was a genuine scientific world model. It was based on astronomical observations, made with instruments to measure angles in the sky. Its theory used the mathematics of that time and it predicted motions of celestial objects with good accuracy. Indeed, the fate of a physical theory is decided by its ability to predict things – a theory of the heavens is put in test practically everyday.

Claudius Ptolemy 87-165

We'll often encounter the "three whales of cosmology": principle, theory, and observation. These elements are clearly visible in the Ptolemaic system. Besides geocentrism, Ptolemy took a further principle from tradition. He wrote: ...*the goal that the astronomer ought to aim at is the following: to show that the phenomena of the heavens are reproduced by circular, uniform motions.* Thus uniform motion in a circle had become a cosmological principle among the scientists who regarded it as the only suitable motion for the heaven. Every sunny day and every starry night appears to enforce this idea. Such a luxury is not enjoyed by our modern cosmological principles whose validity can be tested only using the largest telescopes. This tradition of thinking was so strong that Copernicus, The Revolutionist, was faithful to it. Even Galileo could not admit other than circular motion for the planets.

The present writers can hardly imagine life without the hot, smokey steam and dark, sooted corners of the Finnish "sauna" or the Russian "banya", traditionally built rectangularly. We have found it delightful that the Greek sweating room "laconicum" was built with a *circular* ground plan, apparently being something closest to Heaven on Earth...

1.10 But what is actually rotating?

The mainstream of cosmology eclipsed other brilliant ideas about revolutions of celestial bodies, which were not yet ripe to defend themselves. So Philolaus, about whom very little is known, is said to have taught that the Earth and other heavenly bodies revolve around the fire burning in the center of the world. The fire can not be seen, as the Earth always turns one side towards it. Heraclides of Pontus, a pupil of Plato, taught that the Earth rotates around its own axis. The daily motion of the sky is only an apparent phenomenon for the observer on the rotating Earth.

Philolaus cir.450

Heraclides 388-315

Heraclides was close to becoming the head of the Academy after the death of Plato's successor Speusippos – he got a few votes less in the election than Xenokrates. The question of the Earth's motion must have received ample attention in Plato's Academy. There is even some hint in Timaeus that Plato may have regarded our globe as revolving.

Aristarchus of Samos taught that the Earth and planets revolve around the Sun. Little is known about how "the Copernicus of Antiquity" came upon this idea. Based on Euclidean geometry, he devised clever methods to derive cosmic distances. Though he could measure the distance and size of the Moon reasonably well, his estimate for the distance of the Sun was widely off the mark (too small). Even so he realized that the Sun is much larger than the Earth, which may have been one reason of putting the brilliant torch into the center instead of our modest stony globe.

Aristarchus 310-230

1.11 Towards the principle of no center

The most influential principle of the ancients held that the Earth is the natural center of the universe. But typically for the scientific intuition of the Greeks, they also entertained a quite contrary idea, that of no center.

Epicurus 341-270

Epicurus was born in the island of Samos. He settled at the age of 35 years in Athens and founded a school based on the idea of atomism. Of his extensive writings only the contents of three letters are known. The longest one, to his pupil Herodes, is a small treasury of cosmological ideas.

Epicurus explains that the universe as a whole cannot change. It has always been and will always be as it is today: "There is nothing into which it could change. For there is nothing outside the totality of the universe which could incorporate and provoke its change." The totality consists of bodies and vacuum, and the bodies consist of indivisible atoms. The world cannot have any border, otherwise it would have an "exterior" part.

Though the infinite universe of atomists sounds familiar, an intriguing feature makes it different from our ordinary conception of infinite space. The universe of Epicurus was not yet isotropic, i.e. did not have equal properties in every direction. It had a preferred direction along which atoms were all the time falling. Interestingly, small and big atoms fell alike, with an equal, huge velocity! Sometimes they unexpectedly interacted, which gave rise to the seeds of cosmic structures such as worlds like ours.

Lucretius c.98-55

Lucretius, a Roman poet and admirer of atomism and Epicurus, wrote: *In every direction the world is without end. ... As nothing is outside of the universe, it has no borders and no size and no end. It is not important in which region you stay ... in every direction the universe is infinitely large.*

These words highlight the cosmological principle of no center as it was understood in Antiquity. But the life of this brave idea was shaded by the revolving bowl of night, which made the center of the Earth a special point.

1.12 The Wisdom of Antiquity was kept alive

Ptolemy lived in Alexandria in the epoch of the Roman empire when the cultural heritage of Greece was declining. The scientific center of the world, the Museum of Alexandria, had been founded around 300 B.C. It housed about 500 000 manuscripts which the talented scholars could use in their studies of literature, mathematics, astronomy, and medicine.

In 312 A.D. Constantine the Great embraced the Christian religion which eventually became the only accepted faith in the Empire. The consequences for science were negative. Besides general unsympathy for the study of mundane things, there were extremists who opposed pagan culture. In 390 A.D. a large part of the Alexandrian library was destroyed by Christians. The philosopher Hypatia, whose rare occupation for a woman in Antiquity did not increase her popularity among those opposing pagan doctrines, was murdered in 415 A.D. After this bloody deed many scholars migrated to the Academy in Athens and to Constantinople. The final death-blow to the library came in 642 A.D. when the Mohammedans conquered Alexandria. The story tells that the manuscripts kept the 4000 city baths fueled for half a year...

In 529 A.D. the Emperor Justinianus closed Plato's Academy, after nine centuries of operation. In Europe the Dark Ages had begun, and centuries passed there without much interest in science. The thinkers had other ideals, remembering how St. Augustine, in the fourth century, in his *Confessions* warned them of the "disease of curiosity..., which drives us on to try to discover the secrets of nature, which are beyond our understanding... I no longer dream of the stars." But the treasures of the past were partly preserved in the empire of Muhammed. Its able scientists translated Greek texts some of which had luckily endured the hard times.

Cosmas Indicopleustes 6th century

One of the earliest world maps was drawn by Cosmas Indicopleustes on the verge of the Dark Ages, in the sixth century. Cosmas, a native of Alexandria, spent many years as a seafaring merchant traveling in Asia and Africa, till he went into monastic life. There he wrote books on geography, the structure of the world and the Scriptures. Of these has survived us *The Christian Topography*, known for its keen attempt to picture the universe strictly according to the Bible, taking into account what Cosmas had seen during his long journeys. Of course, the resulting cosmography now seems rather odd, with its tabernacle-formed universe, quadrangular earth, and intentional absence of any "pagan" spheres and circular motions. Clearly, the design was much influenced by what Cosmas thought he had found in the Bible, and even his rather wide travels did not contradict this picture.

Modern cosmographers know that the Earth is round, but even they cannot actually travel much further than Cosmas did. Their eyes are no better, but since Galileo they have learned to play with the heavenly messengers, the rays of light descending from the sky. Telescopes help a lot, but even in our days it is not easy to draw a picture of the cosmic landscapes.

1.13 The world edifice of the Middle Ages

In the 12th century, people started to translate the Greek texts into Latin, mostly from Arabic versions. The works of Aristotle and others were received with enthusiasm among the European scholars. They must have felt as if a treasury had been opened before their eyes. Listen to the words of Bernard of Chartres, a scholar living in 12th century France: "We are dwarfs who have been lifted on the shoulders of giants. We thus see more and farther than they do, not because our eyes are sharper or we are taller, but because they hold us in the air, above their gigantic heights..."

St. Thomas Aquinas 1225-1274

At first the Church was not at all happy with Aristotle's firm views on the universe and the natural laws, which even seemed to call into question the unlimited power of God. But when St. Thomas Aquinas could unite the Scriptures and the ideas of "the giants", the result was the special medieval cosmology. This doctrine encompassed God and Man, Heaven and Earth, and made the physics and cosmology of Aristotle the official paradigm taught in schools and universities. The universe of spheres seemed, after all, to fit well with the dogma of the Catholic Church. God had made the immovable Earth, and all the rest rotated around the Man, the summit of Creation.

Dante Alighieri 1265-1321

In his *Divina Commedia* Dante Alighieri drew an unforgettable picture of the cosmology of the Middle Ages. The poem describes Dante's visit to Hell, Purgatory, and Paradise. Hell is a cone extending down to the center of the Earth, while Purgatory is a conical mountain on the opposite side. From the top of Purgatory, Dante finally rises to Paradise, which contains increasingly lovely levels (planetary spheres), ending at the tenth Heaven, Empyrean, which is the most blessed place of all, the dwelling place of God.

Dante, the poet, scarcely mentions epicycles and other finesses of the mathematical world system. He pictures the universe in its grand design, as it was commonly thought among those other than astronomers, showing the real importance of its structure for the human being.

Both material and spiritual in essence, Man has two competing natural directions of movement. Depending on the balance between his material and spiritual sides, he will either descend after death into the horrible depths of Hell or ascend to the Heavens, towards God. This unification of science and faith resulted in a magnificent view of the cosmic position of Man, something that was to be lost during the Copernican revolution.

That is why the concrete *astronomical* world was so important for the

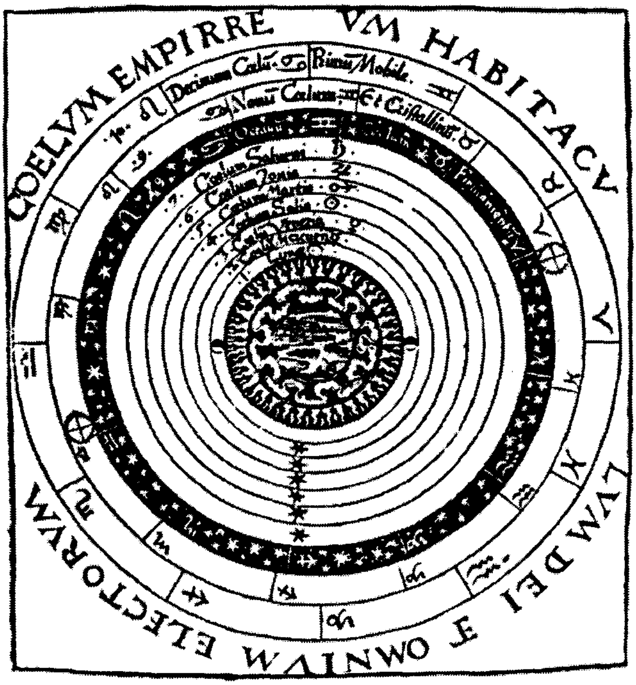


Fig. 1.4 The cosmos of the Middle Ages was bordered by the sphere of Primum Mobile or the first mover, beyond which lay the mansions of God. It was thought that the celestial world is physically quite different from the Earth and its atmosphere – a mortal being could not live there for a moment. If, however, one could somehow climb towards the "outer border", one would see how the physical reality changes and space and time lose their familiar meaning. According to Dante "distance does not decrease nor increase immediately there where God rules; the law of Nature does not exist there".

medieval mind. But the Church was tolerant to alternative *mathematical* models, as long as they did not claim to change the real order of things.

The science of the Middle Age, scholasticism, was concerned with thinking. People wanted to know what they mean by their words. The amalgamation of Aristotle's physics and the Bible had fixed an understandable and "ultimate" paradigm for cosmology. What remained for a scholar to do, was to study that strange reality formed by our language, using the method of logic, also founded by Aristotle. Though, scholastic analysis raised questions about the physical paradigm, too.

In the 14th century Jean Buridan and his pupil Nicole Oresme, criticized Aristotle's notion of force. It was a corner stone of his physics that "everything that is in motion must be moved by something". An arrow flies forward, because it is pushed by the air. The giant spheres of planets rotate, because they are all the time acted on by a force streaming from the sphere of the fixed stars. But why then doesn't an arrow with a blunt end fly faster than one with a pointed tail? Buridan suggested that something is added to a body when it is thrown into its trajectory. This something he called *impetus*.

Jean Buridan c.1297-1358

In a remarkable step Buridan applied his impetus to the motions of the spheres. It was common to think that the stellar sphere was kept in rotation by angels. However, Buridan noted, the Bible is silent about this. So perhaps God gave the spheres their motions when He created the World. Having got their impetus, they have rotated on their own, and without any friction, showing impetus at its purest. Recall how the dynamics of Newton was later discovered in the frictionless Solar system...

Nicole Oresme c.1320-1382

Oresme did not accept Aristotle's proofs for the motionless Earth. He argued that every motion is relative. The Earth may rotate around its axis, giving the starry sky the appearance of rotation, "as a man in the moving ship thinks that it is the trees outside of the ship that move". Aristotle had argued against this by noting that a stone thrown directly upwards falls down on the same spot. The surface of the rotating Earth has meanwhile moved hundreds of meters aside. Oresme saw the test in the light of impetus: the stone just preserves its own share of impetus that it has together with the moving Earth.

* * *

The result of much study, effort and thought, a world view is a child of its time, dear to its contemporaries, and carries with it elements which at that time appeal to both soul and reason. So Jean Buridan and Nicole Oresme accepted that the Earth is motionless, even though they criticized the proofs: one should not use poor arguments in defending the truth. That it had become possible to imagine the rotation of the giant heavenly spheres without the pushing angels, might be called an adjustment of the cosmological model to better fit new ideas developed by physicists down on the Earth. However, the road of small adjustments had come to its end.

Chapter 2

The gate into cosmic order

Nicolaus Copernicus was a miracle man: he stopped the sky which had revolved from times immemorial. And he displaced our Earth from the time-honored center position, and around the movement of this precious stone, modern science started to grow. In 1543 the Polish astronomer and canon of the bishop's council in Frauenburg published, after years of hesitation, his great life work *De Revolutionibus Orbium Coelestium.* With this book, "On the Revolutions of the Heavenly Spheres", the Earth became just a mediocre planet. In his own words:

Nicolaus Copernicus 1473-1543

All the motions of the Sun that we see, do not belong to it, but are possessed by the Earth and our sphere with which we go around the Sun as every other planet... In this way this one motion is sufficient for explaining a large number of apparent irregularities.

This was the discovery of the Solar System, our true home built on universal laws, and the initial link in the chain of the cosmic hierarchy.

* * *

2.1 Roots of De Revolutionibus

What led a peaceful and rather timid servant of the Catholic Church to his startling idea on cosmic order centered on the Sun? Neither the sky nor the Earth offered very compelling evidence against the Ptolemaic system. Thomas Kuhn, who took the Copernican revolution as an example of the "paradigm breaking", thinks that the old system had become an intolerably complex and clumsy "monster". Religious and philosophical movements having Sun worship in their program may have also paved the way.

Raimo Lehti, a Finnish mathematician, has studied this question, and concludes that the *solar relations* were the key. Planets perform their dance around the sky as if conducted by the Sun. For example, the loops of planets always happen when the Sun is on the opposite side of the sky. Perhaps Copernicus got his idea of the central Sun from these regularities which were earlier regarded as a wonder that God had put on the motions of planets. This led him to search for a new system on the mathematical guidelines drawn by Ptolemy who remained for Copernicus his great Master. "Copernicus did not find heliocentricity by observing Nature, but by studying the Almagest of Ptolemy", writes Lehti. Almagest, in itself based on observations of planets, contained the germs of the future cosmology.

Copernicus also received other messages from the past. He referred to ancient Greeks who had pondered alternative cosmological ideas: *Taking advantage of this I too began to think of the mobility of the Earth; and though the opinion seemed absurd, yet knowing now that others before me had been granted freedom to imagine such circles as they chose to explain the phenomena of the stars, I considered that I also might easily be allowed to try whether, by assuming some motion of the Earth, sounder explanations than theirs for the revolution of the celestial spheres might so be discovered.*

The "absurd" ideas, luckily transmitted from the past, may have been the treasure that had to be found among the dusty texts: "the stone which the builders rejected, the same is become the head of the corner"...

2.2 New understanding on matters celestial

In Antiquity, it was a great simplification to picture the sky, previously flat, as a giant sphere carrying the stars, with the Earth fixed in its center. However, then a complicated epicyclic machinery was needed to explain the dance of planets. Copernicus realized that if one allows to the Earth real motions, one may *understand* celestial phenomena in a simple way:

* *the daily rotation of the starry sky*
* *the yearly wandering of the Sun around the sky and the seasons*
* *the regularly repeated loops of the planets*

In fact, Copernicus was following Ptolemy who considered "it a good principle to explain phenomena by the simplest hypothesis possible, in so far as

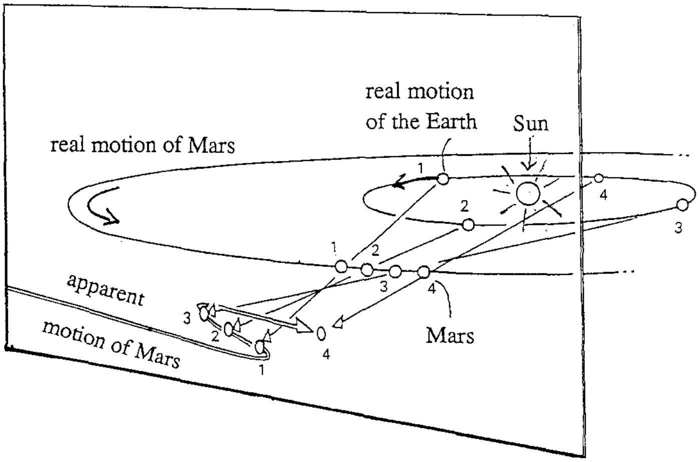


Fig. 2.1 From time to time planets make loops when they wander relative to the fixed stars. For Ptolemy the dance of planets (in the figure the 'apparent motion of Mars') was the reason to add epicycles in his system, while for Copernicus this phenomenon revealed the motion of the Earth around the Sun. Aristarchus proposed that the Sun is at the center. Did he apply his idea to the dancing planets? We do not know.

there is nothing in observations to provide a significant objection to such a procedure." (Ptolemy might agree that to let the Earth move would simplify the model, but he would not accept its movement.) True, the new system, for us conceptually simple, was at the time technically almost as complex as its honored predecessor: plenty of spheres, circular motions and epicycles explained the details of celestial motions. It was received with some enthusiasm by mathematicians who were able to go through the difficult De Revolutionibus - book. (In fact, the book was no best seller; its first edition of a thousand copies was never sold out.) The Catholic Church was first rather indifferent, not to speak of the Orthodox Church which did not regard the movements of the physical Earth relevant at all, – angry protests against the catholic clerk came rather from the Lutherians. Only in 1616, seven decades after its publication, did the Holy Office take action.

Perhaps the long time when any Catholic could read Copernicus was thanks to a preface added to the book without the author's consent. It explained that the book offers a new method to calculate the positions of planets and does not claim that the Sun really is the center of the universe.

Such a statement reflected a long tradition. The astronomer may devise complex models which explain the appearances in the sky, without thinking that all the spheres really exist in the heavens. Medieval Aristotelians did not attribute a concrete reality to the epicycles which were a mathematical machinery for reproducing what is observed in the sky. What about Copernicus? His own preface to De Revolutionibus makes it clear that he presents a new *physical* world model, according to which the Earth is really moving in space. *

2.3 The young Rheticus visits the old Copernicus

Rheticus 1514-1574

The mathematician Rheticus was 23 years old when he decided to go and meet Copernicus personally. His ideas about the world structure had reached the young scientist in the form of a short "samizdat" manuscript that had circulated among astronomers. Rheticus wished to persuade Copernicus, then 66 years old, to publish his great work in totality. He even had as seductive presents beautiful books on mathematics, with white pigskin covers. The visit took longer than planned, almost two years. Rheticus was very excited by the new ideas, and started to spread the Copernican system, though still anonymously. As a result of the efforts by Rheticus and another friend, Copernicus finally agreed to publish his extensive manuscript.

Probably Copernicus started thinking about the new system during his university years in Italy, where he studied theology, law, and medicine, and also became acquainted with astronomy. This rather shy man with a versatile renaissance education did not spread actively his revolutionary ideas, on the contrary, De Revolutionibus could well have been left unpublished without interference by the younger generation. When he received the freshly printed first edition of his book, at the age of seventy, Copernicus was already mortally ill in his bed, but the great mission was complete.

---

*One may add to the above success list the precession of the vernal equinox. In the course of the year the Sun wanders around the sky and intersects the equator two times, once in the spring at the time of the vernal equinox when it goes from the northern to the southern half of the celestial sphere, and once in the autumn (autumnal equinox). The points of intersection shift slowly along the zodiac from one constellation to another over thousands of years, as was shown by Hipparchus (cir. 190-120). This motion reflects a slow spindle-like wobbling of the Earth's axis.

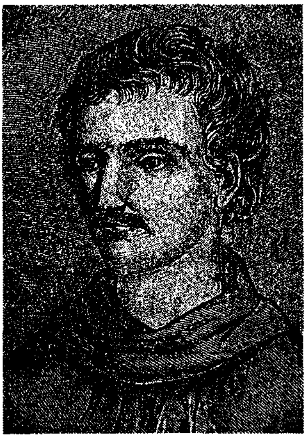


Fig. 2.2 Giordano Bruno made huge mental steps towards a new picture of the cosmos. He presented in a clear manner the cosmological principles of No Center and Universality of Earthly Laws.

2.4 Bruno breaks the stellar sphere

When Copernicus stopped the revolving sphere of the stars, he did not question the sphere itself and left the stars fixed on it. During the following one hundred years people began to realize that such a crystal globe without its original function of making the stars go round is of no use any more. Heated by the central Copernican Sun it evaporated before the eyes of one or two generations, dispersing the stars into the remote depths of space.

Giordano Bruno 1548-1600

In the writings of Giordano Bruno, possibly inspired by his contemporary Thomas Digges (see below), one may clearly see this outcome of the new world order: "As soon as we realize that the apparent celestial rotation is caused by the real daily motion of the Earth ... then there is no reason to make us think that the stars are at equal distance from us".

Bruno ardently supported Copernicus and went even further: he said that the Sun is only one of the stars and these are scattered in an infinite universe. "As the universe is infinite, there must be more suns ... one can assume that there is an infinite number of suns, many of which are seen for us in the form of small bodies; and many may appear for us as small stars".

Giordano Bruno had in his youth gone to a Dominican monastery. His original thinking caused him troubles and at the age of 28 he escaped and spent years strolling around Europe and teaching in universities, usually causing stormy protests among the more conservative part of the listeners.

In 1591 Bruno fatefully returned to his native Italy, invited by a young aristocrat whose eagerness to learn the secrets of philosophy turned out to be a superficial hunger for exotic things. The disappointed pupil led his teacher to the hands of the Inquisition. Bruno was arrested and accused of heresy: he had not only claimed that the prevailing view on the cities of God and Man is erroneous, but more importantly, he regarded God as a pantheistic spirit and he denied transsubstantiation and the immaculate conception, the central doctrines of the Church. Bruno languished in the prison of the Inquisition for seven years, before he was burned at the stake in Rome, at the Square of Flowers (Campo dei Fiori) in the spring of 1600.

These tragic events make us share with the reader a few lines of "Albertino's prayer" from Bruno's book *About Infinity, Universe, and Worlds*: *Convince us of the doctrine of the infinite Universe! Break these imaginary vaults and spherical surfaces, which border so and so many heavens and elements. Give us the doctrine of the universality of Earthly laws in all the worlds and of the uniformity of cosmic matter. Open for us the gate, through which we can look at the countless, everywhere similar stellar worlds.*

Such inspired words might well be signed by any cosmologist of our day – they convey the spirit of how we today think about the universe. Inspecting them closer, one may see three aspects: 1) The cosmological model of an infinite universe, instead of the old model with spherical surfaces. 2) The cosmological principle of universal laws and similar matter everywhere (later we'll encounter Bruno's principle of no center). 3) The understanding of the need for novel observational means to probe the depths of space.

The "three whales of cosmology", theory, principle and observation, already lived a simple life in Bruno's thought. Though no astronomer, Bruno was aware of the difficulties which hamper attempts to observe distant celestial bodies. Stars are like our Sun, but so distant that they look like points of light. Around them are planetary systems, but the planets are too faint to be seen. Bruno also argued that even in our Solar System there may be other planets which we cannot see for various reasons, e.g. they may be very distant, or they may be small in size, or they may be poor reflectors of sunlight. Having to base his cosmological thinking on scarce observations, Bruno explained the absence of direct evidence as a result of *selection effects* – this concept is still quite important in modern cosmology.

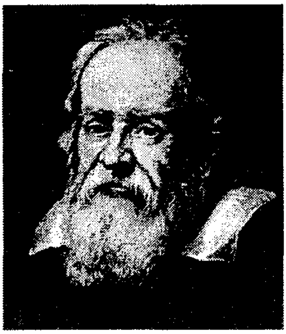


Fig. 2.3 "There are more things in heaven and earth, Horatio, than are dreamed of in your philosophy", wrote a certain William a decade before Galileo pointed his telescope to the sky. Galileo showed one way to unveil those unexpected things, just by looking at deep space. But he also paved another way, well suited for the things down here on Earth, that of testing by experiment philosophical claims about matter and motion.

2.5 ... and Galileo opens the gate

Galileo Galilei made an *astronomical telescope* and observed the structure of the universe much better than was possible by the naked eyes of all previous philosophers. This professor of mathematics at Padua opened the gate to the heavens, a few years after Bruno's death.

Galileo Galilei 1564-1642

By modern standards, Galileo's magnifying tube was modest, having a diameter of a couple of centimeters. [†] Nevertheless, at the moment when it was pointed at the sky, the power of the human eye jumped to a new level, and unexpected information started to flow in. There are mountains on the Moon and satellites around Jupiter. And the face of the Sun has spots and the Milky Way is a gigantic cloud of faint stars.

Galileo was also the founder of a new physics, based on experiments. So he concluded from balls rolling down tilted plates that a body conserves its steady motion when no friction is acting on it. This insight, crucial for the forming, new cosmology, explained why the atmosphere may rotate together with the Earth without intolerable winds. The notion of a free body, absent in old physics, was adopted later by Newton as one cornerstone for his mechanics.

[†] Galileo had heard that in the Netherlands a lens grinder had built a device which made distant objects look nearby. He soon succeeded, in 1609, in building such an instrument himself. Two telescopes made by Galileo are preserved in Istituto e Museo di storia della scienza, in Florence. They have main lenses with diameters of 16 and 26 mm.

2.6 The blurred new view through the magnifying tube

The geocentric view was, naturally, deeply rooted in society. Even in scientific circles it took time to look at the work of Copernicus as a new cosmology rather than as another clever way to calculate the almanac. Also, the system was still based on the ancient spheres and uniform circular motions. These had to be "modernized". In this period of change the names of Galileo, Tycho and Kepler shine as bright stars.

What Galileo saw through his telescope was clearly in favor of Copernicus. Jupiter's satellites revolved around Jupiter and not around the universal center, the Earth. Venus had phases like the Moon, which is possible only if it goes around the Sun. Such things were absolutely new and radical, and could not be immediately accepted by many who had only Galileo's word to believe. And not all who had the possibility to look through the small magnifying tube could see the same things as Galileo – with its blurred, shaking picture it was far from user-friendly. Any small modern pair of binoculars would give a better view – it is a good idea to find Jupiter in the sky and feel the joy of seeing its big moons Io, Europa, Ganymede, and Callisto! In any case, the discoveries by Galileo were a sensation in the Europe of the 1610's and he became a famous man.

In 1616 the doctrine of the Earth's motion was declared absurd and heretical by the Catholic Church. This step was preceded by a complicated chain of events, including the jealousies of lay professors, disputes between the fiery natured Galileo and university officials, and a plan to draw Galileo into controversy about the system of the world and the statements given in the Bible. As a result, the book of Copernicus and another book were "suspended until they are corrected". [‡] One argument by the religious community, and in those times quite a valid one, was that the Earth's motion had not been proved. This radical theory of reality had to fight on two intertwined fronts, in science and in society.

In 1632–33 the famous trial of Galileo before the tribunal of the Inquisition in Rome took place. Fortunately, the seventy years old scientist was treated well throughout. He was never put into a cell and was never tortured. The immediate reason for the trial was the *Dialogue Concerning*

[‡] A book by Foscarini, a Carmelite Father, was totally forbidden – he had tried to show that the moving Earth is in accordance with the Bible. It was not until the 1835 edition of the Index of forbidden books that Copernican ideas were no longer suppressed.

the Two Chief World Systems.

The Pope Urban VIII, who showed interest in matters celestial, had encouraged Galileo, his old friend, to write the new book on cosmology. But he told him that the Copernican system should be described only as a hypothesis (as allowed by the Decree of 1616). Nevertheless the book clearly tried to prove the Earth's motion and the trial could not be avoided.

Galileo's trial, along with those of Socrates and Bruno, has come to symbolize the struggle for freedom of thought. But it would be too simple to describe it as a collision between science and religion. The work of Copernicus was regarded as absurd and heretical by contemporary religious leaders since they had adopted the hypothesis of Ptolemy among their doctrines, resulting in an "illegal marriage of science and religion". [§]

The trial can be also seen as a sharpening of the cosmology crisis. The verdict forced Galileo to declare in public that after all the Earth did not move. But time could not be wound back.

2.7 Kepler's laws of cosmic order

Careful visual observations of the planets were made over many years by Tycho Brahe. He had obtained from the king of Denmark the island of Hveen, where he built a magnificent observatory, Uraniborg and received continuous support. All this was fairly expensive – a few percent of Denmark's national income went into "The Castle of Heavens". But the money was a good investment. It raised observations of the sky to a totally new level, even if it was built before the invention of the telescope. It also led to the next phase in the Copernican revolution, when Tycho's accurate observations were studied by Johannes Kepler. From a painstaking analysis of the observations of the planet Mars, Kepler discovered the mathematical laws of how the planets move around the Sun, in a sense completing the task that Plato had set two millennia earlier.

Tycho Brahe 1546-1601

Johannes Kepler 1571-1630

For Kepler the universe was still finite with the stars sitting on the last sphere. Whatever was beyond that sphere, inside it was our world, subject

---

[§] The situation was thus characterized from the Orthodox side by Metropolit Kirill of Smolensk and Kaliningrad, on the Russian Central Television on July 11, 1997. Of course, the scientific revolutionaries Copernicus, Kepler and Galileo, as well as Newton, were believers in God, like their contemporaries in general in Europe, and they did not regard the Bible as contradictory to science.

to the mathematical laws of nature. This was the message of Kepler, whose one foot was in the past, while the other one stretched towards modern astrophysics. He no longer believed in the material reality of the planetary spheres. Planets moved in empty space, supported by the influence of forces, and following what are called Kepler's laws of planetary motion. ¶

It was a long history how Kepler arrived to the new, revolutionary view on the motions of planets. He first tried to understand the motion of Mars following the old principle of circular motion. But after years' of struggle with circles and epicycles he finally found that only an elliptical orbit could explain the observations of Mars. It all depended on a small deviation of 8 stubborn arc minutes which Kepler could not tame with the perfect circles.

Ellipses were known from the time of Apollonius who studied these curves together with other *conic sections* (the hyperbola and the parabola). It is a curious coincidence that he was also the inventor of the epicyclic theory of planetary motions. It did not occur to him, nor anyone else before Kepler, that planets could move along ellipses. Kepler's discovery was quite unexpected. Man had finally entered the Cosmic Laboratory.

That the planets move on closed orbits, was intriguing. How can they find their way back to the same point in space and then repeat the same elongated orb? With circular motion this seemed to be easier. Kepler pictured two forces: one drives the planet along a circle and another, "magnetism", suitably makes it deviate from the circle, resulting in an ellipse.

The mystery of ellipses obtained a great solution, when Newton some fifty years after Kepler passed away, showed that one force, the universal gravitation, suffices to explain the laws of planetary motions. But this has taken us a little ahead of our story – let us look back and see how the ancient principle of no center came up again.

2.8 Nicholas of Cusa: the center is everywhere

Plotinus
205-270

Far back in the 3rd century the influential "neo-platonist" Plotinus wrote on his spiritual cosmology in *Enneads.* In a fascinating section titled The Heavenly Circuit he wrote "the heavens, by their nature, will either be

---

¶ Kepler's Laws are: I. The planets move round the Sun in a plane along elliptic orbits with the Sun occupying one focus of the ellipse. II. The radius vector from the Sun to the planet sweeps out equal areas in equal times. III. The squares of the orbital periods of the planets are proportional to the cubes of the semi-major axes of their orbits.

motionless or rotate". And without realizing it himself, he almost entered the Promised Land: "The center of the circle is distinctively a point of rest: if the circumference outside were not in motion, the universe would be no more than one vast center." But the heavens continued to revolve.

Then in the 15th century, around 1440, the German Cardinal Nicholas of Cusa stated in his philosophical treatise "Of Learned Ignorance":

> *The universe is a sphere of which the center is everywhere and the circumference is nowhere.*

Nicholas of Cusa 1401-1464

Nicholas of Cusa came to this idea from his attempt to characterize a God who is essentially incomprehensible. In fact, the context in which he states this principle, is the relativity of motion. As there cannot be any absolute rest except for God, even the Earth must have some kind of motion. Furthermore, "every man, whether he be on Earth, in the Sun, or on another planet, always has the impression that all other things are in movement whilst he himself is in a sort of immovable center". In consequence, "there will be a machina mundi whose center, so to speak, is everywhere, whose circumference is nowhere, for God is its circumference and center and He is everywhere and nowhere".

Nicholas of Cusa had a struggle in explaining that there is no absolute center of circular motions of celestial bodies:

Take, then, all these various images you have formed and merge them into one, so that the center becomes the zenith and vice versa; and your intellect, which is aided so much by the ignorance that is learning, then sees the impossibility of comprehending the world, its movement and form, for it will appear as a wheel in a wheel, a sphere in a sphere without a center or circumference anywhere...

In a sense, his vision of no center had led Nicholas of Cusa to the outskirts of the fractal world which filled him with awe and admiration.

2.9 Digges, Bruno, and the Copernican Principle

Nicholas of Cusa did not draw any definite picture of the astronomical universe. Copernicus, who was born a decade after Nicholas of Cusa passed away, did not speculate on the world beyond the distant material sphere of the stars – his great task was to cast light on the order of the things inside the sphere. But he gave a tremendous impulse to look with fresh eyes at

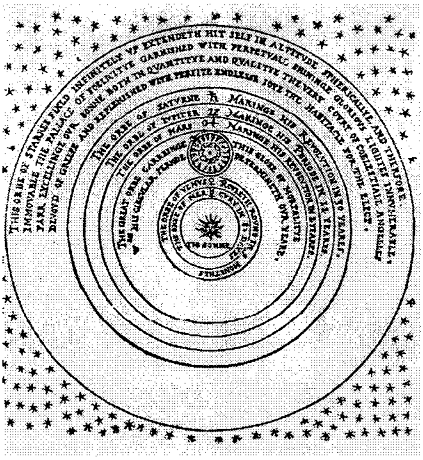


Fig. 2.4 In 1576 Thomas Digges scattered the stars outward from the surface of the outer sphere and into infinite space. He considered that the state of rest is eminently suitable for an infinite structure. Bruno gave stars the physical status of distant suns.

Thomas Digges 1543-1595

the stars. The man who realized that the immobile sphere of the stars has no function any more was born in the same year as Copernicus died. The English astronomer Thomas Digges was in 1576 inspired to publish a map of the universe where the stars were detached from their sphere and dispersed into space. He wrote: ***This orbe of starres fixed infinitely up extendeth hit self in altitude sphericallye ... with perpetuall shininge glorious lightes innumerable, farr exellinge our sonne both in quantitye and qualitye.*** But it seems that Digges still preserved a special place for the Sun in the center of the infinite stellar world. And he did not yet state that stars are just distant suns. Digges no longer regarded circular motion as the only suitable one for the heavens, he now argued that rest and the absence of movement and change is a more noble and divine state than that of the restless Earth, especially for the infinitely thick sphere of stars.

Anaxagoras c.500-428

It seems that Giordano Bruno was the first to imagine and clearly assert that stars, faint points on the sky, are really other suns. In Antiquity, Anaxagoras came very close, too. In the fifth century B.C. when Athens became the center of Greek culture, he is said to have imported philosophy and science from Ionia. Anaxagoras taught that the things of the sky and those on the earth may be similarly understood. According to later commentators he claimed that the Sun is a huge flaming stone and the

Moon is earth. Also the stars are fiery stones, set in fire by the rotating aether. We do not feel the warmth of the stars, because they are so far away from the Earth.

Giordano Bruno happened to live in London, during his non voluntary absence from the native Italy, when Digges's ideas were being discussed. We cited above from Bruno's book which he wrote in London. There he expressed emphatically the principle of uniformity of natural law in all the universe: "Give us the doctrine of the universality of Earthly laws in all the worlds and of the uniformity of cosmic matter." Here "uniformity" means similarity – the matter in the heavens is similar to the matter on Earth. This view echoed the ancient opinion of Anaxagoras and the atomists.

An ardent propagator of the idea that the universe is infinite and filled with stars and planets, Bruno formulated his cosmological principle as:

In the universe no center nor circumference exist, but the center is everywhere.

This means that in the universe all places are alike. This was in flagrant contradiction with the old cosmology where the center existed, occupied by the Earth. In modern cosmology the non-existence of any center is regarded as the natural fundament.

Sometimes one speaks about the *Copernican Principle*, when one has in mind that we are not in a special place. As Hermann Bondi wrote "This principle has become accepted by all men of science, and it is only a small step from this principle to the statement that the Earth is in a *typical position...*" In fact, Copernicus thought that the Sun is in the center (or almost so), but the Earth is not. Hence, though this is partly a question of definition, strictly speaking the Copernican Principle is not the same as the Principle of No Center, if we want to attach the name of a person to what he thought. In any case, the abandonment of the natural central position of the Earth was such a drastic change that the name Copernican Principle is quite justified along with the term Copernican Revolution. The Polish cosmologist from Cracow, Kondrad Rudnicki, has formulated it as follows, capturing both the modern spirit and the real astronomical situation in the time of Copernicus:

The universe as observed from any planet looks much the same

Giordano Bruno also argued that if the Earth were the only planet carrying living beings, this would make it a preferred place, a kind of center of the universe. This sounds familiar – modern bioastronomy is based on the universality of the laws governing both inanimate and organic Nature.

2.10 The first steps on the cosmic distance ladder

Eratosthenes c.275-195

The geographer and librarian of the Alexandrian Library, Eratosthenes made the first documented measurement of the size of the Earth, utilizing the hypothesis of its spherical form, together with observations of the Sun. His result for the length of the circumference was equal to 250 000 stadia, approximately 40 000 km, close to the true value. Thus in Antiquity the form and size of the Earth were rather well known. It took more than two millenia for the people to rise on a spacecraft far above the Earth and to see clearly by eye that it is really a globe. Eratosthenes demonstrated that it is possible to measure the size of something which you cannot see in totality. This is what modern cosmologists also attempt to do with the universe. They also are compelled to make an assumption on its "form" or geometry.

In the old days, when the size of the Earth no more was a mystery, there were serious attempts to measure the distances of the Moon and the Sun. Aristarchus and other Greek astronomers rather well knew the distance of our close companion. However, the Sun is too distant for their clever geometric methods, and its distance remained badly underestimated until the 17th century. The radius of the large sphere of the fixed stars, "the size of the world", could not be based on anything really measurable.

The small table shows the state of affairs up to Copernicus and Kepler. The distances were expressed with the Earth's radius as the natural unit:

	solar distance	distance of the stellar sphere
Aristarchus	1520	"much more distant than Sun"
Ptolemy	1210	19865
Copernicus	1142	"immense"
Kepler	3469	"immense"
Today	23500	—

The solar distance was still poorly known to Copernicus and Kepler and the size of the stellar sphere was simply unknown. That the stars do not show any swings when the Earth makes its journey around the Sun, was for them, as it was for Aristarchus, a proof that the stars must be very distant.

Copernicus made the solar distance important: all the distances in the Solar System could now be expressed with this measuring rod as a unit. One would also like to know cosmic distances in earthly units of length, used by physicists in their experiments. But it is not so obvious how to measure the distance of the Sun, so concrete in the sky. Here one must turn to Kepler's laws of planetary motion. The third law is the relation between the orbital period and the size of the orbit (the *square of period* is proportional to the *cubic of orbit size*). Knowing all periods one can draw a map of the Solar System, though the *scale* is still unknown. Then, if one can measure at least one distance from the Earth to another planet, all the distances become immediately known, among them the distance of the Sun.

Giovanni Cassini 1625-1712

From the 17th to 19th centuries the size of the Solar System was a central astronomical question rather similar to the cosmological distance scale nowadays: different methods were tried and expensive expeditions were sent to remote regions of the Earth. The first successful method, the determination of the so-called horizontal parallax of Mars, gave in 1672 that the solar distance is about 21000 Earth radii (G. Cassini, J. Flamsteed). The presently accepted value of the average solar distance, or the *astronomical unit* (AU), is about 150 million kilometers:

James Flamsteed 1646-1719

distance of the Sun = 1 AU = 149 597 892 km = 23500 Earth radii

2.11 Stars are remote suns

One result of the Copernican revolution was a new attitude towards the stars. Tycho Brahe still thought that stars have angular sizes of about one arc minute, one thirtieth of the Sun's disc. When he combined this value with the immense (though unknown) distances required by the Copernican model, he derived fantastically large true sizes for stars. This paradox of enormous stars, one objection against Copernicus, dissolved when Galileo cleverly showed that stars are much smaller than the naked eye suggests. He stretched a cord against the starry sky and noted at which distance the cord hides the star behind it. This corresponded to a size of 5 arc seconds (1/12 of an arc minute). The actual sizes of the stars are even very much smaller than this: the Earth's atmosphere smears the sharp images.

Giordano Bruno referred to stars as other suns, without real evidence to support this conjecture. Kepler made already a distinction between the physical nature of stars and planets: "to use the words of Bruno, the former are suns, the latter are moons or earths", i.e. the stars emit their own light, while the planets reflect the light of the Sun. But in order to equate the Sun and stars, one must know at least one thing: are the stars as powerful emitters of light as the dazzlingly bright Sun? This requires knowledge of stellar distances. [‖]

Galileo did not share the view of Kepler that the stars are on a thin sphere. Some of them may be two or three times more distant than others, and nearby stars should show regular yearly movements against the background formed by distant stars. In the 17th century, astronomers started searches for such swings of the stars. This change of position or *parallax* would tell the distance of the star. It would be a proof for the Earth's motion, making the search for the parallax also a cosmological test.

To see the phenomenon of parallax, just look at your finger against a background of wall paper by only one eye. Then close the eye and open the other one. You will see how the finger shifts. The shift gets larger when you move the finger closer to your eyes. Similarly Thales of Miletus could measure by triangulation the distance of a ship far from the shore, without stretching a measuring tape to the vessel.

Friedrich
Wilhelm
Bessel
1784-1846

The shifts of even the nearest stars are very small and hard to detect even with the widely separated astronomer's "eyes" (the size of the Earth's orbit). Only in 1838 Friedrich Bessel had success with a star in the constellation of Cygnus. Its parallax of 0.3 second of arc puts it at a distance of 650 000 times the Sun–Earth distance. This proved that stars are so remote that in order to be visible in the sky, they must be pouring out as much or even more light than our Sun.

Parallaxes give astronomy one of its odd units, *parsec*: a star is at a distance of one parsec, if its annual parallax angle is one second of arc. This definition is cleverly hidden in the name of the unit (*par*allax = 1 arc*sec*). How much is one parsec in terms of the Earth's distance to the Sun (or

---

[‖] A star's ***luminosity*** tells how much light energy it sends every second into space. The energy flies in every direction at the speed of light. At a distance R from the star, the energy is evenly distributed on the surface of a sphere with radius R. So the ***flux*** f of the light at a distance R is equal to luminosity divided by the area $4\pi R^2$ on which it falls: $f = \frac{L}{4\pi R^2}$ This is a most important formula in astronomy. The flux f is what the astronomer can measure, luminosity L is a property of the star. If the distance of a star and its flux are measured, one can calculate its luminosity.

AU)? The answer is 1 pc = 206265 AU. All known stars are more distant than 1 parsec, so the deflection in the sky (parallax) is always less than one second of arc. The restless air, which spreads the image of a star into a fuzzy dot, limits parallax measurements to stars closer than 50 parsecs.

Nowadays stars beyond 50 pc may be reached when one rises above the atmosphere, where the stars look sharp. Thus the HIPPARCOS satellite could in the 1990's measure ten times more distant stars, a total of 100 000. Even 500 pc is only a small fraction of the size of our Milky Way galaxy. In the 2010's, GAIA will reach 100 000 pc! In order to push the meter stick deeper into space, to distant galaxies, other methods must be used.

2.12 Understanding the new cosmic order

Ideas which at first impression seem absurd, like a moving Earth, may later come to be accepted as perfectly reasonable, when viewed through the lenses of the new generation, in possession of new arguments, theories, and observations. And so the previously sound notion of a rotating sky and resting Earth became absurd... Thus one reads from the first astronomy book published in Finnish and intended for the general public, especially for farmers ("A short story about the Heaven and the Earth, the Moon and the Stars etc." by an anonymous author in 1836): "That the earth should stand, but our sun and all the planets and other suns which we call stars should revolve around our earth, is such a stupid and useless belief that even the simplest man cannot but laugh, knowing how far away the sun is from us."

The new world picture did not come to be as natural as its predecessor until the accumulation of evidence for the Earth's motion (e.g. stellar parallaxes), and the invention of a new theoretical framework, Newtonian mechanics. Aristotle did not know of free inertial motion, the new basis of mechanics. Galileo had inferred its existence from experiments with balls rolling down inclined plates. In his mind Isaac Newton got rid of the mundane conditions hampered by friction, and thought of a space where a particle will keep its constant velocity if it is not acted on by any force. Then what was the meaning of force in the new mechanics?

Isaac
Newton
1642-1727

The presence of a force is seen as a change of the velocity of the particle. It experiences an acceleration as expressed in Newton's Law of Motion: *Force = mass × acceleration.* Newton's law was in sharp contrast to the old mechanics. It says that when the force is turned off, the body still moves, though with a constant velocity. According to the old concept of force the body should then stop, a view already criticized by Oresme.

The concepts of inertial motion, forces, and accelerations imagined by Newton and described in his 1687 *Principia* provided us with a splendid model of the physical world (though, as all models, it has its limitations). The physical phenomena occur in an uncorruptable stage, absolute space, which is infinite and Euclidean. Our Solar System is as if a sphere cut from the big world, a fair sample where all the universal laws function and can be studied. **

Robert Hooke 1635-1703

As a planet does not move along a straight line, there must be a force acting on it. What is the nature of this force? Before Newton, Robert Hooke had proposed that planetary motion is caused by a force *towards* the central body, the Sun. This force makes the planet fall towards the center, deflecting it from its rectilinear motion. In a letter Hooke asked Newton what properties such a trajectory would have.

Newton did not write an answer. Still, Hooke's question perhaps was the impulse which led a few years later to the Law of Gravitation which describes the force acting between two bodies as proportional to the product of their masses divided by the square of the distance between them. With this law of gravity and his other laws of motion Newton explained why a planet dances around the Sun as it does, both "floating" in absolute space.

The law of gravity was another great outcome of the process pushed in motion by the quiet canon of Frauenburg. The old mechanics stood helplessly before the new elliptical orbits. Newton revealed the secret as to why a planet can travel along one and the same ellipse forever: it is the inverse square law of the gravity force. In fact, even a slightly different law would inevitably lead to a more complex orbit. Such a "rotating ellipse" (that of the planet Mercury) was in store for the future.

**Newton's three laws of motion are: I. If the total force acting on a body is equal to zero, then the body is either at rest or moves inertially i.e. along a straight line without acceleration. II. The force acting on a body is the product of the mass and the acceleration. III. If a body A exerts a force on another body B, then the body B will exert an equal, but oppositely directed force on A. The law of gravity is the fourth law of Newtonian mechanics.

2.13 The triumph of Newton's universal gravity

The 300 years of gravitation were celebrated in 1987, and not only for historical reasons: Newton's mechanics is still to-day excellent for applications in physics and astronomy. The orbits of spacecraft are safely calculated from the old laws. Newton's theory is a prototype and Queen of scientific theories. She still flourishes, though knows Her limits.

The spectacular prediction of the arrival of Halley's comet in 1758 brought much publicity for the theory. Another sensation came when astronomers predicted the existence of an unknown planet, Neptune, which was found in 1846. The planet Uranus – thought to be the most distant one – did not move as a good planet should. It deviated slowly, but surely from its calculated orbit, so that in 1845 it was 2 arc minutes away from its expected place in the sky. This was a large error. (Even for Kepler using less accurate *visual* observations, a few arc minutes was a signal of alarm in his struggle with Mars.) Something was wrong with the picture of our Solar System. An unknown mass must lure Uranus from its regular course.

Urbain Le Verrier from France and John Adams from England calculated, independently of each other, where such a planet should be in order to cause the misbehavior of Uranus. The final discovery by telescope was made by Johan Galle in Germany. Neptune wandered in the sky only about 1 degree away from the predicted spot. Needless to say, a long dispute prevailed in British and French press about the honor of the discovery ...[††]

The prediction of the arrival of Halley's comet and the discovery of Neptune were the offspring of the new science of *celestial mechanics.* It can predict planetary positions in the sky a thousand times more accurately than the recipes of Ptolemy or Copernicus. This suggests that looking at Nature through the eyes of Newton's theory, we now understand it better. In fact, Newton's mechanics does much more than the old models. It is a physical theory that may be applied to a plethora of situations. When it is said that with our science we have achieved power over Nature, this means the "magic" of knowing, say, in which direction and when one must launch a rocket in order to reach a planet.

[††] John Adams (1819-1892) and Urbain Le Verrier (1811-1877) retained mutual respect – it even happened that the former as President of the Royal Astronomical Society presented the latter, Director of the Paris Observatory, with a gold medal.

2.14 Just add one particle more...

Amidst all the admiration for Newton's mechanics, one must admit that the motion of bodies can be accurately predicted only in very simple situations. One can tell how two particles revolve around each other. However, just add one particle more and one encounters the enormously more complicated *problem of three bodies.* One cannot say exactly beforehand, how three gravitating bodies, started at known positions and with known velocities, will be moving at some moment in the distant future. Even though each body, feeling the gravity of the other two, faithfully follows the laws of motion, three bodies form a *chaotic system.* In practice, its long-term behavior can be foretold only in a statistical sense. A triple system decays sooner or later, when one of the masses escapes. However, one cannot say when this will happen. In the end of the 19th century, Henri Poincaré was the first to study the chaotic behavior of dynamic systems. In doing this he discovered mathematical objects which are now called strange attractors and which are related to fractals.

* * *

Pierre Laplace 1749-1827

The new physics inspired great optimism on the chances of drawing an ultimate picture of the world. Pierre de Laplace envisioned the universe as a huge, totally predictable clockwork, and wrote the memorable words:

An intelligence which, for a given instant, knew all the forces by which nature is animated and the respected positions of the beings which impose it, and which besides was large enough to submit these data to analysis, would embrace in the same formula the motions of the largest bodies of the universe and those of the lightest atom: nothing would be uncertain to it, and the future as well as the past would be present to its eyes.

But it became gradually clear that the universe as a whole is not simple to grasp. This is so even for the almighty Newtonian intelligence, even if one, recalling the problem already with three bodies, waives the dreams of a clockwork and a theory of everything. If one just contemplates the infinite universe, uncomfortable paradoxes start creeping in.

Chapter 3

The Paradoxal Universe of Sir Isaac

Infinite absolute space, imagined by Isaac Newton as a stage for physical phenomena, became also a part of a brave attempt of the mind to grasp the universe as a whole within one mathematical frame. Astronomical observations and celestial mechanics joined together, and offered the ground for the study of the universe for two centuries. Newton himself felt that "I seem to have been only like a boy playing on the sea-shore, ..., whilest the great ocean of truth lay all undiscovered before me". But to many of his predecessors it may have seemed that the final truth was already found and all future science would have pedagogical value only.

But the victory was not complete, and over the years it was realized that an infinite universe, so tempting after the closed medieval globe with its mysterious outer border, is not simple at all, but hides deep secrets.

* * *

3.1 Structure of the Heavens

The Copernican revolution was truly far-reaching. Paradoxally, one had first to put the Sun into the center of the universe. Only then it became possible to get rid of the sphere of the stars and let stars wander in space. This dramatic turn motivated and eased the mental effort to detach also the Sun from its privileged position and to give it the status of an ordinary star. The second courageous step into the deep universe from the Solar System to the realm of stars was made by Bruno and other thinkers, before any evidence that faint stars are really very distant suns.

In Principia, Newton did not care much about stars. He just said that they are so distant that their gravity may be ignored when one calculates how the planets orbit around the Sun. Only later he got interested in stars as constituents of the large universe. This happened when in 1692 Richard Bentley, a young chaplain (later Professor of Theology in Cambridge), needed advice in matters cosmological for a sermon. *

Richard Bentley 1662-1742

He asked the opinion of Newton about the behavior of matter spread evenly in space. What would happen to it under the effect of its own gravity? Bentley also posed other knotty questions: What is the lifetime of the Sun and the stars? What is the difference between the Sun and planets? What is the agent of gravity from one body to another? This correspondence had a great influence on Newton's thoughts on cosmology, a topic that he had neglected in his younger age.

It was for Newton a deep puzzle to the end of his life, how the stars can remain at rest even though they are pulling each other. The fixed stars had always been a symbol of unchangeability, also for Newton. The first detection of a star's movement was made only by Edmond Halley in 1718. Sirius, the brightest star, had moved amidst the other stars about half a degree (the size of full moon), since the time of Ptolemy.

Edmond Halley 1656-1742

Newton liked to think that the immobility of the stars is dictated by the structure of the stellar universe. Any one star is pulled by other stars from all sides, so perhaps these forces cancel each other. It seemed to him that behind this stable arrangement must be a Divine hand. It also helps the matters, if the stars are far away from each other.

Only parallaxes of stars, in the 1830's, gave direct geometric evidence of their distances. Though, there had been "educated guesses" based on their faintness, and Newton knew about the vast interstellar distances. The Scottish mathematician James Gregory had suggested in 1668 the method of "standard candles". If all stars are like our Sun, then more distant stars look fainter than nearby ones, and one can infer their distances. But it is difficult to compare the light of the dazzling Sun to that of a faint star (considering they do not particularly often appear together in the sky!). Gregory's method utilized a planet as an intermediate step (the planet's

James Gregory 1638-1675

---

*Richard Bentley is the same man whose words on the (only) apparent irregularity in Nature were cited in the beginning of the 2nd chapter in Benoit Mandelbrot's *The Fractal Geometry of Nature*: "All pulchritude is relative... We ought not... to believe that... the mountains are out of shape, because they are not pyramids or cones; nor that the stars are unskillfully placed, because they are not all situated at uniform distance."

brightness depends on the reflected light of the Sun). Thus Newton could calculate the distance to Sirius, with the help of Saturn. It turned out to be about 1 million times more distant than the Sun. This is two times too large, but gave the correct idea of the enormous distances.

Newton's view of stars is encapsulated in his words from a manuscript: *The Sun is a fixt star & the fixt stars are scattered throughout all the heavens at very great distances from one another & rest in their several regions being great round bodies vehemently hot & lucid & by reason of the great quantity of their matter they are endued with a very strong gravitating power.*

3.2 Newton's Cosmology in a Nutshell

Newton began to think about the structure of the heavens in his old age and his cosmological views remained fragmentary. However, many key questions of cosmology were raised by him. Though he did not leave any complete cosmological system, we may recognize three major elements:

* *The cosmological Principle* – there are no preferred points and the stars are uniformly distributed in infinite space.
* *Observations* of the positions of stars from ancient times showed to Newton that the stars are very far away and do not move.
* *Theory*, as Newton's mechanics and gravity, valid everywhere.

God created the universe to be infinite, and arranged the eternal stars uniformly in their fixed positions in an absolute space. Such a universe, infinite in space and time, is intuitively attractive and forms a natural start for cosmological thinking. But if one sits down, lights one's pipe, and seriously tries to imagine such a world, strange questions spring to mind.

3.3 Cosmological paradox

What is a paradox? This term comes from the Greek 'paradokson', unbelievable or beyond what is thought. Originally it was a statement seemingly self-contradictory or absurd but (and this is important) in reality expressing a possible truth. Paradoxes of that kind may relate to strange predictions of established theories, such as special relativity and quantum mechanics. They do not prove that the theory is incorrect, but they dramatize the points where it deviates from the old view. An example is the Twin Paradox, appearing when velocities are very high (Ch.5).

A cosmological paradox is something else; it is a consequence of far-reaching cosmological theories, a statement resulting from basic cosmological assumptions, which contradicts plain observational facts. Here the contradiction is genuine, and it may be so striking that one cannot but admit that something is badly wrong with the theory.

3.4 Why do we not feel an infinite gravity force?

If the stars fill infinite space, then we are influenced by an infinite number of masses from all directions. Then where are the signs of tremendous forces tearing us apart? Why are the stars moving so majestically slowly? How can infinities cancel each other and result in zero?

Different aspects of this gravity paradox have been emphasized by its expositors. Hugo Seeliger, a German astronomer, intended to show that "Newton's law when applied to the infinitely large universe, leads to unsurpassable difficulties and unsolvable contradictions, when one assumes that the matter dispersed into the universe is of infinite amount".

Hugo Seeliger 1849-1924

Seeliger's argument may be described rather simply. Instead of thinking at once of infinite space filled with stars, let us start with a spherical uniform star cloud in empty space. What is the force that affects a star inside the cloud? Newton's mechanics tells that the force pulls the star towards the center. Its strength is directly proportional to the distance (so a particle two times farther from the center, but still inside the spherical cloud is attracted by a force two times stronger). [†]

Imagine that more and more layers of stars are added to the star cloud. In such a huge cloud there is room for a very large sphere with its center coinciding with that of the cloud. The force acting on a star at the surface of the embedded sphere is also very large.

Now we make the critical (and rather demanding) step and imagine that the cloud is infinitely large. One may select its "center" to be anywhere, and the embedded sphere, with a star on its surface, may be as large as you like. If one takes the center to be at an infinitely large distance, the resulting

---

[†] The force F is proportional to M/R^2 and the mass M of the sphere depends on the cubic of the radius R, hence the force F is directly proportional to the distance. If R increases without limit, then also the force F strives towards infinity.

force is infinitely large. However, one is as well free to select the middle point quite close to the star, which leads to a minute force in the same direction. There seems to be no unique Newtonian gravity force inside an infinite matter distribution, the force may be either infinitely large or very feeble, depending on the choice of the line of argument. Often especially this ambiguity is called the Newtonian gravity paradox. [‡]

3.5 How to tame the infinite gravity?

One might say that in reality the star is pulled by forces from all directions, and these forces cancel each other, both the finite and infinite ones. But is it permissible to subtract two infinities and obtain zero? In his first cosmological letter to Bentley, Newton stated that "if the Matter was evenly disposed throughout an infinite Space it could never convene into one Mass", but in the second letter he wrote that there is a difficulty in Bentley's argument that "every Particle of Matter in an infinite Space, has an infinite Quantity of Matter on all Sides, and by consequence an infinite Attraction every way, and therefore must rest in Equilibrium, because all Infinites are equal". Newton wished to make clear that infinite is not an ordinary quantity with which one can make routine calculations, say, subtracting one infinite from another. However, in specific situations this is possible. He has in mind a perfectly isotropic distribution of matter, which even if extending into infinity results in a complete cancellation of contrary acting infinite forces on a body. And "if to either of these Forces you add any new finite attracting Force, that new Force, how little fo[r]ever, will destroy their Equilibrium, and put the Body into the same Motion into which it would put it were those two contrary equal Forces but finite, or even none at all; so in this Case the two equal Infinities by the Addition of a Finite to either of them, become unequal in our ways of Reckoning".

A view resembling Newton's considerations was expressed by Svante Arrhenius in 1908. The Swedish physicist noted that for any single star the total gravity force from all other stars may be divided into two parts. One

Svante Arrhenius 1859-1927

[‡] One aspect of the "unsurpassable difficulties" comes to light when one considers the gravitational potential in Poisson's equation $\nabla^2\phi = -4\pi G\rho$, in which ρ is the matter density. In a uniform world the potential, as all other physical quantities, should be constant. But then the equation implies $\rho = 0$, i.e. it excludes a uniform non-zero density! In order to resolve this paradox one has to use relativistic equations for gravity.

part is caused by the neighboring stars and it fluctuates from one place to another. The second part is due to all other stars, up to infinity. It is equal to zero, because of the symmetry of all directions. This conclusion was mathematically confirmed by Jan Holtsmark in 1919. The Norwegian physicist assumed that particles have been scattered randomly in space, so that their distribution is on the average uniform, but with small differences in the number of particles from place to place (so-called Poisson law). Holtsmark's formula says that in the infinite universe of such interacting particles a finite average force acts on any particle. This force comes from the nearest neighbors, as Arrhenius had proposed. The importance of Holtsmark's result for the dynamics of stellar systems has been untiringly advocated by Tateos Agekyan of St.Petersburg University, a founder of the study of motions in star and galaxy clusters. [§]

Jan Holtsmark 1894-1975

Rather than reject infinite matter, Seeliger felt compelled to propose that Newton's law of gravitation is not universal. If the gravity force decreases with increasing distance faster than what the law says (having a finite range of influence), then there are no infinities that should be subtracted from each other and no need for a perfectly isotropic world. It is also interesting to read what Bruno, the advocate of infinity, wrote:

Question: But what do you say about the interaction between the finite and infinite matter, like e.g. between the Earth, a cold body, and the sky and innumerous stars? Do you not think that it should follow, as Aristotle maintains, that the infinity would absorb and destroy the finite?

Answer: Not at all... For while the material force extends and spreads across the infinite space, the infinite matter would not act on the finite one with an infinite force, but only by such a force which can be radiated from a limited number of parts and from certain distances against the finite body, it cannot influence by all its parts everywhere, but only by its nearest parts.

Bruno says that bodies cannot influence each other beyond a finite distance. If liberally interpreted in modern terms, he was actually also proposing finite range of gravity as a solution to the paradox of infinite forces.

One should also note that in Newton's theory gravity is an action-at-

---

[§] Holtsmark considered electric particles of the same sign, producing repulsion. But as the electric force has the same $1/r^2$ dependence as gravity force, his analysis is valid also for gravity, as S. Chandrasekhar later showed. Holtsmark's force has an infinite dispersion around the finite average value, because of close encounters between point-like masses. A pupil of Agekyan, Irina Petrovskaya (1938-1999) generalized in 1986 the Holtsmark distribution for finite-sized stars. Then the dispersion becomes finite.

distance. This means that gravity propagates infinitely rapidly and cannot be shielded or absorbed. These features in fact lead to the gravity paradox, as the physicists say, on the level of the equation for the gravitational potential. Without modern relativistic and quantum physics the gravity force in infinite space remains paradoxical.

3.6 If, however, a uniform infinite cloud of stars exists, why has it not collapsed?

Though the visible stars are not so helpful in this respect, Newton preferred to assume that the stars in general are quite uniformly distributed and, furthermore, at rest as if initially hammered by a nail in absolute space (the God who created the absolute space naturally could achieve this feat). Then, what happens if the nails are quietly detached from the space?

In a letter to Richard Bentley, Newton wrote:

> ... *if the Matter of our Sun and Planets and all the Matter of the Universe, were evenly scattered throughout all the Heavens, and every Particle had an innate Gravity towards all the rest, and the whole Space throughout which this Matter was scattered, was but finite; the Matter on the outside of this Space would, by its Gravity tend towards all the Matter on the inside, and by consequence fall down into the middle of the whole Space, and there compose one great spherical Mass. But if the Matter was evenly disposed throughout an infinite Space it could never convene into one Mass, but some of it would convene into one Mass and some into another, so as to make an infinite Number of great Masses, scattered at great Distances from one to another throughout all that infinite Space. And thus might the Sun and fixt Stars be formed, supposing the Matter were of a lucid Nature.*

Here Newton makes a difference between the behaviors of finite and infinite clouds of matter. A finite cloud, originally at rest, has always a center into which it will inevitably collapse. The middle of the cloud differs radically from the other points: there the gravitational forces from other parts of this cloud cancel each other. However, if one allows an infinitely large cloud, then a miracle occurs: there is no longer any single center and one wonders how any collapse could start in the first place.

As stars exist and not one collapsed mass only, Newton concluded that our universe must be infinite. Furthermore, he outlined a process of star formation in it. Matter uniformly spread in an infinite universe is unstable:

condensations tend to form from small density seeds. ¶ Under the inherent gravity the initially uniform and monotonous scenery starts to "live its own life", evolving to a complex, structured landscape.

Newton compared the unavoidable instability with the enormous difficulty of making an infinite number of needles stand accurately poised upon their points on a hard, infinite surface. He, however, admitted that it is possible, at least by a divine power. By the way, Newton's needle has become a popular example in the physics of chaos: even the tiny atomic vibrations would make a real needle fall in about 4 seconds.

3.7 Why is the night sky so dark?

In order to avoid the collapse of the stellar universe into one clump, Newton suggested that it is infinitely large. But alas, this leads to another paradox. In the universe uniformly filled by stars, the whole celestial vault should blaze as brightly as the surface of the Sun.

Heinrich Olbers 1758-1840

This inference is usually known as *Olbers's paradox*, named after a German physician and astronomer. Its history goes back to Kepler who still thought that the stars lie in the celestial sphere. It is curious that Newton did not notice this problem in his own infinite cosmos, though the calculations that he wrote down, show that he was but one step from the paradox.

What happens in the night sky of an infinite stellar universe? Think about a forest. If it is deep, you cannot see all the way through it. In any direction your line of sight hits a trunk of a tree, before reaching the limits of the wood. This is because the trunks are not thin mathematical lines. So, if the trees grow 10 m apart from each other and are 1 m thick, you cannot see much further than 100 m. A similar thing happens in the universe uniformly filled with stars. No matter how large is a typical distance between the stars, in every direction the line of sight sooner or later hits a stellar disk. Stars are not mathematical points.

Stars are typically separated by 3 10^{13} km. The size of an average star (say, our Sun) is around 1000 000 km. With some effort, one may calculate the distance after which all the sky would become covered by stellar disks,

---

¶ James Jeans in 1902 found a criterion for the onset of instability in gravitating uniform matter. In regions larger than a critical size $R_J = v/(G\rho)^{1/2}$ structures start to grow with a characteristic time $\tau = 1/(G\rho)^{1/2}$. The density of the uniform matter is ρ, and v is the initial chaotic velocity of the particles (for a thermal gas $v^2 \approx kT/m$).

Fig. 3.1 The sparseness of the stellar forest in the Milky Way allows one to see the Andromeda nebula. First observed by the Arab astronomer Al-Sufi (903-986), this nearby Island Universe is visible to the naked eye as a faint fuzzy dot just above the constellation of Pegasus. For Newton, the peaceful stellar scenery whispered that the universe was infinitely large and filled with stars. But for a few other minds, it was not the glimmering stars but rather the darkness of night which was a cosmological mystery.

making it as intolerably bright, as the surface of our Sun. This distance is huge, but finite, about 10^{15} times the distance to the closest star. As the size of our Milky Way galaxy is only 10^4 such length units, we can easily look through the "stellar forest" of the Milky Way and see other galaxies.

As the night sky is far fainter than the Sun, some assumptions of Newton's cosmology must be wrong. Many thinkers have tried to understand that innocent looking phenomenon – darkness at night. Kepler saw the dark sky as evidence for a thin stellar sphere. Otto von Guericke, known for his air pump and experiments with vacuum, regarded that the stars form an immense, though finite island in the infinite, otherwise empty universe. He thought that one stares between the stars at the dark and endless void.

Otto von Guericke 1602-1686

Or perhaps the stellar universe is infinite, but the visible universe is finite? Thus Olbers thought that starlight is absorbed when it traverses the vastness of space – a cosmic fog limits the visibility: "The Almighty with benevolent wisdom has created a universe of great yet not quite perfect transparency, and has thereby restricted the range of vision to a limited part of infinite space." Though absorption of starlight is an attractive explanation of the dark sky, it was later realized to be too simple.

A better way to restrict the visible universe is to say that the stars have not existed always, but were switched on some finite time ago: when we look far in space, we also look back in time, at the ancient era when stars were not yet shining. This idea was discussed by William Thomson, professor at Glasgow University for 53 years and better known as Lord Kelvin ('Kelvin' derives from the river flowing near his university). If the stars have been shining no longer than one hundred million years, then the radius of the visible universe is at most 100 million light years. Ages of this order appeared in a theory, popular at that time, which ascribed the hotness of the Sun to the energy released when it gradually contracts under its own gravity. We note in advance that such a solution for Olbers's paradox is offered by big bang cosmology. The paradox also gave early motivation to consider hierarchic organizations of stars.

William Thomson 1824-1907

3.8 The riddle of the shining stars

The law of conservation of energy was discovered in the 19th century by a few ardent students of the nature of heat. In a rudimentary form it had been expressed already in 1748 by the Russian scholar Mihail Lomonosov, who in a letter to Leonard Euler wrote: "...all changes in Nature happen so that if something is added to one thing, then this is taken away from an other thing. So, as much as matter is added to some body, as much of it is lost by another, ...As this is a general law of Nature, it applies also to the laws of motion: a body which by a bump puts another body in motion, equally much looses of its own motion..." Or as later Antoine-Laurent Lavoisier put it: "Rien ne se perd, rien ne se crée, tout se transforme."

Mihail Lomonosov 1711-1765

Antoine-Laurent Lavoisier 1743-1794

The first to state the conservation of energy in all generality, in 1842, was a German physician. Julius Mayer's start with physics was unusual. While working as a ship's surgeon on a voyage to Java, he noticed that the blood of the sailors was redder than in his cooler home country. He related this observation to the theory proposed by Lavoisier that body heat is generated by a burning process for which the blood gives oxygen. The blood was redder because less burning and oxygen was needed in the tropics. This inspired Mayer think about the relation between heat and mechanical work performed by muscles. He reasoned that heat and work are two forms

Julius Mayer 1812-1878

of energy. There are different kinds of energies. Their total sum is conserved in a physical process, and ultimately, in the whole universe. [||]

If stars shine thanks to the finite store of energy hidden in the mass they contain, and if the universe has an infinite age, then all the fuel should have been consumed! The sky would be coal-black day and night. Our modern knowledge says that the mass in a star can give out no more energy than allowed by Einstein's famous formula. A star pouring out energy, cannot live longer than the life time given by the total energy divided by emission power. For our Sun this is 20000 billion years, respectable but finite. In an eternal universe the stars would be only a flash of light. [**]

3.9 What has saved us from the ultimate heat death?

Still another problem with an eternal world is the awe-inspiring heat death. Thermodynamics states that when a physical process goes on without interaction with the external world the *entropy* of such a closed system always increases. This quantity characterizes the level of order: more entropy means more chaos. Roughly speaking, entropy is the number of separate units in a system: what is whole in the beginning, tends to break into pieces in the end. Things left on their own gradually disrupt into a pile of dust.

The entropy law is familiar. When you drop a cup, the entropy grows, and it is not a good idea to wait until the pieces by themselves ascend from the floor and form the cup again. Or heat always goes from a hot body to a cooler place, allowing us to warm the home by a stove. This means that in real life there is an arrow of time, even though in mechanics the direction of time does not exist: colliding billiard balls look the same if the film is run forward or backward, but people never get younger.

Hermann von Helmholtz 1821-1894

In 1852 Hermann von Helmholtz held in Königsberg a lecture where he presented the idea of the heat death of the universe. The increasing entropy means that there is a continuous degradation of energy until all motions in the universe would have ceased, i.e. the heat death has arrived.

[||] Mayer's ideas were published in private pamphlets and were not especially welcomed by his contemporaries. The skilfull experiments by James Joule (1818-89) were needed for the scientific community to accept the law of energy conservation.

[**] The maximum life time of a star can be calculated from its internal energy reservoir $M_* c^2 \approx M_\odot c^2 = 1.8 \times 10^{54}$ ergs. If the star loses its energy at the rate $L_* \approx L_\odot = 4 \times 10^{33}$ erg/sec, then its age never exceeds $\tau_{max} = M_* c^2 / L_* \approx 5 \times 10^{20}$ sec.

Ludwig Boltzman 1844-1906

Ludwig Boltzman, the founder of statistical mechanics, expressed the law of entropy as follows: A system consisting of a large number of particles strives towards a more probable state. You are likely to agree that all the air of your cabinet (about 10^{24} molecules) has never gathered in one corner, leaving the part where you are sitting and contemplating, empty. Such a distribution of the air molecules is extremely improbable. The most probable state is the thermodynamic equilibrium, when all the molecules are distributed uniformly in space. In the eternal universe of Newton, there is ample time to achieve such a state of equilibrium with no stars, no planets, no complex structures, no life...

In order to explain why Nature has been gentle to us, Boltzman suggested that what we see is actually a gigantic fluctuation in the world of molecules. Clearly, the probability that such a deviation from uniformity occurs accidentally, is very small. However, one may speculate as the Greek atomists did: enough time, enough space and anything may happen. But it should be warned that the heat death is a bold extrapolation to the big universe. Boltzman did not take into account that gravity and other forces between his particles leads to the growth of cosmic structures.

One natural reaction to the paradox is again to question the infinite age: since the universe is not yet in thermal equilibrium it must have existed only for a finite time. This was regarded as physical evidence for the creation event already before the big bang cosmology was known.

* * *

The paradoxes of Newton's infinite and eternal cosmos have inspired cosmologists up to our day. Indeed, the horrible predictions for the entire universe, from infinite gravity and blazing night to the heat death, nevertheless sprang from a theory which worked wonderfully in the Solar System. Although the old physics could not be extended arbitrarily deep into space, one element of Newton's world, the cosmological principle of the uniform distribution of matter, was kept alive by Einstein and is retained in modern big bang cosmology. But there is another view of the cosmic arrangement of matter which has been entertained by thinkers since the time of Newton. While this has been partly a response to the Newtonian paradoxes, the origin of this view is rooted in tempting analogies with the Solar System and in the notion of self-similarity. Faint nebulous spots in the starry sky lent wings to philosophical excursions into cosmic hierarchies.

Chapter 4

The dream of a hierarchical world: protofractals

The 18th century was the time of the Enlightenment, when every observed phenomenon, from the human body to the clockwork of the universe, became amenable to scientific scrutiny and explanation. By the "siècle des lumières" the luminaries of the night sky had scattered into the depths of space. But what is the shining band of the Milky Way?

One clear night Johann Lambert stared at the sky, hoping to see some order: "Last night I again inspected the starry sky, as I have never been able to find any definite *symmetry* in its appearance. Once again in vain. Then I noticed that the stars of first, second, and third magnitude are very unevenly distributed, somewhere densely together, while elsewhere in the sky there are large empty spaces hardly containing a few stars of sixth magnitude. So, I thought, would for us appear the solar *system*, if we could see all the planets and comets at the same time."

What a wonderful night! Structure was discerned in the sky, telling about the stars in *space*. Maybe they form a large, flattened system similar to our swarm of planets around the Sun? At the time of Lambert, the third dimension, depth in space, achieved a new fascination.

* * *

4.1 Stars and nebulae

Isaac Newton regarded it as highly desirable that stars be uniformly distributed in infinite space. Otherwise it was hard to understand the peaceful scenery where we are living, not having collapsed into a huge mass nor feeling intolerable forces emanating from the abysses of space. Uniform matter was needed to cancel the gravity force. But the real stars in the sky posed a problem. The most conspicuous feature of the stellar vault is the Milky Way. Galileo had seen through his magnifying tube that its pearly glow comes from innumerable faint stars.

Newton tried to show that if one goes deeper and deeper into space, i.e. looks at fainter stars, then uniformity is encountered. Modern cosmologists have a similar hope: looking at large enough parts of the universe, one should finally see a uniform world of galaxies (the new basic units of matter in the large). Newton's attempts remained inconclusive. This is not surprising. Nearby stars actually form a large flattened cloud. Newton also did not know how the magnitudes of stars depend on their distances, and hence how many faint stars there should be, in comparison with the bright ones, if stars are uniformly dispersed in space.

Charles Messier 1730-1817

By the time of Newton, keen observers had begun to look with their telescopes also at other things than stars. In the 18th century, comet hunting became popular. However, not all diffuse blobs in the sky are comets. There are a lot of immobile nebulae which annoyed the comet hunters because one easily confused them with the slow moving comets. Charles Messier decided to collect a useful catalog of such nuisances. During his career Messier discovered some twenty new comets and was invited to become a member of the French Academy. Though, he is now remembered less for his misty comets than for his nebula catalog which contained 103 objects, their positions in the sky and descriptions of their appearance. Astronomers still refer to M31 (or Messier 31) when they speak about the Andromeda nebula. Also many other beauties of the night are known by their Messier–numbers. Though Messier was not at all interested in nebulae *per se*, some minds started to wonder what they are.

4.2 Emanuel Swedenborg

Emanuel Swedenborg 1688-1772

Emanuel Swedenborg was a Swedish scientist and visionary. His father was a professor at Uppsala University and later the bishop of Skara. The young Emanuel studied languages and natural sciences and after graduation in 1709 he embarked on travels in Europe. There he became acquainted with the scientific circles of the time and built a reputation as a skilled engine constructor. After returning home, he worked as the editor of Sweden's first

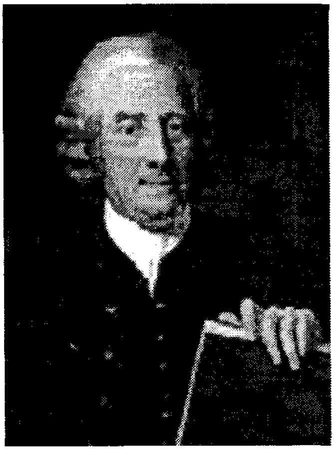


Fig. 4.1 Emanuel Swedenborg expressed his principle of self-similarity in different ways. One was "Nature seems to admire her own charms".

scientific periodical and was appointed, at the age of 28, assessor extraordinar in the College of Mines. This high position he held for over thirty years. He was also the corresponding member of the Imperial Academy of Sciences of St. Petersburg. The remains of this unusual man rest in the cathedral of Uppsala.

Swedenborg was a productive thinker and writer, who discussed practically all fields of science of his time and wrote treatises in physiology, zoology, chemistry, geology, mineralogy, physics, and astronomy. * He appears in our book because of his ideas on the structure of the universe. In 1734, in his *Principia*, Swedenborg put forward the remarkable views of self-similarity and cosmic hierarchy: elementary particles form celestial bodies which form systems which in their turn may be elements of systems higher in the hierarchy, and so on. This view reflected his general opinion that everything in the world is constructed according to a common plan.

Swedenborg's work in science and engineering has often been overshadowed by the last three decades of his life when he studied things which are rather strange to us: he became a clairvoyant, had conversations with spirits and angels, and wrote books on the spiritual world and theology.

---

*For instance, it is nowadays recognized that Swedenborg's writings on the brain and sense organs were "observationally more accurate and theoretically more profound than those of any other eighteenth-century scholar" (cited from *Nature* 394, p. 144 (1998)).

But remember that the great alchemist Newton devoted a majority of his time on other than what we now call scientific activities – the age when Newton and Swedenborg lived was different from ours...

4.3 Cartesian physics

René Descartes 1596-1650

Swedenborg was much influenced by the ideas of René Descartes. This French mathematician, physicist, and philosopher paved the way for modern science at a time when it was still not always healthy to speak about Copernicus. Descartes had a considerable family fortune, which allowed him to combine scientific pursuits with travel around Europe. He lived and died outside his native country, residing twenty years in the more peaceful and liberal Netherlands. The last months of his life he spent in Stockholm where he was invited by Queen Christina. The chilly Nordic winter and the very early morning philosophical teachings ordered by the queen were too much for his health, which had always been fragile.

In fact, already during his school years the frail boy was advised to lie in bed as late as he pleased in the mornings. There is a story that thanks to this lifelong habit (which he combined with thinking) he invented analytical geometry. One morning his eye caught a fly crawling on the ceiling of his bedroom. How could one describe the path of the fly mathematically? The answer was given by x and y coordinates: in his imagination Descartes labeled each point of the ceiling by an (x, y) pair of numbers. Geometry and algebra became happily married. And the matchmaker came to be called the father of modern mathematics.

In 1619 Descartes experienced three dreams which he interpreted as an invitation to reconstruct and raise human knowledge to the level of certainty possessed heretofore only by mathematics. One result was a doctrine, the Cartesian system. It enjoyed high popularity in the seventeenth and eighteenth centuries, but also was attacked by the Church (both Catholic and Protestant) as poorly disguised Copernicanism.

A starting point of Cartesian physics was the law of inertia, previously discussed by Galileo, but which Descartes clearly formulated for a particle residing in an infinite universe. Without contact with other particles, it either would keep its initial state of rest or would move with a constant speed along a straight line, until deflected by a collision with another particle. In the light of this principle (in which one recognizes Newton's subsequent first law of motion), the various changing movements in the real world are caused by some impact. There is no vacuum and no mysterious action-at-distance. Bodies are all the time in contact with other bodies.

Descartes interpreted visible phenomena in terms of microscopic interactions. So, the attraction of a magnet for a piece of iron would be caused by invisible screw-shaped particles which are emitted from the magnet and enter screwed channels existing in the iron. The motions of the planets are forced by an ethereal vortex around the sun, somewhat as bits of cork are caught up in a whirlpool. Similar perpetual vortexes exist around other stars. The space between stars is not empty, but filled by particles of the ocean of ether.

4.4 The Swedenborg self-similar universe

Swedenborg was well versed in Newton's theory which had not yet been known in Descartes's lifetime, and which was destined to replace Cartesianism. He even brought to Sweden the mathematics of the new physics, the calculus. However, he was attracted by the basic ideas of Descartes (material interactions, vortices, no vacuum) and attempted to draw a world picture on this foundation. He was guided by a particular assumption: that everything in the world, the small as well as the great, is made according to similar principles. This idea of self-similarity is the golden thread in Swedenborg's Principia. On it he bases his attempt to build a consistent theory of particles and celestial bodies. He was fascinated by the possibility to find by a scientific method ("experience, geometry, rational reasoning") the secrets of the invisible world of elementary particles.

Swedenborg was much interested in the properties of magnetism, especially the concentrations of iron particles along, as they are now called, lines of force. The concentration is heaviest close to the axis of a magnet. Because magnetic force is much stronger than gravity, one might see a signature of its cosmic analogy also in the structure of Heavens:" ... and from a vile stone of the earth and its magnetic powers, contemplate what is similar on the largest scale".

Dante wrote in his Divina Commedia: "Pricked out with less and greater lights, between the poles of the universe, the Milky Way so gleameth white as to set very sages questioning". The Swedish sage paid attention to the fact that stars are packed most closely along the Milky Way and reasoned

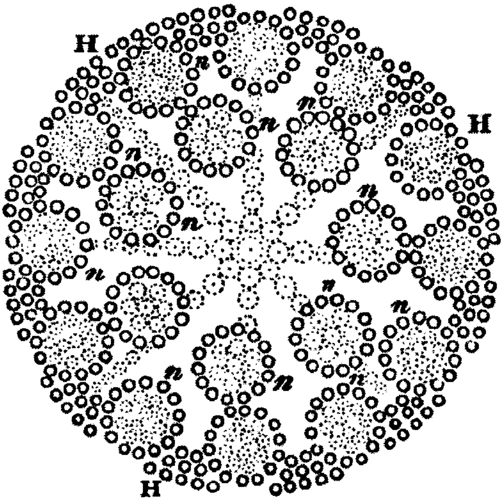


Fig. 4.2 Swedenborg's model of elementary particles. The surface of the particle is made of smaller particles which inside the particle form identical copies of the latter.

that this should correspond to an axis in the system of the stars. Swedenborg's text suggests that he pictured the Milky Way as a straight column in space. A less likely interpretation is that he described a ring (like the circular lines of force around a ring-shaped magnet). In any case, he was perhaps the first to realize that what we can see projected in the sky, may offer information on the large scale distribution of matter.

The axis of our Solar System, which also is a big magnet, deviates from the cosmic axis defined by the Milky Way. This means that we reside somewhat aside from the main concentration, and so we can see it as a column in the sky, according to Swedenborg.

Swedenborg viewed the microcosm as formed by progressively smaller and smaller particles. But the series does not extend to infinitely small sizes or arbitrarily fine substances. There is the "first finite" or the substance where "geometry begins", born from the "natural point", which is a kind of singular state between the unexplorable Infinite and our world.

There is no vacuum and no independent space (as Newton postulated), but the concept of space refers to relations between particles. Looking at the macrocosm, the world of planets and stars, Swedenborg extended to

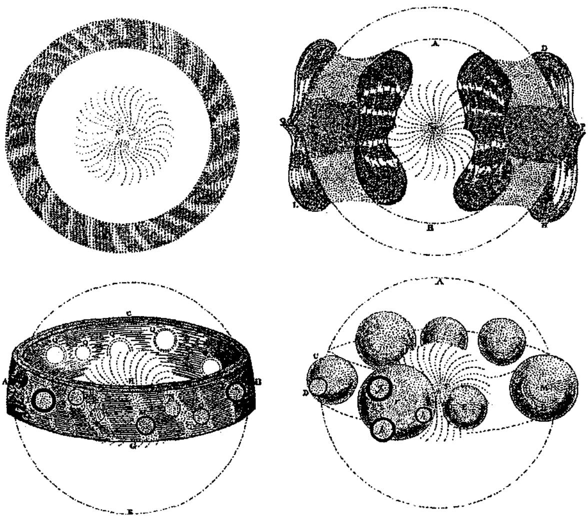


Fig. 4.3 Swedenborg's view of how the planets were born by condensation from a ring of matter which collapsed upon itself around the primordial Sun. The revolving ring explains why the planets are in a same plane and orbit in the same direction.

those large scales the multilevel structure which he imagined to exist in the microcosm. The Milky Way, formed by other suns, is just one element of a larger system which is an element of a still larger one, and so forth. This grandiose view was some decades later taken up by Kant and Lambert.

Time has almost forgotten the ambitious world architecture of Swedenborg, based on magnetism rather than gravity, and ignoring the then unknown quantum laws of the microworld. Nevertheless, his *Principia* was a result of much penetrating thinking and is still fascinating reading.

4.5 Towards the origin of the Solar System

Emanuel Swedenborg presented a sketch for the birth of planetary systems (of which ours was known, and others assumed). The origin of the flat Solar System, with the planets revolving on almost circular orbits, appeared as a great enigma in Newton's letters to Bentley. Descartes had thought that planets had previously been wandering stars in the cosmic space between other stars. After they were extinguished, their ethereal vortex had weakened and they were threatened by the attraction of the Sun's whirlpool.

With fine intuition and assisted by the idea of self-similarity, Swedenborg found it simpler to imagine that the origin of the planets and moons lies in the primordial solar mass itself. He suggested that the swift rotation of the solar mass makes, as a result of centrifugal force, its outermost parts to be cast out.

He then proposed that the planets are formed by condensation out of the detached ring of matter. Later Kant and Laplace presented a roughly similar picture for the origin of planets, already dressed in Newtonian clothing, the celebrated Kant–Laplace nebular hypothesis. These old ideas are still a part of our astronomical world view, even if the physical processes turned out to be more complicated.

Swedenborg even pointed out evidence for such an evolutionary process: the starry sky is not quite unchanging. Sometimes a new star is born and then vanishes. This happens when the broad belt of condensed matter covers the young star from sight. Now we know that such exploding and dimming new stars actually are a late phase in the life of massive stars.

Georges Buffon 1707-1788

It is often said that the French scientist Georges Buffon was the first to present a scientific theory of the formation of the Solar System. However, his theory, based on the idea of material torn out of the Sun by a comet, was published ten years after Swedenborg's Principia. †

4.6 Hierarchies of Kant and Lambert

Immanuel Kant and Johann Lambert both also pondered the appearance of the night sky above their respective native towns Königsberg and Mülshausen (the latter in Swiss Sundgen, at present French Alsace).

In a letter written by the mathematician and physicist Lambert to the philosopher Kant, one can read how a deep though simple idea may be born: "contrary to my habits then, I went into my room after the evening meal, and looked through the window at the stellar sky, and especially at the Milky Way. The insight, which I had then, to see it as an ecliptic of the fixed stars, I wrote down on a quarto page"

---

† Principia received the honor of being placed on the Index of the Catholic Church. The translator of Principia, the Rev. Augustus Clissold speculates that this happened because the proposed formation of the Solar System appeared to oppose the doctrine that God created all things out of nothing and was also difficult to reconcile with the literal interpretation of the first chapter in Genesis.

In his *Cosmological Letters*, Lambert describes the thoughts which the Milky Way inspired. He marveled at the great number of faint stars in that narrow strip across the sky. He found it hard to believe that they could lie tightly packed side by side at a similar distance from us. They should be distributed in depth, and the brightness of the Milky Way said to him that in that direction the rows of stars must be much deeper than outside it. "Briefly said, the edifice of stars is not spherical, but flat, even very flat."

Johann Lambert 1728-1777

At about the same time as Lambert inspected the sky (1749), the self-taught theologian and scientist Thomas Wright arrived at a remarkable view of the heavens. For years he had aimed at making a world model which would contain God and which would explain the appearance of the starry sky. For the latter purpose he imagined that the Sun is situated in the middle plane of a layer of stars. When one looks in the direction of the layer, many stars are seen, similarly as Lambert thought. However, Wright preferred something greater than a flat cloud. He pictured a huge spherical shell formed by stars which revolve around a distant center containing a "Primum Mobile", a large gravitating mass. The center is also the mansion of God. If the stellar shell is relatively thin, then in the vicinity of the Sun it is almost planar up to some distance, producing the great circle of the Milky Way on the sky. Wright also regarded it as probable that there are other similar Milky Ways, "many cloudy spots, just perceivable by us".

Thomas Wright 1711-1786

Lambert's and Wright's reasoning about the strip of the Milky Way in the starry sky is a nice example of how a natural phenomenon, visible to everyone from times immemorial, gains new significance, literally a new dimension, after a bright flash of idea.

Kant happened to read a newspaper report which described a book written by Wright. This gave Kant food for thought, and he also draw a cosmological picture where the Milky Way is a flattened cloud of stars and the Sun is a member of this cloud. Further, similarly as Lambert and Wright did, he presented the hypothesis that the pale elliptical nebulae, observed by astronomers, are other Milky Ways. Their great distances do not permit us to see they consist of stars.

Immanuel Kant 1724-1804

Kant and Lambert also suggested that the stellar systems form a hierarchical structure, so they are not distributed uniformly. Kant wrote: *It might further be conjectured that these higher universes are not without relation to one another, and that by this mutual relationship they constitute again a still more immense system ... which perhaps, like the former, is yet again but one member in a new combination of numbers! We see the first*

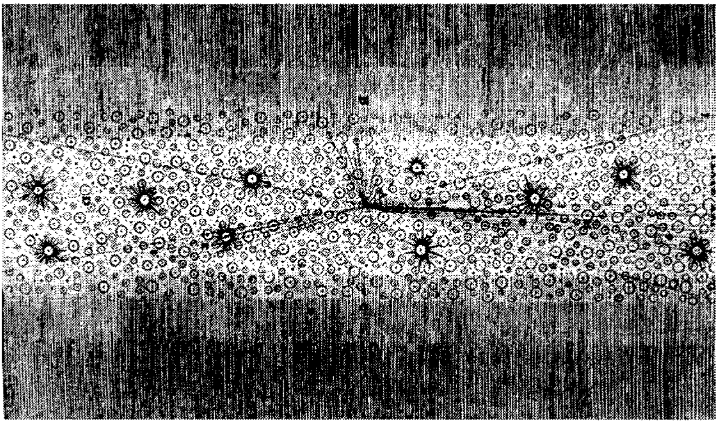


Fig. 4.4 Thomas Wright explained how a thin layer of stars gives rise to a strip of stars on the sky, the Milky Way. Imagine yourself placed in the middle of the layer.

members of a progressive relationship of worlds and systems; and the first part of this infinite progression enables us already to recognize what must be conjectured of the whole. There is no end but an abyss ... without bound.

Swedenborg wrote about his cosmic spheres: *That there may be innumerable spheres or sidereal heavens in the infinite universe, that these may be colligated one with the other, like the spheres of two magnets; and the whole visible sidereal heaven is perhaps but a point in respect to the universe.*

4.7 Finite or infinite?

There was a difference between these world models. Kant, and perhaps Swedenborg, imagined that the hierarchy continues without end towards larger and larger levels of celestial systems: it was an infinite hierarchy. Lambert thought that after a large (he cites 1000 as an example), but finite number of steps, the hierarchy ends. He thought that the stellar systems are kept together by the gravity of dark stellar masses: "... in the end you arrive at the middle point of the whole world structure and there I find my ultimate mass which governs the whole creation."

In his *Universal Natural History*, Kant explains why there should be a hierarchy of celestial bodies. Universal gravity makes smaller bodies orbit around greater masses. Planets circling the Sun is an example and similarly the stars of the Milky Way should according to Kant revolve around some mass that resides in the center of the stellar system. Analogously, the Milky Way is an element of a much larger system, again revolving around its center. The universally penetrating gravity keeps both small and large systems tied together, making possible the world structure. Hierarchic system seemed to offer a solution to the problem of stability of the universe, which was so puzzling thing in Newton's uniform stellar world. Stars and their systems do not collapse, but rotate around distant centers.

Kant speculated on how such a hierarchy could have formed. He takes the universe to be spatially infinite and initially filled with a thin substratum. He proposes that some finite time ago there was a moment of creation whereby at some point of the infinite universe there appeared a large mass. This mass, defining the "center" of the universe, put in motion the matter around it and agglomerations of other masses started to occur around the center, this process spreading as a chain reaction farther and farther from the primordial center. The process leads to systems of different order.

4.8 Emerging protofractals

Kant was very excited about this view of a developing matter distribution. However, his discussion shows that the system (similarly to the hierarchy of Lambert) has a preferred center which not everyone is ready to accept as a nice cosmological picture. Such structures may now be termed as *protofractals*, predecessors of the concept of the fractal. ‡ They exhibited self-similarity: the structure in large is repeated in the structure in small.

Already the oldest drawings, as found in caves tens of thousands of years old, testify the pleasure which people have always felt in looking at geometric patterns revealing similarity and symmetry. This affection culminated in the cosmological model based on spheres and circular motion, as studied in Plato's Academy. The world was divided into heavens, planetary spheres, and the sublunary region. When the Jewish philosopher Philo wrote the impressive words: "Heaven alone is unchangeable and self-consistent and similar to itself", he referred, not to fractals (!), but to the spherical heaven.

Philo
c.20 B.C.-
c.50 A.D.

This view was beautifully illustrated by a drawing, made widely known by Benoit Mandelbrot, from a Bible Moralisée written between 1229 and

---

‡The term "protofractal" for such early examples of structures resembling modern fractal was suggested by Benoit Mandelbrot to the authors of this book at a conference in Paris in 1999. The Greek word "protos" means "first" (c.f. elementary particle proton).

1250. There God the Geometer creates from an amorphous mass the "sky and earth, sun and moon and all elements". The whole world is a sphere, and in its interior Mandelbrot sees "circles, waves, and wiggles", the first two forms being the subject of most mathematics and science, while the "wiggles" have only in modern times become tractable as fractals.

After Copernicus, the heavenly spheres started to erode. Before Newton's explanation of planetary motion, a new view of the cosmos and its forces was drawn by Descartes. His vortices made the large scale universe closer connected with our complex immediate neighborhood. One could see nearby analogies of cosmic processes in the eddies of the restless river and in the whirlwinds shaking the trees. Descartes regarded that every star, including our Sun, is in the center of a matter vortex.

Bernard le Bovier de Fontenelle 1657-1757

Inspired by Descartes's cosmology of an infinite stellar world, Bernard le Bovier de Fontenelle, a French scholar, bravely went a step further. In his *Conversations on the Plurality of Worlds* of 1686 that passed through 31 editions during the author's long lifetime, Fontenelle supposed that stars have around them planetary systems, and each planet is also surrounded by a vortex inside the major "turbillon" around the star. These smaller vortices may drive moons around the planets. Fontenelle's charming book was an early attempt to popularize science. Its concept was simple enough: a man of science walking in the moon-lit garden with a sweet lady, explaining to her the secrets of Nature. He "chose from all Philosophy the topic which most of all arouses curiosity", i.e. the structure of the world and whether there are other worlds also inhabited by living beings. To the last question Fontenelle answered emphatically "yes".

Leonardo da Vinci 1452-1519

Another hierarchy of vortices, though unquestionably real, had been studied by Leonardo da Vinci in his superb drawings of turbulent water. His eye caught eddies inside eddies inside eddies, where someone else saw just a mess of surge and foam. Nowadays fluid "turbulenzia" – the word first used by Leonardo – is a major example of a hierarchic, fractal system.

Ivan Aivazovskij 1817-1900

The painter Ivan Aivazovskij became famous for his sea-theme, which is also interesting for glimpses of the "pre-fractal" view of Nature. The sea is not just a field of regularly spaced waves, but – as a scenery of mountains – big waves are superpositions of smaller waves. The clouds hanging over the troubled waters have edges with details in large and small.

These painters reproduced with their skilfull hand what their acute eyes saw in Nature. Jumping a little ahead in our story, we note that excursions of another type to the "wiggle" world were made by modern abstract painter

Jackson Pollock in his "action painting". He did not copy Nature, but used its own methods, which produced structures that were later called fractals. We come back to his work in Chapter 14.

4.9 Inwards and outwards

One likes to think that understanding something means an ability to imagine it as a whole. The simplest picture that Man has learnt to draw of the universe is the uniform matter distribution which continues without limits. This seems to be easy to imagine, and non-Euclidean geometry has even waived the need to extend the distribution into infinity.

But why do hierarchies also please the mind as structures of the physical reality? Why were some thinkers attracted by such a non-uniform picture even when observations did not offer evidence? Perhaps because one does not have to travel far away in order to see whole new worlds, in particular if hierarchical levels exist below you. Just change your size, or in practice, the scale of the inspected part of the World. In the words of William Blake:

William Blake 1757-1827

To see a World in a Grain of Sand
And a Heaven in a Wild Flower,
Hold Infinity in the palm of your hand
And Eternity in an hour

Or as mathematician Hermann Weyl put it, the space is "inwardly infinite", thus a potential mansion of rich structures.

The notion that there is a correspondence between the macro- and microcosm, has of course appeared before Swedenborg. So Gottfried Leibnitz entertained this idea in his *Monadology* from 1716: "Every portion of matter may be conceived as like a garden full of plants and like a pond full of fish. But every branch of a plant, every member of an animal, and every drop of the fluids within it, is also such a garden..." A less idyllic scene was described by Bruno (with his tongue in his cheek, certainly). When one of his heros tells that "...from the inspection of the macrocosm it is easy, by making necessary conclusions from the similarity, to learn about the microcosm, the particles of which correspond to the parts of the former.", the other one replies: "So that we can discover inside You the Moon, Mercurius, and other stars; France, Spain, Italy, England, Calcutta, ...?"

Gottfried Wilhelm Leibnitz 1646-1716

Swedenborg was guided by the beneficial aspect of self-similarity. Referring to the different levels of systems, he was happy to say that "he who has learnt the nature of one will have learnt the nature of all". He took self-similarity as a cosmological principle which permeates the cosmos on its all scales. Only later, when one could observe both the micro- and the macroworld, did it become possible to check how widely this idea applies.

Kant made an interesting prophecy: "...the formation of all celestial bodies, the cause of their motions, in brief, the origin of the whole present arrangement of the world edifice, will sooner be understood than the production of a single herb or of a caterpillar will come evidently and completely clarified from mechanical reasons". Kant discusses the question how a man, who easily errs in the small things around, could be able to study the large scale things. He asserts that among all tasks of the study of nature none can be solved more correctly and with more certainty than the structure of the world edifice at large. Why is he so confident? He points out that celestial bodies populate an empty space and are separated by great distances. There is only the attractive gravity force influencing their motions. All this is much simpler than what makes the things work here on the earth. Thus Kant foresaw that on widely different scales different physical laws dominate – "he who has learnt the nature of one" does not necessarily know the nature of all.

* * *

Isaac Newton, Kant's master in physics, had speculated that in order to understand small things, one has to assume that down in the microworld there are both attractive and repulsive forces which arrange tiny atoms into molecules and these into larger hierarchic systems up to visible things. Solid crystals and beautifully regular, but shortlived snowflakes indicated for him the presence of such forces. Modern physics has proven true Newton's idea that the "Hate" of Empedocles works together with his Love in molding a rich physical reality.

The ideas of self-similarity, fractality and chaos became an important part of scientific culture towards the end of the 20th century. But before that happened science made illustrious other advances.

PART II

COSMOLOGICAL PHYSICS FOR THE REALM OF GALAXIES

This page is intentionally left blank

Chapter 5

The new world of relativity and quantum forces

Cosmological physics of the 20th century radically changed our view of the cosmos. Surprising experiments in the laboratory and discoveries at observatories drove classical physics into a blind alley. The puzzling velocity of light, which is always the same, the discrete spectral lines from glowing gas, and the embarrassing extra forty arc seconds in the motion of Mercury eventually transformed physics. New theories appeared, strange concepts surfaced, such as relativity, quanta, boiling vacuum, curved space...

The road from Laplace's apparently clear, but too simple clockwork universe to the strange, but closer to true quantum-relativistic world has been littered with paradoxes and new concepts which are hard for human imagination. The deepest upheavals have gripped the microcosm and the classical notion of particle. Fundamental physical forces are now understood as processes of the exchange of "force carriers" – the quanta of fields. The interplay between such forces is the reason for the rich material structures we see in the diverse conditions of plasma, gas, fluids, and solid matter.

* * *

5.1 The Principle of Relativity

At the dawn of the 20th century a few bright minds realized that there is a difference between the familiar physics of slowly moving bodies and physics at very high speeds. In the works of Poincaré and Einstein, a new theory was developed, which united space and time into space-time.

Henri Poincaré 1854-1912

Henri Poincaré introduced the *Principle of Relativity* into physics. In a lecture in 1904 at St.Louis, titled *The principles of mathematical physics*, he outlined the necessity to base physics on

the principle of relativity according to which the laws of physical phenomena should be the same, whether for an observer fixed, or for an observer carried along in a uniform movement of translation; so that we have not and could not have any means of discerning whether or not we are carried along with such a motion, and to develop

an entirely new mechanics which would be, above all, characterized by the fact that no velocity could surpass that of light ... this observer would not use the same clocks as a fixed observer, but indeed, clocks marking "local time".

Poincaré said that the crisis of the old physics arose when physicists started to make accurate measurements of phenomena in new conditions. He was inspired by the attempt by Albert Michelson and Edward Morley, in 1887, to measure the velocity of the Earth relative to the aether (we are orbiting the Sun with a speed of 30 km/sec). Physicists had assumed that there is a unique frame of reference, termed the aether, relative to which the speed of light has its measured value. If so, classical physics says that it should be possible to measure our velocity relative to the aether. But Michelson and Morley showed that this is not so.

Albert Michelson 1852-1931

The principle of relativity means that making experiments in a closed room we cannot determine whether the room is stationary or moving with uniform velocity in some direction. This was already clear to Galileo, but Poincaré had in mind not only mechanical tests, like dropping a ball, but also electromagnetic phenomena. This profound idea, which also inspired Einstein, liberated physics, so to speak, from the fetters of absolute space and time, and opened the way for very fruitful new theories of physical reality. One may see an analogy with the principle of no center, which made all places of the universe equivalent and led to modern cosmology.

5.2 The relativistic physics of Poincaré and Einstein

In 1905 Henri Poincaré submitted two articles, one dated June 5^{th} to *Comptes Rendus de l'Académie des Sciences* and the second dated July 23^{rd} to *Rendiconti del Circolo Matematica di Palermo* under the title "On the dynamics of the electron". In these studies he gave the complete mathematical formulation of the new relativistic mechanics and utilized the four-

Fig. 5.1 Founders of new physics at the 1911 Solvay Congress. Sitting by the table Mme Marie Curie with Henri Poincaré. Behind them standing Albert Einstein, and on the left James Jeans and Ernest Rutherford (a detail from the conference photo).

dimensional space-time, nowadays called Minkowski space. He coined the name *Lorentz invariance* to the novel property of physics, which binds space and time together into space–time. Even earlier he had emphatically stressed the non-existence of absolute space and time.

Albert Einstein 1879-1955

On June 30th of the same summer, Albert Einstein, at that time a young unknown physicist, submitted his article "On the electrodynamics of moving bodies" to *Annalen der Physik.* * In this famous study he also arrived, though in another way, at the physical foundations of the new mechanics, and stated in a clear manner that the aether was no longer needed. Einstein's presentation received more attention among physicists, and usually one speaks about Einstein's *special relativity.* It is startling how these two great physicists simultaneously arrived at the novel ideas on space and time, the mature Poincaré who worked in all fields of mathematics and physics and the novice Einstein whose name was to become the symbol of 20th century science.

---

*The year 1905 was important for Einstein, a Technical Expert at the Patent Office in Bern, Switzerland: Annalen der Physik published three epoch-making articles by him.

5.3 Velocity of light

In relativity theory, a pivotal role is played by a fundamental constant of Nature, the velocity of light in a vacuum. It is the maximal speed of propagation of any influence or information. [†] And it has the same value for any observer, independently of motion:

$$c = 299\,792.5 \text{ km/sec}$$

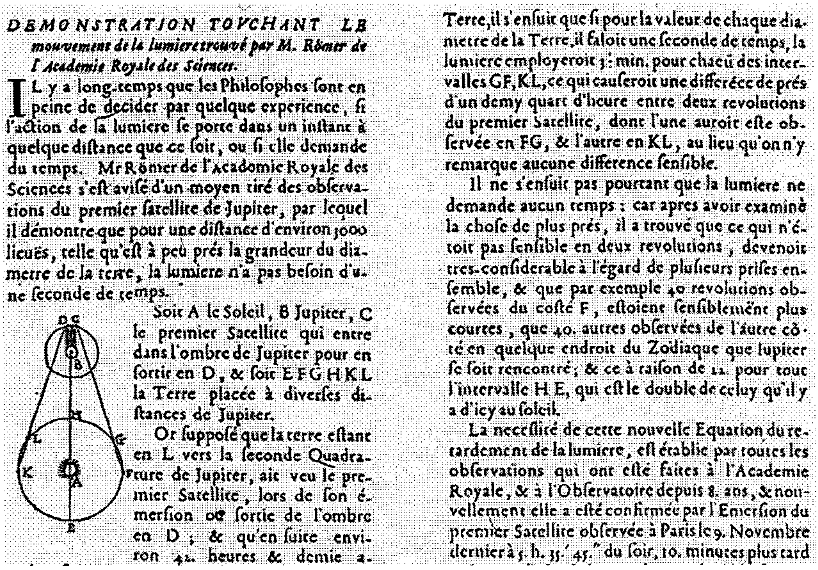
The speed of light is huge compared with familiar motions on the Earth. It was first measured in cosmic conditions where, even for light, it takes a noticeable time to cover large distances. Working in the Paris Observatory, the Danish astronomer Ole Rømer studied Jupiter's innermost moon Io in the role of a clock that could be used at sea for determination of geographic longitude, as Galileo had suggested. But this time keeper was not as accurate as had been expected. Sometimes it was fast, sometimes slow.

Ole Rømer 1644-1710

Rømer was sure that this variation of about 22 minutes, was caused by the finite velocity of light. In 1676 he published a report that, in effect, was the first determination of the speed of light. In fact, his report did not contain an explicit calculation of the speed, but rather a discussion of how he detected the apparent variation in Io's motion and it conveyed his conviction that the reason is the finite velocity of light. The variation of 22 minutes is the time which the light spends traversing the Earth's orbit. In modern units this would correspond to a speed of about 227 000 km/sec (the error was due to problems with timing Io's motion).

The conclusion aroused little enthusiasm, as it was generally thought that light rays travel instantaneously – an exception was Newton who in his Principia, a decade later, announced that the speed of light is finite, as measured by astronomers. In Paris, the things were complicated by the fact that Rømer's boss, Giovanni Cassini, had earlier proposed a similar interpretation for Io's peculiarities, but had withdrawn it, probably as too speculative. The younger colleague gave wings to that explanation and raised it into a testable scientific hypothesis. Rømer's report came after his succesful prediction that an eclipse of Io would occur 10 minutes late. This great observation was also cosmological – the effect arises, because the Earth moves along its orbit around the Sun.

[†] For all kinds of interactions $v \le c$. In his 1905 paper Poincaré emphasized that gravitational influence also must travel with the speed of light. For light and gravity $v = c$.



DEMONSTRATION TOVCHANT LE mouvement de la lumiere trouvé par M. Römer de l'Academie Royale des Sciences.

IL y a long-temps que les Philoſophes ſont en peine de decider par quelque experience, ſi l'action de la lumiere ſe porte dans un inſtant à quelque diſtance que ce ſoit, ou ſi elle demande du temps. Mr Römer de l'Academie Royale des Sciences s'eſt aviſé d'un moyen tiré des obſervations du premier ſatellite de Jupiter, par lequel il démontre que pour une diſtance d'environ 3000 lieuës, telle qu'eſt à peu prés la grandeur du diametre de la terre, la lumiere n'a pas beſoin d'une ſeconde de temps.

Soit A le Soleil, B Jupiter, C le premier Satellite qui entre dans l'ombre de Jupiter pour en ſortir en D, & ſoit E F G H K L la Terre placée à diverſes diſtances de Jupiter.

Or ſuppoſé que la terre eſtant en L vers la ſeconde Quadrature de Jupiter, ait veu le premier Satellite, lors de ſon émerſion ou ſortie de l'ombre en D; & qu'en ſuite environ 42. heures & demie a-

Terre, il s'enſuit que ſi pour la valeur de chaque diametre de la Terre, il faloit une ſeconde de temps, la lumiere employeroit 3½ min. pour chacũ des intervalles GF, KL, ce qui cauſeroit une differẽce de prés d'un demy quart d'heure entre deux revolutions du premier Satellite, dont l'une auroit eſté obſervée en FG, & l'autre en KL, au lieu qu'on n'y remarque aucune difference ſenſible.

Il ne s'enſuit pas pourtant que la lumiere ne demande aucun temps: car apres avoir examiné la choſe de plus prés, il a trouvé que ce qui n'étoit pas ſenſible en deux revolutions, devenoit tres-conſiderable à l'égard de pluſieurs priſes enſemble, & que par exemple 40 revolutions obſervées du coſté F, eſtoient ſenſiblemẽt plus courtes, que 40. autres obſervées de l'autre côté en quelque endroit du Zodiaque que Jupiter ſe ſoit rencontré; & ce à raiſon de 22. pour tout l'intervalle H E, qui eſt le double de celuy qu'il y a d'icy au ſoleil.

La neceſſité de cette nouvelle Equation du retardement de la lumiere, eſt établie par toutes les obſervations qui ont eſté faites à l'Academie Royale, & à l'Obſervatoire depuis 8. ans, & nouvellement elle a eſté confirmée par l'Emerſion du premier Satellite obſervée à Paris le 9. Novembre dernier à 5. h. 35.′ 45.″ du ſoir, 10. minutes plus tard

Fig. 5.2 A part of the communication to the French Academy by Ole Rømer where he reported the detection of the large, but finite speed of light.

5.4 From classical space and time...

René Descartes in the 17th century realized that physics should be built on inertial motion. The concept of the inertial frame derives from the remarkable property of a free particle: it preserves its uniform motion along a straight line. "Free" means the absence of force or the exact compensation of all forces. "Perfect" circular motion is not free, and old attempts to base the science of mechanics on spherical motion were doomed to fail.

Dynamics is a science of forces, and requires a uniformly moving stage which displays what happens without and with forces. This is the inertial frame. You can play billiards as well on the ground or in an airplane if the airplane flies straight and steady. However, the game is spoiled if the pilot makes a sudden turn and the frame of the game is no longer inertial. The Principle of Relativity means that in all inertial frames physical laws are the same, and one cannot pick up any single frame as truly special.

In the physics of Newton, space, time and particles defined a simple model for understanding physical reality. Space was Euclidean. Time flowed uniformly. Distance and time are the same as measured by solid rulers and firm time-keepers riding on every inertial frame. Such a model of space and time, in which both elements have a kind of absolute reality, is a good stage for phenomena occurring in many physical experiments and in our everyday life. This view prevailed in physics before the theory of relativity.

5.5 ... to relativistic space-time

Absolute space and time are gone, but we are not left free and wild in mere nothingness. Relativity theory defines a single concept of space-time, instead of Newton's separate space and time. In everyday life, where velocities are much smaller than that of light, we like to share Newton's view. The velocity of an aircraft, 300 m/sec, is very small indeed, just one millionth part of the speed of light!

Newton's world may also be thought as a space-time. But relativistic space-time is different: for instance, a particle cannot freely wander everywhere in space and time, because it does not reach speeds larger than c. One cannot at will change the distance (in space) of a particle in an arbitrary interval (in time). The velocity barrier reflects a subtle connection between space and time, the above mentioned Lorentz invariance.

For further insight on space–time, recall that the velocity of light is the same as measured by all observers: one who is at rest relative to a shining lamp measures the same speed for the light emanating from the lamp as another one who moves at a high speed away from the lamp. The light ray passes both of them at the same speed. If space and time are independent this is incomprehensible. The conclusion of relativity is that space (distance) and time measurements are intertwined.

Hermann Minkowski 1864-1909

One of the founders of the space-time was Hermann Minkowski, who was born in Russia and later taught mathematics in Zürich (Einstein was one of his pupils). In 1902 he accepted a chair in the University of Göttingen. In his famous lecture of 1908, Minkowski announced that "henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality". Only such a united space and time allows one to understand experiments that physicists do every day. Even your home TV is made according to the requirements of relativity: electrons of the beam drawing the images travel nearly at the speed of light.

5.6 Time travel into the future with a one-way ticket

One famous property of the new space-time is that differently moving observers get their own measures of space and time. The clock carried by a moving person measures a shorter time between two events than a similar clock held by a stationary observer. This slowing down of time, an experimental fact, allows one to dream about a kind of time-travel.

Make a high speed trip to a nearby star. When you come back, your clock, calendar, and grey hairs tell that the trip took, say, 20 years. However, on Earth 25 years has lapsed since your departure, and your twin brother has celebrated five more birthdays than you did on board the rocket. In this sense relativity theory makes possible a travel into future, or more exactly, to slow down your time relative to the time ticking on earth. [‡]

5.7 Rest mass energy: $E = mc^2$

Einstein's formula connecting the energy and mass of a material body has become a familiar, impressive, even enigmatic symbol of science. It gives the maximum energy that can be extracted from any body which is at rest. Because the square of the speed of light is very large the energy is also huge. The energy in one gram of any stuff would power a family house in the cold northern weather for one thousand years. In the everyday life of unhurried motion and weak forces, this energy is quietly conserved in the rest mass and does not manifest itself, but lies dormant.

The mass–energy can be awakened in conditions where interactions between particles are very strong. In nuclear weapons and, more happily, in the Sun, a part of Einstein's energy is released, in fact only 1 percent of it.

Relativity divides all particles in Nature into two classes depending on whether their rest mass is zero or non-zero. Massless particles always move at the speed of light and cannot be stopped. Light itself provides an example – the photon. The other class of particles, those with non-zero rest mass, may only dream of the speed of light. When the velocity of such a bit of

---

[‡] The Twins Paradox comes to light when one notes that the space traveler may regard himself as at rest while the Earth is departing at a high speed (or approaching on homebound journey). Hence, the twin on Earth is moving and should age slower than the astronaut, totally contrary to what was first concluded! The solutions of the puzzle are usually based on the observation that after all, the men are not doing the same things. The astronaut made an about-turn, while his brother preferred to stay at rest.

matter is forced to approach that of light, its total energy grows beyond limit, and one can never close the gap to the speed barrier.

5.8 Light, electricity, and magnetism

Light is all around, it pours out of the Sun, from camp-fires in the forest and from fire-flies in the summer night. As the name says, it is "light", in fact without rest mass. Its constant velocity the ruler of space-time, light is a relativistic phenomenon which we can see with our own eyes. From the deep soil of relativity arise other things, which one at first sight does not suspect as close relatives of light. These are electricity, the old "amber effect", and magnetism, also known from ancient times. [§]

Charles Coulomb 1736-1806

Charles Coulomb found in 1785 that the electric force obeys, similarly as gravitation, the "inverse square law". The force between two charged particles is inversely proportional to the square of the distance between them. After it had been for long suspected that there is a link between electricity and magnetism, Hans Ørsted discovered in 1820 that an electric current (moving charges) gives rise to a magnetic field.

Hans Ørsted 1770-1851

Michael Faraday 1791-1867

The concept of the field for the description of forces arose in physics, when Michael Faraday introduced the notion of force lines and in his experiments visualized the magnetic field. The field has an important function: it carries force through space, instead of a mysterious action-at-distance.

James Maxwell 1831-1879

And in 1865 all the empirical knowledge on electricity and magnetism was compressed into short mathematical laws by James Maxwell. He united the electric and magnetic phenomena in one physical entity, the electromagnetic field. Maxwell's theory encompassed also light: light is an electromagnetic wave propagating with a high velocity, as if very closely spaced ripples of the electromagnetic field. This field is a material agent which transports energy and momentum by "action-through-distance".

Woldemar Voigt 1850-1919

It is curious that the first theory which is relativistic was Maxwell's electromagnetism – it was constructed well before relativity theory itself! Indeed, when he invented his famous equations, Maxwell did not know

---

[§] "Electricity" and "electron" come from the Greek word for amber, which attracts bits of paper when rubbed with fur. The formula for the Coulomb force between two charges Q_1 and Q_2 is $F_{el} = Q_1 Q_2 / r^2$, remarkably similar to Newton's gravity law $F_N = GmM/r^2$. Also, both influences propagate with the speed of light. Modern physics explains these facts as due to the zero rest masses of the carriers of both forces.

that they were hiding a treasure: the theory of relativity. Now we understand that this had to be so, because electromagnetism and light are relativistic phenomena. Only in 1887 did Woldemar Voigt and later Hendrik Lorentz recognize that Maxwell's equations have other space and time properties than Newton's equations, thus laying out the groundwork for relativity. Nowadays every physics student is fluent with what are called Lorentz transformations in relativity theory.

Woldemar Voigt 1850-1919

Hendrik Lorentz 1853-1928

5.9 Least action, symmetry, conservation laws

The classical equations of motion may be derived from one law, the *Principle of Least Action.* Action is a quantity which characterizes how much work a physical system can perform during a time interval. It is a remarkable feature of this theory which so well describes the world, that a particle always "chooses" such a path for which the action has a minimal value. A system develops in such a way that the energy and time consumption is the most economic one. "Nature does nothing in vain" was already said in Antiquity. The Principle of Least Action is so deep in Nature that it has remained as a cornerstone of relativistic and quantum physics, too.

The symmetry of space and time is simple, but profound. Emmy Noether proved in 1918 that conservation laws are direct consequences of symmetries of space and time. In her proof of this theorem, she inspected the quantity Action. The invariance of Action, when the physical system is displaced in space or in time, is the very heart of the conservation laws. This was one of the greatest achievements of 20th century physics.

Emmy Noether 1882-1935

Energy conservation corresponds to uniformity of time (any moment is as good as any other) and conservation of *momentum and rotational momentum* is connected with uniformity and isotropy of space (any place and direction is as good as any other). Isotropy and uniformity are the basic symmetry properties of Euclidean space. [¶]

[¶] Isotropy means the symmetry of all directions: a rotation around any fixed point does not change anything. Uniformity is the symmetry of all sites: a shift from one point to another does not change anything. Space and time, if considered separately, have in special relativity the same symmetry properties as they have in classical dynamics. Isotropy and uniformity guarantee the conservation of energy E and momentum $\vec{p}$, and the latter are united into one four-dimensional entity, the energy-momentum vector, $\vec{P} = (E, \vec{p})$. Noether's Theorem contains even deeper things: the Minkowski space is the cause for the conservation of the energy-momentum tensor T^{ik}.

5.10 Quantum physics of the microworld

Einstein liked to think of the 17th century as the happy childhood of science and others have said that in the 19th century science finally came of age. At that optimistic epoch of steam engines and electric light it seemed that all the phenomena of the material world might be understood on the basis of classical particles moving in absolute space and interacting by gravity and electromagnetism. The world appeared quite "transparent" to the viewer.

However, two dark clouds were looming on the horizon. One was that experiments to detect the aether had failed. This led to relativity theory. The second was the lack of explanation for the thermal radiation from hot bodies, the so-called black-body radiation. Theoretically the radiation should be increasingly strong at short wavelengths ("the ultraviolet catastrophe"), while actually the intensity turns down after a maximum.

Max Planck 1858-1947

In 1900, Max Planck announced that the concept of discrete energy states or *quanta* explains the radiation of hot bodies. Energy does not appear as a continuum which may be divided into arbitrarily fine slices. Rather, it possesses grainy structure, consisting of little packets of energy ("quantum" means "amount"). This was the beginning of quantum physics.

Planck linked the energy of the quanta to a property of the emitted electromagnetic radiation: the energy quantum is proportional to the *frequency* of the radiation. The constant of proportionality is a very small number in usual units, it has 27 zeros after the decimal point:

$$\textit{Planck's constant} = h = 6.626 \cdot 10^{-27} \text{ erg sec}$$

The value of this fundamental constant of Nature was first determined by Planck himself in his epochal article, just after accurate new measurements of the spectrum of thermal radiation by two teams of German physicists.

The next step to the quantum Nature of the microcosm was made by Einstein in 1905. His article on "the creation and transformation of light", which was to earn him the Nobel Prize in 1921, explained all the main properties of the photoelectric effect where light shining on certain metal surfaces kicks off electrons. Einstein concluded that light behaves as if it consisted of separate radiation quanta, photons, also when traveling far from its source.

Spectral lines bring important messages from the microcosm. The spectrum emitted by hydrogen atoms consists of narrow lines corresponding to different frequencies of light. In 1885 Johann Balmer, a Swiss mathematics teacher, had found that the frequencies of the hydrogen spectral lines have not haphazard values, but follow a mathematical law based on integer numbers. Balmer's law remained a mystery until 1913 when Niels Bohr from Denmark proposed his "planetary model" for the atom: electrons move around the nucleus like planets around the Sun. The spectral lines are formed when electrons jump between the orbits (energy levels), absorbing or emitting photons. This model looks nice and easy, but it is flagrantly non-classical! According to classical physics, the charged, radiating electrons would spiral into the center. Something must keep them in their orbits. And the orbits can have only certain sizes in order to produce the observed spectral lines. These sizes – that is, the structure of the atom – are determined by quantum laws. The atom is a genuine quantum thing.

Johann Balmer 1825-1898

Niels Bohr 1885-1962

For Newton's mechanics and relativistic electrodynamics such phenomena were incomprehensible. Only the quantum theory which unified particle and wave properties of matter and light, managed to explain these. Below we make a brief plunge into these strange seas of thought and Nature.

5.11 Heisenberg's Uncertainty Principle: nebulous particle

In 1923 Louis de Broglie, a French prince and physicist, generalized Einstein's idea on the dual nature of light to any material particle. A wave is not a property of light only! Even particles have their wave-like side. [‖]

Louis de Broglie 1892-1987

What do we normally have in mind, when we want to describe a particle? Its position in space, its velocity, and its mass. Quantum phenomena have forced us to recognize that such a classical particle is a too idealized picture of reality, a too tight "mathematical mask" for the real entity.

In classical physics, and also in relativity theory, the motion of particles is "ordinary", i.e. they move along sharp trajectories and measurements can in principle be as accurate as you like. These assumptions were for long regarded as self-evident. But not in quantum physics!

Heisenberg's Uncertainty Principle says that *it is impossible to measure*

Werner Heisenberg 1901-1976

[‖] Light is simultaneously a wave with wavelength λ and frequency f and a particle with energy $E = hf$ or momentum $p = E/c$. De Broglie introduced an intimate relation between wavelength λ and momentum p for any material particle: $\lambda = c/f = h/p$. Such matter waves explain why electrons can move only on certain orbits: an orbit can exist only if it accommodates an integer number of the matter waves of the electron.

simultaneously the exact position and velocity of a particle. The product of the measurement uncertainties are never less than Planck's constant. The better you measure the particle's position, the less accurately you know its velocity, and vice versa. This is not due to some defect in the measuring apparatus. Simply, the intuitively appealing "classical" particle is no longer an adequate picture for the entities of the microworld. ** On the small scales of space, new laws appear.

The old physics had no objection against a free particle which moves with a constant, exactly known velocity. But then the Uncertainty Principle tells that we do not know *anything* about its position – it is anywhere and nowhere in the infinite universe! A classical particle simply cannot exist in the quantum realm. This is a dramatic change in our world view.

The need for quanta was not felt in classical times. The new properties appear only when one looks at very small sizes, much less than microbes and cells visible from the 17th century through microscopes. Because Planck's constant is so small, quantum haziness does not meddle in everyday affairs. Though we may live happily without being aware of the quanta, they are necessary for understanding many things around us. We may sit and look at our blazing star and wonder without end how it shines. Like it or not, quantum physics is needed to grasp what is going on in the Sun.

5.12 The search for genuine atoms

Physics is based on experiments, which are a way of examining the world by active influence on it, by organizing special conditions in order to see in pure form phenomena predicted by physical theories. A child learns about things around him by touching and playing. The ever curious physicist learns about the world by playing with the concepts in his theory and watching the phenomena which appear in his experiments.

True, in this way one can probe only a tiny part of the universe. One cannot move stars or make galaxies collide. But astronomers can plan their observations to test theories about stars and galaxies. Somewhat similarly,

---

**More exactly, there is the momentum p of a particle, which in slow motion is velocity multiplied by mass $p = mv$. The uncertainty principle has two forms, for position and momentum $\Delta x \times \Delta p > h$, and for energy and time, $\Delta E \times \Delta t > h$. Δx and Δp are the uncertainties in the position and momentum of a particle. ΔE is the accuracy with which one can measure the energy of a system during time interval Δt.

physics explores the world in the small which is also impossible to reach directly. Big accelerators create high energies in very small volumes, and physicists observe what happens in that outlandish microworld. Thus they have penetrated into previously closed realms and at every step found new particles which at the time seemed indivisible ultimate portions of matter.

Henri Becquerel 1852-1908

In 1896 Henri Becquerel found that uranium salts emit penetrating radiations. Then, patiently handling a ton of a uranium ore called pitchblende in their very modest laboratory in Paris, Marie and Pierre Curie discovered a new strongly radiating element, Radium. This radioactivity (the word invented by Marie Curie) offered a key to the atomic nucleus, as something was ejected from the nucleus, in fact three types of rays termed *alpha, beta*, and *gamma*. It turned out later that alpha-rays are nuclei of Helium, beta-rays are electron streams, and gamma-rays are very energetic photons.

Pierre Curie 1859-1906

Marie Curie 1867-1934

In 1910 Ernest Rutherford's famous alpha-ray scattering experiment showed that atoms have a very small core surrounded by swarms of electrons. [††] In the 1920's he and others found that sometimes the alpha-ray was absorbed by the nucleus, and caused the atom to spit out hydrogen nuclei. They realized that the hydrogen nucleus is one of the fundamental building blocks of matter. It was given the name "proton".

In the 1930's a fourth type of radioactivity was found. It was an electrically neutral particle, different from the negatively charged electron and positive proton. In 1932 James Chadwick determined its mass, which turned out to be very close to the mass of proton. And because of its lack of charge, it was named the "neutron".

James Chadwick 1891-1974

Nuclei of different elements are made of different numbers of protons and neutrons. The Helium nucleus consists of two protons and two neutrons. Each element is characterized by the total number of protons, the *charge number*, and the sum of protons and neutrons, the *mass number*. The charge number is the main identifier of an element, giving also the number of electrons around the nucleus. So one uses for the Helium nucleus the designation $_2He^4$. It is remarkable that even though we know that the atoms on the quantum level are not sharp-lined miniature "solar systems", we can characterize them by the integer numbers of their "planets". Pythagoras was here!

[††] Ernest Rutherford (1871-1937), professor at Cambridge University, worked in the same Cavendish laboratory where Joseph Thomson (1856-1940) had in 1897 discovered electron as the constituent of cathode rays.

Table 5.1 The elementary particles at the end of the 20th century. These are the main constituents of Nature as seen by the contemporary quantum field theory. Protons, neutrons, and other ***baryons*** are no longer elementary, but they are made of quarks and gluons. Note that rest mass is expressed in energy units, 1 eV = 1.6×10^{-12} erg.

Particle				rest mass	charge	spin
fermions	*leptons*	electron	e	0.511 MeV	−1	1/2
		muon	μ	0.106 GeV	−1	1/2
		tau	τ	1.78 GeV	−1	1/2
		e–neutrino	ν_e	< 2.5 eV	0	1/2
		μ–neutrino	ν_μ	< 2.5 eV	0	1/2
		τ–neutrino	ν_τ	< 2.5 eV	0	1/2
	quarks	up	u	6 MeV	2/3	1/2
		down	d	10 MeV	−1/3	1/2
		strange	s	0.25 GeV	−1/3	1/2
		charm	c	1.2 GeV	2/3	1/2
		bottom	b	4.3 GeV	−1/3	1/2
		top	t	174 GeV	2/3	1/2
bosons	scalar	Higgs	H	< 220 GeV	0	0
	vector	photon	γ	0	0	1
		W boson	$W^\pm$	80 GeV	±1	1
		Z boson	Z^0	91 GeV	0	1
		gluon	gl	0	0	1
	tensor	graviton	g	0	0	2

Enrico Fermi 1901-1954

A good example of the special way in which particle physics develops is the discovery of ***neutrino***. In 1934, Enrico Fermi, an Italian physicist, explained the beta-decay of nuclei as a result of a ***weak interaction***. He realized that a new particle was needed in order to explain the puzzling violation of the conservation of energy and momentum which was observed in the reaction where a neutron decays to a proton plus an electron plus "something". The ***neutrino*** has a very small rest mass and no charge.

The microworld is rich in different types of particles which may be described by a number of properties: mass, electric charge, spin, and lifetime. For each particle, there is a corresponding ***antiparticle*** class. A particle and its antiparticle have opposite charges, but the same mass. Some of these entities make up the more or less permanent matter around us. Others are just brief visitors to our reality and swiftly decay into more stable forms.

5.13 Quarks hide inside protons

For some time protons and electrons were regarded as indivisible true "atoms". But Nature again turned out to be not that simple. The proton lost its status of the "primordial particle". It is made of *quarks.* Quarks were not discovered in the way you may encounter some new insect species in the jungle – the microcosm is much less hospitable for such expeditions! In fact, nobody has ever seen a free quark. The quarks were first figured out in the 1960's by Murray Gell-Mann and, independently, George Zweig, as convenient mathematical tools to make calculations in the complexities of elementary particle physics. Perhaps they were also real physical particles? They are, but this only became accepted later, when the existence of small charged grains inside protons was observed in high-energy experiments in which a proton was hit by a beam of electrons. (This reminds us of how Copernicus's new planetary model was at first viewed as just another way of calculating the positions of the planets. In fact, also the quarks first encountered much disbelief – for example, Zweig's article on the new theory was rejected for publication.)

Remarkably, the quarks have fractional electric charges ($+\frac{2}{3}, -\frac{1}{3}$ of the proton's charge). Six types of quarks have now been discovered (see Table 5.1). Another novel property of quarks is that they cannot exist in isolation, but are as it were bound together by a rubber band – the attractive force between them *grows*, when the distance increases! They are not only nebulous, but also no longer true individuals.

While an electron has electric charge, each quark has besides the electric charge also the so-called color charge. The carriers of color forces between quarks are called gluons. Now the proton and neutron are no longer elementary particles – they consist of quarks strongly bound by gluons. A proton is made up of three quarks (two up-quarks, one down-quark). A neutron contains one up-quark and two down-quarks.

Those particles called leptons (electron etc. in Table 5.1) are not made of quarks. Quarks and the not-so-shy leptons are in a sense cousins in the world of elementary particles. As far as we know, they make up all matter and do not have substructure. The third category of particles, bosons, mediate interactions between other particles.

5.14 The quantum nature of fundamental forces

Modern physics says that all physical phenomena can be understood as effects of four fundamental forces between elementary particles: the *strong force*, the *weak force*, the *electromagnetic force*, and the *gravitational force*

This order corresponds to their decreasing strength and, hence, importance in the microworld. The strong force appears only at very small distances inside atomic nuclei, and disappears outside. It binds together the protons and neutrons forming the nucleus, and is much stronger than the electric repulsion between protons. Gravity is the weakest force, usually of no importance for elementary particles. But unlike the other interactions, the gravity force grows with the number of particles, and finally becomes dominant in the astronomical world.

Table 5.2 The fundamental forces of Nature

Force	Relative strength	Range	Phenomena
Strong	2000	10^{-13} cm	atomic nuclei
Weak	10^{-8}	10^{-15} cm	neutron decay
Electromagnetic	1	infinite	atoms, molecules, solids
Gravitational	10^{-43}	infinite	stars, galaxies, universe

The concept of the field is a key player in the theory of fundamental forces, the carrier of force through space. The new theory of *quantum relativistic fields* unites quantum and relativistic properties of matter, including radiation, in order to describe physical interactions in all their richness. The theory explains in terms of quanta the processes that give rise to the observed physical forces.

The Standard Model of elementary particles includes the theories of electromagnetic, weak, and strong forces. These are called, respectively, quantum electrodynamics (QED), Weinberg–Salam–Glashow electroweak model, and quantum chromodynamics (QCD). The last term reminds us that color (chroma in Greek) is a new type of charge in strong interactions.

Elementary particles are quanta of fields, either fermion (half–integer spin) or boson (integer spin) fields. In the microcosm fermions represent "matter", while bosons take care of "forces" between the bits of matter. So physical interactions can be imagined as the exchange of force carriers.

In the Standard Model the strong, weak and electromagnetic forces are described as independent entities. However, in the last few decades new theories have appeared which unify these interactions. E.g. in the Grand Unified Theories (GUT) there is one quantum field which at low energies appears like a three-headed eagle, giving rise to three fundamental forces.

The next more radical step attempts to see matter and forces as two sides of one still more fundamental field. The Supersymmetric Theories unite fermions and bosons into one "superentity".

5.15 "Fur coat" of virtual particles and the boiling vacuum

The Standard Model of elementary particles is complicated, but we offer a glimpse of how the force between particles may be understood. Take two electric charges (say, electrons) that are at rest. They feel an electrostatic force between them. What transfers this force from one body to the other? How does one body know about the second body, separated by empty space?

The classical theory says that an electron is surrounded by an electromagnetic field which influences the second electron. Quantum electrodynamics instead says that the electric force is transmitted by a mutual exchange of photons (quanta of light). Why, then, is an electron at rest not observed to emit electromagnetic waves (or photons)? There are photons exchanged by electrons at rest, but they are "virtual", which means here "not directly measurable": the Uncertainty Principle masks them from our view. Virtual photons explain the pure static force between two charges, when there is no acceleration and hence no radiation of real photons.

Richard Feynman 1918-1988

Virtual photons were invented by Richard Feynman. He is best known for his work on the quantum theory of electro-magnetic field. His *Lectures of Physics* are legendary textbooks. Every physicist of our generation has enjoyed their clarity and enthusiasm as they guide the reader towards a deeper understanding of physics. "Poets say science takes away from the beauty of the stars – mere globs of gas atoms. Nothing is 'mere'. I too can see the stars on a desert night, and feel them. But do I see less or more? ... What is the pattern, or the meaning, or the why? It does not do harm to the mystery to know a little about it. For far more marvellous is the truth that any artists of the past imagined!"

Virtual photons and other virtual particles form "Feynman's fur coat" around an electron. The electron at all times emits them in all directions and also absorbs others arriving from all around. In spite of this, there is

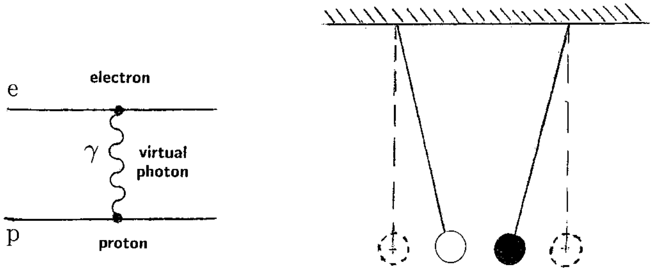


Fig. 5.3 The Feynman diagram explaining the Coulomb electric force $F = Q_1 Q_2/R^2$ between two charged particles. Virtual photons are the cause for the "action-at-distance" deflecting the hanging charges. Richard Feynman received his Nobel Prize in 1965.

no measurable change in the energy or mass of the electron. Why?

Virtual photons exist only so long as the Uncertainty Principle permits them to exist without being detected. The less energetic the virtual photon is, the longer it can exist and the farther it can travel from the electron. Because any particle cannot have less energy than its rest mass energy, the range of the force carried by a virtual (massive) particle cannot be arbitrarily large.

But photons are devoid of mass. This means that an electron may send such zero-energy virtual photons up to infinite distances. That is why the electric force has an infinite range of action, and has the inverse square distance dependence exactly as the empirical Coulomb law. [‡‡] But the strong force is strong only inside atomic nuclei. Caused by quanta having a non-zero rest mass, its grasp is limited to distances less than 10^{-13} cm.

Indeed, the strangest inhabitants "down there" are probably the *virtual particles*. Even empty space is filled by virtual particles, but their direct observation is prohibited by the Uncertainty Principle. Empty space, virtual particles – is this just another Much Ado About Nothing? Not at all. There are many observable consequences of the virtual particles, even if these culprits themselves keep out of sight. Empty space may even be a new source of energy which may be extracted from vacuum fluctuations. Hendrik Casimir, a Dutch physicist, suggested an experiment to show the

---

[‡‡] Feynman's fur coat of virtual particles has thickness $R \approx c\Delta t = ch/\Delta E$ where the lifetime Δt of a virtual particle comes from the Uncertainty Principle: $\Delta E \times \Delta t \geq h$.

reality of the vacuum force by using two parallel metal plates. Between the plates the vacuum changes, which is seen as a force moving the plates closer together. This experiment was first successfully performed in 1958.

5.16 Spiraling down into the microcosm?

Modern physics has found several fundamental limits in Nature, which restrict the imaginable properties of physical phenomena. The speed of light turned out to be the highest velocity with which energy or information may be carried. Classical particles with definite positions and velocities, our familiar friends from Newton's theory, no longer exist in the strange quantum microworld which started to surround us. Even the vacuum, previously taken to be empty space, is now a bubbling ocean of virtual particles which constantly appear and disappear.

The atom, "indivisible", turned out to consist of protons, neutrons, and electrons. Then it took half a century to find that the proton, literally the "first element", itself is made of new atoms, the quarks. One wonders if this was the last step into the depths of the matter. Could quarks have internal constituents? The next generation tool for particle physicists, the CERN Large Hadron Collider, will be completed in 2005. It may unearth new information about this quirky species of the underworld.

We cited Hermann Weyl in Chapter 4 about the infiniteness of mathematical space "inwards". A similar idea as applied to the real world has fascinated imaginative thinkers who have populated the microcosm with whole worlds, sometimes complete with flourishing life. The discovery of tiny atoms and their nuclei has shifted such possible other levels of the universe to very "down" indeed: the size of an atomic nucleus is about 10^{-12} cm. This ratio of the Earth to an atom is thus about 10^{21}. And the emergence of the quantum laws shows that we cannot think about the microcosm in similar terms as we do in the macrocosm, putting a kind of barrier between us and the world on very small sizes. Something like that can be said of very large cosmical scales, as we'll see.

An apparently ultimate wall in the microcosm is encountered at the size called the Planck length, about 10^{-33} cm, where space itself must be considered as a quantum porridge, still incomprehensible for modern physics. There "geometry begins", to borrow the expression by Swedenborg, whose intuition did not permit a structured world below a certain level or scale.

It is curious to think that the size ratio "the Earth to a nucleus" is not far from the ratio "a nucleus to the Planck length". One huge step leads us down to the quantum world, and a second one takes us before the gate of the totally unknown.

5.17 From terrestrial to cosmic laboratory

In all its exoticity, the quantum field theory predicts very successfully phenomena which physicists observe in their accelerators and astrophysicists in high-energy cosmic rays. The physicists are confident of the validity of the quantum field theory and there is a sincere hope that all the forces of Nature may be described by one unified theory. Within the capabilities of contemporary physics, it belongs to the realm of theory – in words of the physicist Abdus Salam – whose time has not yet come.

The tests of the Grand Unified and Supersymmetric Theories demand extremely energetic particles, in order to penetrate deeper and deeper into the microworld. However, experimental particle physics seems to have reached its natural limits in producing high-energy particles in terrestrial laboratories. The great variety of scales existing in the astronomical world gives rise to extraordinary phenomena and objects which possess almost unlimited energies. Hence, a new epoch of physics has opened when the astronomical universe has become a part of physical laboratory.

* * *

In our times, physical experiments have reached the cosmic scales. It is hoped that this gives the chance to understand the physics of fundamental forces, including gravitation, the force making the structures in the large universe. This is no easy task, but as Albert Einstein once said "God is subtle, but not malicious". Hopefully, Nature does not in some crooked way try to resist scientists who want to know Her secrets, something which apparently did happen to the poor astronomer Malyanov in the fine specimen of science fiction *One billion years before the end of the World* by Arkadij and Boris Strugatskij... Experimental method is scientist's manner of asking the taciturn world, whether his theory is correct.

Chapter 6

Gravity – the enigmatic creator of order

A pencil falls to the floor, a most familiar and at the same time a very puzzling phenomenon! Gravitation is ever present on the Earth. But its true mansion is the cosmos. Indeed, Newton's law of gravitation was revealed by the planets, and for Einstein's theory the first success was to explain why the innermost planet Mercury did not like to stay on its closed ellipse. In the microcosm of elementary particles gravity is much weaker than other forces. Only heavenly bodies contain such huge numbers of atoms that gravity becomes the number one force, which binds cosmic matter, making planets, stars, and galaxies. And a cloud of gravitating particles is a much richer system than just a chaotic swarm of colliding ping-pong balls – it is a system in which all the particles are as it were colliding all the time.

But what is gravitation, the untiring architect, after all? Is gravity like the other quantum forces or perhaps not a force at all, but rather a property of space–time? Answers will soon come from astronomy, which is presently experiencing the renaissance of gravity physics, with observations of exploding supernovae, gravitational waves, and quasars. The nature of gravity will likely turn out to be deeper than we yet understand.

* * *

6.1 The nature of gravity

What is the cause of the attraction between two masses? Why does an apple actually fall? Newton was rather silent on the matters raised by this simple sounding question. Though, in a letter to Richard Bentley one reads:

That Gravity should be innate, inherent and essential to Matter, so that one body may act upon another at a Distance thro'a a Vacuum without the Mediation of any thing else, by and through which their Action and Force may be conveyed from one to another, is to me so great an Absurdity, that I believe no Man who has in philosophical Matters a competent Faculty of thinking, can ever fall into it. Gravity must be caused by an Agent acting constantly according to certain Laws; but whether this Agent be material or immaterial, I have left to the Consideration of my Readers.

The note reveals that Newton found it very difficult to regard gravity as a "pure" action at a distance through empty space. There must be some Agent or carrier of gravitational force, the nature of which he left open.

Newton's theory of gravity gives a good description of the motion of planets about the Sun. However, it does not contain a mediator of the gravity force from one body to another. Moreover, when a pine tree drops a pine needle, the change of gravity is at once felt by all the galaxies in the universe. Such a swiftness violates relativity theory.

General relativity was the first relativistic theory of gravity, created by Albert Einstein in the 1910's. Since then physicists have understood gravity as an effect of the geometry of space. Newton's apple falls because the space around the massive Earth is curved. There is no need for a gravity force acting "thro'a a Vacuum". Einstein's theory explains why there are small deviations from Newton's predictions in the motions of planets and binary neutron stars. But it does not include quantum effects.

Field gravity theory was presented by Richard Feynman in the 1960's as a quantum description of gravity, similar to other fundamental forces of the microphysics. It explains gravity as a force, caused by the exchange of quanta, or gravitons, through space. The proverbial apple falls because there is a traffic of gravitons between the apple and the Earth. Gravitons play the role of Newton's secret agents.

6.2 Newton's law and the gravitational constant

The gravity theories of the future will always contain the constant G which determines the attractive force (GmM/r^2) between two masses in the usual cosmic conditions of weak gravity. G is rather small so that things in our near environment, such as shoes and refrigerators, cause only tiny forces on each other. That is why more than a century elapsed after Newton's

Principia before his law of gravity was confirmed in the laboratory and the value of G measured: in 1798 Henry Cavendish "weighed the Earth". This wealthy private experimentalist used a torsion balance apparatus constructed by John Michell (who, by the way, was the first to think about strong gravity objects from which light cannot escape – today's black holes). Modern measurements give the gravitational constant the value:

Henry Cavendish 1731-1810

$$G = 6.67 \cdot 10^{-8}\ \mathrm{cm}^3/\mathrm{g\ sec}^2$$

Measurements in laboratories, the Solar System, and binary neutron stars support the view that G is universal for different heavenly bodies. It cannot change faster than a mere 1 percent in 10 billion years, otherwise this would be noticed by astronomers studying the evolution of stars.

Theoretically speaking, Newton's gravity force is an ordinary force whose effect is felt in an inertial frame. But what is inertia itself? It appears in Newton's 2nd law of motion, where the force required to accelerate a body, relative to an inertial frame, is proportional to the "inertial mass". In 1872, Ernst Mach put forward the idea that the inertia of a body is determined by all other matter in the universe. This *Mach's principle* attempts to link inertia, a local property, to the global distribution of matter. But still in our times the origin of inertia remains a deep riddle.

Ernst Mach 1838-1916

6.3 The riddle of inertial and gravitating masses

Four centuries ago a discovery was made, which in essence says that a heavy mill-stone and a tiny mustard seed fall with an equal acceleration. If dropped simultaneously from the top of the leaning tower of Pisa (placed in a vacuum...), they both hit the ground at the same moment. This observation, often ascribed to Galileo, fore-shadowed the end of Aristotelian physics and paved the way for the modern understanding of gravity.

Actually the far-reaching experiment was made by Simon Stevinus. The Dutch–Belgian mathematician reported in 1586 that bodies with different masses fall with the same acceleration. Smiling astronauts floating in their spaceship are a modern example. This "weightlessness" is *not* due to a weak gravity force (the distance of the satellite from the Earth's center is only slightly larger than the Earth's radius). The reason is that they share with the spaceship the same acceleration in the gravity field of our planet.

Simon Stevinus 1548-1620

That the gravity force equally speeds up different masses, has a particularly deep consequence: the inertial mass in the law of motion is the same mass that appears in Newton's gravitation law – such different things and yet as if identical! * This law of equal inertial and gravitating masses has been tested with accurate torsion balances in experiments utilizing the daily rotation of the Earth. The Hungarian Baron Lorand Eötvös hung two weights, one made of wood, the other of platinum, from the ends of a 40-cm beam suspended by a fine wire. The Earth's gravity imparted the same acceleration to both wood and platinum, with an accuracy of one part in 10^9. Robert Dicke, in Princeton, and Vladimir Braginskij, in Moscow, have confirmed that inertial and gravitating masses are equal with the high precision of one part in 10^{12}. Employing the Moon–Earth–Sun system, the modern accuracy record is ten times higher.

Lorand Eötvös 1848-1919

Robert Dicke 1916-1997

6.4 Relativistic gravity emerges in our Solar System

Observations in near space first raised suspicion that Newton's majestic theory of gravity was not the final word. Nowadays we know in the weak † gravity field of the Solar System several effects, tiny but measurable, which cannot be explained by Newton's theory. These phenomena are the basis for the construction of relativistic gravity theories. They are:

- *an extra 43 arc sec perihelion advance of the orbit of Mercury*
- *the bending of the light of distant stars by the Sun*
- *the gravitational shift of spectral lines*
- *the delay of radio signals passing by the Sun*

Orbiting so close to the Sun that it is hard to see from the Earth, Mercury feels the great mass in a manner that calls for new understanding. One of the discoverers of Neptune, Urbain Le Verrier found that Mercury did not travel along its orb quite as expected. The orbit was like an ellipse which slowly rotates. During one century, the point of the closest approach

Urbain Le Verrier 1811-1887

---

[*] The force in the Law of Motion $F = m_i a$ is now the gravity $F = Gm_g M/R^2$. Gravitating and inertial masses are the same ($m_i = m_g$), so they cancel out and the acceleration does not depend on the mass of the falling body: $a = GM/R^2$ where M is the Earth's mass and R is the distance from the Earth's center to the body.

[†] The gravity field is weak, if it gives a body a velocity which is much less than the speed of light, or the gravitational potential energy of the body $E_{pot} = -GmM/R$ is much less (in absolute value) than its rest mass energy $E_0 = m_o c^2$, i.e. $|E_{pot}| << E_0$.

to the Sun (the *perihelion* point) advances an angle of 575" relative to the inertial frame defined by distant stars. The Newtonian gravity forces of the other planets account for 532" of this shift. The remaining 43 seconds of arc were a deep mystery until relativistic gravity arrived.

The solar gravity deflects the light ray from a star passing by the edge of the Sun. The bending angle is predicted to be 1.75 seconds of arc by relativity theory. Because stars are seen close to the Sun only during an eclipse, this test is not easy. The bending was first detected in 1919 by British expeditions to the village of Sobral in Brazil and to the Portuguese island of Principe off the west coast of Africa, organized by Arthur Eddington. At that time only two competitors to general relativity were known. The relativistic scalar field theory of the Finnish physicist Gunnar Nordström did not predict any bending of light. The Newtonian theory did predict a deflection, but only half of the general relativistic value. Hence, a limit of validity of these theories was found in the Solar System, in the same cosmic neighborhood in which classical physics was revealed to us.

Gunnar Nordström 1881-1923

The news of the result were enthusiastically received by physicists as a triumph for general relativity. This also made the name of Einstein widely known among the general public.

The shift of spectral lines in the Earth's gravitational field was first measured by Robert Pound and Glen Rebka in 1959 in a 24 meter high tower at Harvard University. As reported in their article "Apparent weight of Photons", the effect was verified to an accuracy of 1 percent. It is hard to reach a higher precision, because the wavelength shifts are minute: about 2.5×10^{-15} in the Earth's weak gravity. The short ($\sim 10^{-5}$ sec) delay of radio signals which pass close to the Sun was detected by Irwin Shapiro in 1968, in a radar experiment with Mercury and Venus as reflectors.

6.5 Geometry of curved spaces

General relativity is based on the new mathematics of *curved spaces*, which was worked out in the 19th century by Gauss, Lobachevskij, and Riemann. This concept appeared after mathematicians had for two millennia tried to derive Euclid's fifth postulate from the other four ‡ and had all failed. This

---

‡The other postulates: 1. A straight line can be drawn from any point to any point. 2. A finite straight line can be produced continuously in a straight line. 3. A circle may be described with any center and distance. 4. All right angles are equal to one another.

Parallel Postulate is stated as follows:

- *Through a given point in a plane one can draw only one line parallel to a given line in this same plane.*

To be parallel means that the two lines in the same plane do not intersect anywhere. At first the enigmatic importance of the Parallel Postulate may elude the eye, and it may even seem almost painfully obvious and non-interesting for the "Euclidean mind". That Euclid included it among his basic axioms is a testimony to his greatness as a mathematician.

Nikolaj Lobachevskij 1793-1856

In 1829 the Russian mathematician Nikolaj Lobachevskij, professor and rector of the university of Kazan, constructed a logically consistent geometrical system in which Euclid's parallel postulate was replaced by another assumption, now not at all so "obvious". Lobachevskij's postulate was:

- *Through a given point in a plane, there can be drawn an infinite number of lines, which do not intersect a given line in the plane.*

Lobachevskij called his revolutionary geometry "pangeometry" (from the Greek "pan" meaning "all") and entertained the philosophy that any field of mathematics, no matter how abstract it is, will eventually find application in some phenomena of physical reality. But reality was not yet ready for these (or perhaps some more dangerous) ideas, and he was relieved of his job as head of the University, with no explanation whatsoever.

Janos Bolyai 1802-1860

The same mathematical discovery was made independently in 1832 by Janos Bolyai, a young Hungarian army officer, whose result was published in an appendix to his father's book. When Lobachevskij's book on geometry was translated into German in 1840, Bolyai got upset because he lost his priority. He published no mathematics again.

Carl Friedrich Gauss 1777-1855

As far as historians of mathematics know, the German mathematician and physicist Carl Friedrich Gauss was the first man who believed in the independence of the parallel postulate. Gauss approached the question using algebra and calculus and he developed the new mathematical discipline of differential geometry.

Differential geometry began as the study of curved surfaces embedded in 3-dimensional Euclidean space. The most simple example of a curved surface is given by a sphere. In 1827 Gauss showed how measurements of angles and lengths on the surface can be used for deriving the radius of curvature of the sphere in the embedding space. The sum of interior angles of a triangle drawn on the spherical surface is greater than the Euclidean value of 180 degrees by an amount which depends on the area of the triangle and the radius of curvature of the sphere. [§]

There is a story that Gauss tested the method by measuring the radius of the Earth. Using three places forming a triangle with an area of 2700 km^2, he measured that the sum of the angles was 180 deg plus 14 arc seconds. His formula then gave that the radius of curvature is 6400 km. This was typical for Gauss who always tried to connect theory and practical physics. All physicists use his system of units *centimeter, gram, second.* Gauss also was the first to construct an electromagnetic telegraph in 1833. The line connected the physical institute and the astronomical observatory (of which he was head) in Göttingen. This new application of electricity was soon to replace the system of the optical telegraph, about which Alexander Dumas so interestingly related in one episode of "The Count of Monte Cristo".

Bernhard Riemann 1826-1866

The differential geometry of surfaces was extended to n-dimensional curved spaces in 1854 by Bernhard Riemann, a pupil of Gauss. The most symmetrical Riemannian spaces are spaces of constant curvature. The properties of such spaces do not depend on position or direction, i.e. they are uniform and isotropic. There are only three types of such spaces, corresponding to three possibilities for the curvature: *spherical* with $K > 0$, *Euclidean* with $K = 0$, and *hyperbolic* with $K < 0$. Lobachevskij's space, with its infinity of parallels, is hyperbolic. If one assumes that there are no parallel lines at all, then one has constructed a spherical space, the prototype of the first modern world model by Einstein.

The most spectacular feature of curved spaces is that in such spaces figures of different sizes lose their similarity. For instance, triangles of different areas on a sphere have different sums of their angles. Similarity exists only in Euclidean space.

6.6 General relativity as geometrical gravity theory

In 1915 Albert Einstein made "probably the greatest scientific discovery" (in words by Paul Dirac) – he found new equations for the description of gravity, which showed that physical space-time is curved by the masses

---

[§] The sum of the angles is $A+B+C = 180^o + S/R^2$, where S is the area of the triangle and R is the radius of the sphere. Gauss introduced the general notion of *space curvature* K, which for a sphere is simply $K = 1/R^2$, and now R is called the radius of curvature.

embedded in it. A curved space-time is no longer the flat Minkowski space of special relativity. In Einstein's theory the amount of curvature depends on the presence of matter in a definite manner:

CURVATURE = constant × ENERGY-MOMENTUM OF MATTER

In the simplest case, energy-momentum is essentially the density of matter. The simple-looking formula hides the deep fact that curvature and energy-momentum are *tensors*, not usual numbers, but 4 × 4 matrices. Instead of one, there are ten independent equations. [¶]

The idea of gravitation as geometry came to Einstein, when he realized the significance of the mill-stone and mustard seed falling side by side. From this empirical fact Einstein made a general conclusion which he called the *Principle of Equivalence.* In essence it states that all effects of a (uniform) gravity field can be reproduced in an accelerated frame. Inside an ascending rocket, with eyes closed, one cannot easily decide if the rocket is speeding up or if by some magic, gravity is suddenly increased in the launch location. And in a rocket accelerating in space, all things are seen to fall towards the floor with the same acceleration, well imitating gravity.

Gravity can not only be "created" in a suitably accelerating reference frame, it can also be eliminated, at least locally, in a frame which is freely falling in that gravity field. For instance, in the famous Einstein's lift, rushing down from the top of, say, the Tower of Eiffel, all physical processes happen exactly like in an inertial frame *without* gravity. That the gravity force can thus be eliminated, while other forces cannot, led to Einstein's wonderful idea of treating gravity as something deeper than an ordinary force – it is inertial motion in curved space.

Vladimir Fock 1898-1974

Historically, the equivalence principle played a key role in the birth of general relativity. However, as Vladimir Fock, the renowned physicist from St. Petersburg, liked to emphasize, the true basis of general relativity is the assumption that space around massive bodies is non-Euclidean. [‖]

[¶] Briefly: $\hat{G} = (8\pi G/c^4)\hat{T}_m$. $\hat{G}$, Einstein's tensor, describes the curvature of space–time caused by the energy–momentum of matter $\hat{T}_m$. This equation was at the same time independently derived by David Hilbert (1862-1943) using a different method, from the Principle of Least Action. The "cap" tells that the quantities are tensors!

[‖] One of us (Yu.B.) had the fortunate chance to attend the last lecture delivered by Prof. Fock at the old Physics Department of St. Petersburg State University on the island of Vasily. Fock said that the correct name for Einstein's gravitation theory is "Chronogeometry", because it essentially deals with space and time.

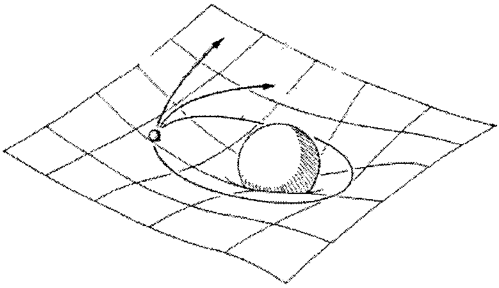


Fig. 6.1 The rubber membrane analogy for the geometrical understanding of gravity. We mention in the text that this analogy is somewhat misleading as it requires an external force (which is the very force we are trying to understand!)

6.7 What causes gravity according to general relativity?

This tricky question is often answered with the analogy of a rubber membrane curved under a heavy weight. A stretched sheet of rubber represents flat space. If one puts on it a lead ball, the membrane gets curved and forms a funnel around the ball. Now, a light nearby object, say a wooden ball, will strive towards the funnel, as the space (rubber membrane) is curved. But such an explanation of geometrical gravity seems confusing, because it is based on an external force which is also the reason why an initially resting particle is put into motion. But *why* does it actually start moving?

We illuminate the geometrical gravity with simple thought experiments. First, imagine that there is no mass nearby: the particle stays at rest. However, in Minkowski space-time it has "motion" along the time axis – this is how one sees things in space-time: even if at rest in *space*, the particle propagates in *time*, hence it moves in *space-time*.

Then, switch on gravity: the "shortest" trajectory in what is now curved space-time is no longer along the time-axis. Hence also distance must change: the particle moves in space towards (or away from) the mass.

The "shortest trajectory" is more exactly termed "geodesic line". It has a so called extremal length between two points of space-time. That a particle follows this trajectory is dictated by the Principle of Least Action. Hence, the motion of particles in curved space-time is described as a *free motion* along geodesic lines. This leads to an exceptionally important conclusion: what is called gravity is actually not a force, but a natural free motion of a particle in a curved space. Other forces, such as the electric force, manifest themselves as deflection from a free (inertial) motion.

For example, if a particle stays at rest near a mass, there must be an ordinary force acting on it in order to keep it from falling. Turn off this force, then the state of the particle changes into natural free motion, which leads it towards the mass.

In precise mathematical language, Einstein's equations tell how a mass curves space and what is the free motion of a particle in it. Two test particles having different, but tiny masses, do not curve space: they follow the same path over the curved landscape. In this simple fashion geometric gravity explains why the mill-stone and the mustard seed fall side by side. This resolves the mystery of equal inertial and gravitational masses.

6.8 Big Bang, Black Hole, Time Machine...

General relativity was born when only one of the relativistic gravity effects, the rotation of Mercury's orbit, was known. It predicted also other tiny effects later measured in the Solar System. New examples of its impressive accuracy come from binary neutron stars where all such deviations from Newton's theory are much bigger than around the Sun. The detected phenomena have demolished any lingering hope that classical physics could describe the cosmos in its entirety. Spectacularly, general relativity opened the door to self-consistent world models, free of the paradoxes of Newton's cosmology. The start was in 1917 when Einstein filled the world uniformly with matter and added his famous cosmological constant to the equations of general relativity. This constant was required by a static distribution of matter. The geometry of Einstein's universe was that of spherical space. It has a finite volume similarly as a sphere has a finite area.

In 1922 Alexander Friedmann found a new type of cosmological model which allowed the space to collapse or expand. Such expanding models are now the basis of big bang cosmology. General relativity says that such universes where all matter is distributed uniformly were born from an enigmatic singularity, where matter, space, and time were created.

The success in explaining the gravity phenomena in the Solar System and binary stars, and in constructing world models has made general relativity the standard tool for studying also the things which occur in strong gravity fields, such as exist near extremely compact celestial bodies. In such situations, the theory predicts extraordinary things which attract physicists, and inspire science fiction: the Black Hole and the Time Machine.

Imagine our Sun squeezed down to a radius of three kilometers. Then its light would be locked in by its gravity and it would be a black hole. Other physical forces cannot withstand infinite geometrical gravity and all matter reaching down to a critical distance from the center, inevitably collapses into the central point, a singularity.

In Chapter 5 we encountered the "twin paradox" in special relativity. There one may travel to the future of one's non-accelerated home by means of a high velocity trip. This reasoning is based on the established laws of special relativity, and is not concerned with time travel to the past. For such trips to the future an honest travel agent can sell one-way tickets only.

Much more is promised by general relativity – a truly weird thing, the time machine. Certain solutions of Einstein's equation are called "wormholes". Using such channels in space, it is in principle permitted to organize journeys in time to the past and to the future. Unfortunately, opening such tourist routes would require a non-Euclidean space with complex topology which is only attainable in the inhospitable vicinity of black holes.

6.9 . . . but riddles still exist

Force, work, and energy are tightly interconnected for all forces in physics. However, as Willem de Sitter emphasized already in one of the first articles on general relativity, geometrical gravity means the elimination of gravity force from physics, and "gravitation is thus, properly speaking, *not* a 'force' in the new theory". Therefore, in general relativity not only force, but also energy (work made by force), loses its ordinary sense for gravity. This has led to a long standing "energy crisis": one cannot define the distribution of the energy of the gravity field around a body. ** In contrast, there is a definite electric field energy in space around an electric charge.

The idea that the gravity force is totally different from other forces is

---

**This relates to the conservation laws. Emmy Noether (Ch.5) showed that conservation of energy and momentum follows from the symmetry of the Minkowski space–time. If gravity curves space, it is impossible to derive a single conservation law for energies of matter and gravity. Technically, this is because energy–momentum of the gravity field is not a tensor, but a pseudotensor. See Sect.101 in Landau & Lifshitz: *The Classical Theory of Fields* (1971), often regarded as the best presentation of general relativity.

revolutionary. And revolutions are accompanied with sacrifice as is all too well known from history – here the victim was the concept of the energy of the gravity field. The search for quantum gravity has made this fact topical. Quantization of other fields (electromagnetic, weak, strong) leads to energy quanta, and if in the theory there is no notion of energy, one wonders what remains to be quantized. As Feynman once noted, if general relativity is left without quantization, this would lead to the paradoxal situation that in Nature at the same fundamental level are living side by side hazy quantum particles and sharp classical trajectories (geodesic lines). But deep obstacles hamper the quantization of *geometrical* gravity. One has to face the quantum structure of space and time, about which physicists have no previous experience. Understanding other interactions has required quantization of matter only, leaving space and time intact. Is it possible to arrive at quantum gravity by this familiar route?

This important question was discussed by Feynman (see below) and has recently received new impetus from experiments with ultracold neutrons, which suggest a deep similarity between electromagnetic and gravity fields. Valery Nesvizhensky and his colleagues at Institut Laue–Langevin in Grenoble (France) have demonstrated for the first time that the gravitational field of the Earth gives rise to quantum states for very slowly moving neutrons. The experiment, which required twenty years of preparations, showed that not all trajectories are permitted in a gravity field, but only those which are compatible with the predictions of quantum mechanics. Recall how a similar thing occurs – on a very much smaller scale – with the electrons orbiting atomic nuclei under the attractive electromagnetic force.

6.10 Feynman's quantum field approach to gravity

In the *Lectures on Gravitation* which he delivered at the California Institute of Technology in 1962, Richard Feynman put forward a view on the physics of gravity, which conceptually differs from the geometric picture.

Feynman's field gravity studies quantum gravity from the view point of particle physics. It regards gravity as a physical interaction described by a quantum field, which so splendidly works for other forces. Space-time is a permanent stage on which the quantum gravity field is a basic ingredient and gravitons are just bits of energy and momentum of the field. Furthermore, the Minkowski stage (space–time) is the most symmetric space

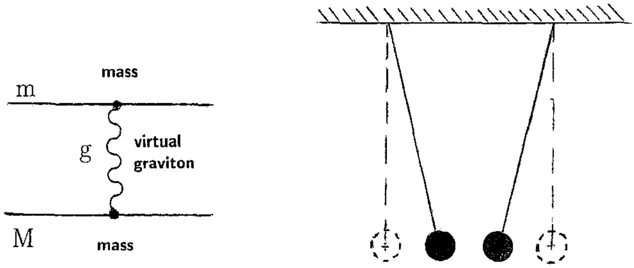


Fig. 6.2 What causes gravity according to Feynman's quantum field theory? As in the case of other fundamental forces, gravity is carried by the quanta of the gravity field, or gravitons, giving rise to Newton's gravity force $F_N = GmM/R^2$.

and, says Noether's theorem, preserves the ordinary conservation of energy-momentum for the gravity field. Feynman emphasizes that "the geometric interpretation is not really necessary or essential to physics", though he wonders why the field theory also has a deep geometric meaning which "is not something readily explainable – it is just marvelous".

In his lectures Feynman tried to show that the field interpretation, though pedagogically useful, will lead to a theory which is identical with general relativity. These two theories do make common predictions for classical relativistic gravity effects. However, as is clear from a comparison with the theory of other fundamental forces, the quantum field description delivers a genuinely complementary understanding of gravity physics, with the concepts of force, energy, and quanta as basic elements.

In fact, the idea of a field approach was guessed already by Henri Poincaré. In 1905 he pointed out that all forces in Nature, including gravity, should be manifestations of the relativistic field and propagate with velocities less than or equal to the speed of light. A field theory may be constructed from different kinds of fields, technically speaking scalar, vector, and tensor fields. For example, electricity is described by a vector field. The theory of Nordström was a scalar theory. The remarkable unique feature of the gravity force is its universality – it interacts with all kinds of matter, feeling their mass and motion (energy-momentum). (By the way, from this universality and the Principle of Least Action, it follows that the inertial and gravitational masses are equal.) How can one treat such an omnipotent force as a field?

George David Birkhoff 1884-1944

The key is that a *tensor field* can adequately describe gravity, as was shown by the mathematician George Birkhoff in 1943. Then all the special properties of gravity naturally appear. Similarly as the force law between charged particles is due to quanta of the electromagnetic field with zero rest mass (photons), the gravity force has the same inverse square law form if gravitons have zero rest mass, moving with the speed of light. As a follower of Poincaré in classical dynamics, Birkhoff also wrote seminal articles on chaotic processes. Interestingly, he made fundamental contributions to both gravitation and chaos, which are so closely related to cosmic fractals.

6.11 Relativistic effects in quantum field gravity

Feynman emphasized that the field gravity theory in the conditions of weak gravity has the same equations as general relativity. Hence it predicts the same classical relativistic gravity phenomena. But the physical interpretation is different. Birkhoff's pupil Marcos Moshinsky, who was born in the Soviet Union in 1921 and early moved to Mexico, showed in 1950 that the interaction of light with the gravity field predicts the same value for the deflection angle as general relativity. However, now the reason for the bending light is that the gravity field around the Sun acts as a kind of lens.

All the other relativistic effects of a weak gravity field are also explained by field gravity. Interestingly, the perihelion shift of Mercury's orbit may be regarded as a measurement of the energy of the gravitational field around the Sun. This energy produces 16 percent of the shift. [††]

In the extreme conditions of very strong gravity the field approach is expected to differ much from the geometrical one. For instance, in field gravity theory black holes do not exist as singularities as they do in geometrical gravity. Another important consequence of field gravity is the existence of a new type of gravitational wave, called scalar waves. Observations may be able to test these predictions in the near future (more on singularities and gravitational waves in Chapter 9).

[††] Even the Newtonian force F_N has a different interpretation in the field theory. In the Newtonian limit the attractive force acting on a mass m in the potential Φ_N is the result of a contest between two processes in which two kinds of virtual gravitons exist: those having spin 2 (an attractive force $F_2 = -3/2\ m\nabla\Phi_N$) and those with spin 0 (a repulsion $F_0 = +1/2\ m\nabla\Phi_N$!). So the sum is $F_N = F_2 + F_0 = -m\nabla\Phi_N$. Applications of the field gravity theory to weak and strong gravity fields in astronomy have been studied in a series of articles by one of us (Yu.B.) together with Vladimir Sokolov.

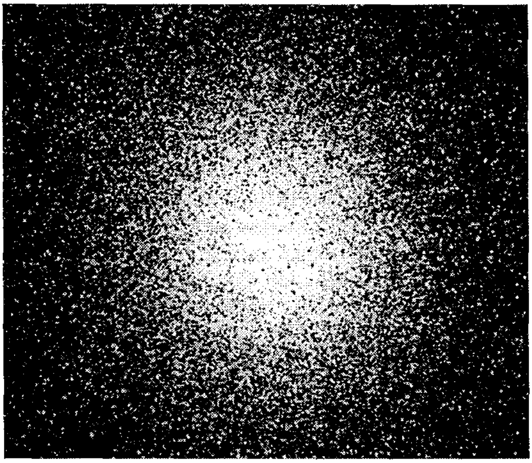


Fig. 6.3 The globular cluster ω Centauri keeps it spherical form thanks to the mutual gravity of about a million stars. Astronomers use such old relics from the youth of the Milky Way to calculate the age of the universe. Photographed by Tapio Korhonen with the Danish 1.5-m telescope at La Silla, Chile.

6.12 Gravity as a builder of celestial structures

Newton's law of gravity is deceptively simple. But as we saw in Chapter 2 even the mutual motion of just three bodies is unpredictable (more on that in Chapter 15!). And nobody can imagine what the detailed evolution of a cloud of a huge number of gravitating bodies is.

Ordinary gas may be viewed as a swarm of particles which interact (elastically as billiard balls) only when they collide with each other. This is a good model for the gas in terrestrial and cosmic conditions. Each gas particle is moving freely most of its life and only for short moments does it violently collide with other particles which happen to cross its way. A hit makes the particle change the direction and speed of its motion. For such briefly interacting particles, thermodynamics predicts that the gas strives towards its most probable state: thermodynamic equilibrium. In this state all particles are uniformly scattered inside the container holding the gas.

What happens when one turns on gravity between the particles? Such a substratum, a *self-gravitating gas*, is quite different from ordinary gas. Now the gravity may hold together the gas, forming beautiful structures, such as globular star clusters (see Fig.6.3). And because the gravity of any one particle is felt by all the others even from large distances, all particles are as if colliding all the time. The resulting collective phenomena may be studied using modern supercomputers. The very efficient computer can follow, step by step, how each particle moves under the gravity force of all its companions. Such calculations show that starting from a chaotic initial state, Newtonian gravity creates complex structures of different sizes.

Another way to the secrets of a self-gravitating gas is to construct the theory of how such a gas behaves. This is a difficult field, but recently advances have been made. Astronomers from Paris, Hector de Vega, Norma Sánchez, and Françoise Combes showed that the state of equilibrium of a self-gravitating gas is not a uniform distribution, but fundamentally lumpy, even "hierarchically clustered". So, a miracle happens in a cloud of particles. Gravity creates "fractal order" from indifferent chaos. We explain in Part IV what fractality is, but wish here to point out, how even the familiar and simple Newton's gravity law is a source of unexpected phenomena. The eagerness of gravity to build clusters and clusters of clusters is exploited in modern cosmology, in which gravity is the maker of very large structures.

6.13 Energy flows and order from chaos

The idea that order may arise from chaos is not new in physics and is not limited to the effect of gravity. Ilya Prigogine, the Russian-born physicist and Nobel-prize winner, has deeply studied the behavior of "ordinary" gas (which can be thought of as ping-pong balls without gravity). He has concluded that in Nature the law of entropy growth (the tendency towards disorder and equilibrium; Ch.3) is not so all-embracing as has been thought. Yes, it holds when in the beginning the physical system is not too far from equilibrium, but if the system has already some structure and significant energy flows inside it, it tends to evolve towards more order and structure in an island of decreasing entropy.

A well-known example of such self-organization is the formation of convection cells in a fluid through which thermal energy flows (say, as happens every day in a pot of water on a hot-plate). A cosmic case is the surface of the Sun where the convection carries energy from the hot interior and produces a pattern of granules with a typical size of 1000 km. Convection transmits energy more efficiently than normal conduction, and Nature likes economy. Thus the formation and maintenance of structure in a system of gas or fluid opens a new channel for the energy to flow.

The most striking example of order existing thanks to a continuous stream of energy, is offered by all of us and the life flourishing around. In the 100 C temperature of a sauna the human body does not passively tend towards a thermodynamic equilibrium with its environment, but actively maintains its normal temperature – and even gains pleasure as a by-product. No wonder Prigogine calls life as "the supreme expression of the self-organization process".

6.14 A star is a self-gravitating nuclear reactor

Classical physics was never able to explain what makes the Sun shine. It was not even known how long the source of energy, whatever it was, should keep the Sun and other stars hot. Now we know, from age determinations of the Earth, that the Sun has radiated almost steadily for about five billion years. The requisite energy cannot have been supplied by ordinary means. For example, if made of coal, the Sun would burn out in 6000 years.

Relativity theory revealed the unexpected thing that any particle contains a huge energy ($E = mc^2$). This was the key. If even a small fraction of the Sun's mass has changed into energy, its energy flow over five billion years could be understood. In fact, a star's life is a continuous struggle against the force of gravity, which attempts to make the star collapse. The outward flowing energy from the very hot interior of the star gives rise to the pressure force which balances the force of gravity.

In 1938 George Gamow organized a meeting in Washington in which scientists could discuss the origin of stellar energy. Hans Bethe, who when he came "knew nothing about the interior of stars, but everything about the interior of the nucleus" (in the words of Gamow), was much inspired by the event and within six months found a mechanism for energy production for stars heavier (and hotter) than the Sun. Independently, Carl Friedrich von Weizsäcker also discovered this carbon-nitrogen-oxygen cycle, where carbon acts as a catalyst in a conversion of four protons into one nucleus of helium. For stars like our Sun or smaller, the major energy source is the *proton-proton reaction*, also leading from protons to helium. Gazing at Sirius, you may see the carbon-nitrogen-oxygen cycle, but on a hot day you cannot but feel the proton-proton reaction!

The stars are gigantic nuclear reactors which release by fusion a small fraction (less than one percent) of the dormant mass energy. Besides producing energy, the nuclear reactions serve as builders of heavy elements such as carbon, nitrogen, oxygen, and others up to iron. In their classical article "Synthesis of the Elements in Stars", Margaret & Geoffrey Burbidge, William Fowler, and Fred Hoyle described in 1957 how this takes place.

The matter in the Sun is a good sample of the normal stuff in the universe. The big planets of our Solar System are also made of "Sun-matter", while the small ones, Mercury, Venus, Earth, and Mars have a different composition. The table below compares the most abundant chemical elements in the Sun and the Earth, shown in percentage by mass.

Sun		*Earth*	
element	percent	element	percent
hydrogen	77	iron	35
helium	21	oxygen	30
oxygen	0.8	silicon	15
carbon	0.3	magnesium	13
neon	0.2	nickel	2.4
nitrogen	0.1	sulphur	1.9
iron	0.1	calcium	1.1

So much hydrogen and helium and so little of the heavier elements – why? The answer given by modern cosmology is that the light ones come from the early universe, whereas the heavy-weights were cooked inside the stars.

6.15 Exploding stars – the end of the fight?

Stars which are heavier than the Sun, live a short time, cosmically speaking. And they spend their whole lives fighting against gravity. What happens when they die?

In July 4th of 1054 Chinese and Japanese star gazers recorded the violent death of a star. An unknown star appeared in the constellation of Taurus. It was so bright that it was visible in the daytime for 23 days and then it waned and disappeared. After inspection of the Chinese accounts, Knut Lundmark proposed in 1921 that this "guest star" caused the Crab Nebula (Messier 1), which is now observed in the sky at the same place.

Indeed, observations have revealed that the Crab Nebula expands with

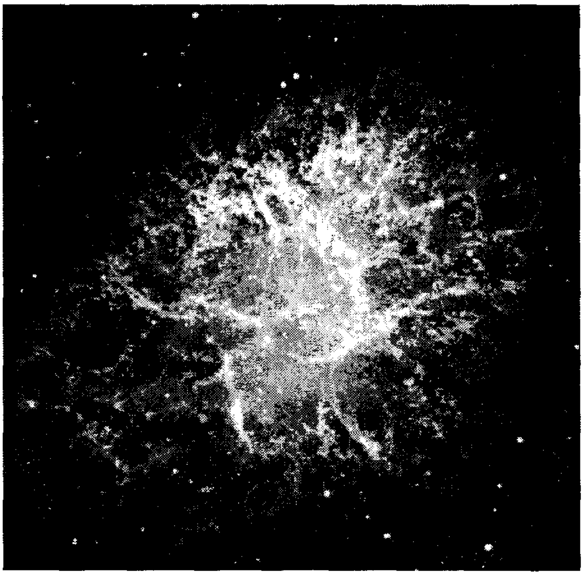


Fig. 6.4 The complex structure of the expanding remnants of the star which exploded as a supernova in 1054. In the center of this Crab Nebula a relativistic neutron star is observed as a pulsar. Photographed by Søren Larsen with the NOT telescope in 1999.

a speed of 1000 km/sec. One can calculate that 950 years ago it must have been compressed into a very small volume. The old witnesses really saw a gigantic cosmic cataclysm, the explosion of a star at the end of its life.

Knut Lundmark had made in 1920, with an "incredible foresight and imagination" as praised by Fritz Zwicky, a division between ordinary novae and 10 000 times more luminous giant novae, now called supernovae, in his study of the distance to the Andromeda galaxy. In 1934 Zwicky and Walter Baade came up with the idea that supernovae signal the death of massive stars, accompanied by an explosive release of enormous energy which produces an expanding nebula. A very compact remnant, a *neutron star*, is left in the middle.

Fritz Zwicky 1898-1974

Walter Baade 1893-1960

When a massive star has devoured its nuclear energy resources, the equilibrium between the internal pressure and the gravity force of the stellar matter is broken and the star collapses. If the star is not very massive, like our Sun, it eventually becomes a white dwarf in which the quantum-mechanical pressure of the electron gas maintains it in balance against gravity (the electrons fill up the available quantum states as compactly as possible). A massive star forms a dense core and ejects most of its matter as a shell expanding into space. Depending on the original mass of the star, the compact relic can be a neutron star or even a black hole as predicted by general relativity. In a neutron star gravity is so strong that it has forced electrons and protons to merge into neutrons; the star is like a giant atomic nucleus consisting of innumerous neutrons. Like in a white dwarf, a balance is maintained in the contest between gravity and quantum physics.

These celestial bodies have masses roughly similar to the Sun, but their sizes and densities are quite different. A white dwarf has a radius of about 10 000 km and density about 10^6 g/cm^3. Neutron stars have radii of only 10 km. Their densities of 10^{14} g/cm^3 are quite fantastic, but true: a spoonful of neutron star matter weighs as much as one million heavy locomotives...

The surface of a neutron star is made of very dense iron and it is almost smooth. But a "star quake" may occur when a few centimeter high "MtEverest" collapses. Such events explain the sudden changes in the regular pulses from *pulsars*, fast rotating neutron stars with very strong magnetic fields. Otherwise their rotation gradually slows down, when energy is carried away by the magnetic field sweeping the surrounding medium.

The neutron consists of quarks. Also quark stars might exist, as globes of extremely dense quark porridge. But the existence of such "superneutrons" depends on the nature of gravity. General relativity sets an upper limit to the mass of a neutron star so that a compact star with mass more than 3 solar masses inevitably collapses and becomes a black hole.

* * *

Betelgeuse, a red giant star in the constellation Orion, is living its last "moments" – it may explode as a supernova any night during the next few million years (indeed, it may have already done so...). At its distance of 200 parsecs this phenomenon, a feat of gravitation, will be seen in our sky as a magnificent star a hundred times brighter than the planet Venus.

The astronomy of the forthcoming decades gives the chance to decide which aspect of gravitation, geometry or quanta, is more fundamental, or if one can combine both within some deeper theory. Of course, the new theory should be dressed up in "Feynman's strait jacket": "At the same time the thoughts are restricted in a strait jacket, so to speak, limited by the conditions that come from our knowledge of the way nature really is. The problem of creating something which is new, but which is consistent with everything which has been seen before, is one of extreme difficulty."

Chapter 7

The law of redshift in the kingdom of galaxies

Two centuries passed after Edmond Halley's list of a total of six objects *Of Nebulae or lucid Spots among the Fix't Stars* from 1715, before it was cleared up whether the nebulae are distant galaxies, or whether they are foggy components of our Milky Way, of another kind than stars. This required larger telescopes, the invention of photography and spectroscopy, and enthusiastic work by generations of astronomers.

When the breakthrough came, in the 1920's, it was quickly realized that Man had entered a new unexplored world of large scales, with its own building bricks and structures, and hitherto unknown laws of Nature. Indeed, one of the greatest discoveries of the last century was the universal redshift in the light coming from remote galaxies. In 1929 Edwin Hubble showed that the redshift grows with distance. Though a modest thing, when inspected from the images of spectra, the cosmological redshift must tell about something magnificent. In modern cosmology redshift is ascribed to the expansion of the universe, a mysterious and majestic process.

* * *

7.1 The Island Universes

Around 1900, most astronomers thought that the nebulae reside within the Milky Way. This view was not based on distance measurements, such being not yet within reach. But there was an intriguing observation. When astronomers plotted on the sky the positions of the nebulae, they found that especially those showing spiral structure had a curious distribution.

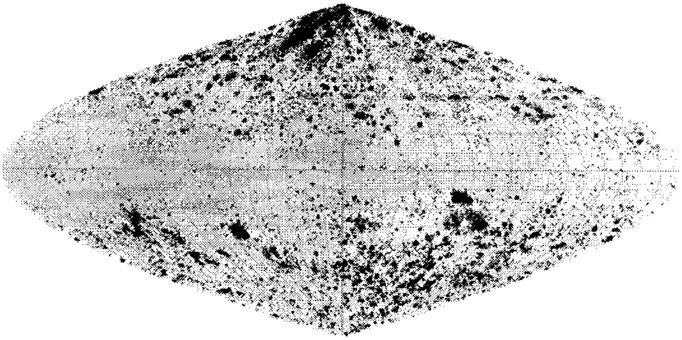


Fig. 7.1 The distribution of 11475 spiral nebulae in the sky map prepared by Carl Charlier in the beginning of the 20th century. Along the equator only few nebulae are seen through the dusty Milky Way. Charlier wrote: "A remarkable property of the image is that the nebulae seem to be piled up in clouds... Such a clouding of the nebulae may be a real phenomenon, [or it may be] caused by dark matter in space."

They were seen all around the sky, with the exception of the Milky Way band which they preferred to avoid! Why?

As the nebulae lie so symmetrically relative to the Milky Way, they must have a relation to it: they probably are a subsystem of the Milky Way and not distant Island Universes. So went a common argument. But for a modern astronomer this same map tells quite another story.

Interstellar space is not empty, but contains a lot of dust. Compact dust clouds form cocoons around baby stars, and a smoother dusty medium spreads between the stars. The cosmic dust is very fine – it is more like smoke, composed of particles having sizes of about 0.0001 mm. It is hard to see anything that lies behind a thick cloud of cosmic smoke. This happens when one looks towards the Milky Way band: the arriving light travels a long distance inside the disk made of stars, gas, and also of dust clouds, piling up one behind another, hence anything beyond the disk is strongly obscured. Armed with this knowledge, one can literarily see the shadow of the dust in the arrangement of the nebulae. The *Zone of Avoidance* becomes a natural thing in the flat and dusty Milky Way we inhabit.

The opinion about Island Universes started to divide during the 1910's. But only distances could decide whether the nebulae are outside the Milky Way. So how to measure the distance to a nebulous blot? One may try to see individual stars, if there are such, and then to measure their distances.

7.2 The distance to the "Little Cloud" is measured

If the nebulae are other Milky Ways, they are made of stars which are probably similar to "our" stars. If stars of a certain kind have identical luminosities in the Milky Way, such stars should have this same luminosity "out there", too. Knowing this luminosity, then the distance of a star can be calculated from its measured flux of light. This was how Knut Lundmark reasoned when he measured the distance to the Andromeda nebula in 1919. Even if named the Little Cloud by its early observer Al-Sufi (Fig.3.1), its large size in the sky (a couple of degrees) makes it a natural candidate for the closest Island Universe. True, two of his older colleagues in Sweden had succeeded, as they thought, in measuring its parallax, putting it at a short distance of 7 pc or at most 20 pc, but Lundmark went his own way.

Knut Lundmark 1889-1958

He knew that American astronomers had detected in Andromeda in the course of years ten novae ("new stars") which all had reached about the same brightness. As they all must lie at a similar distance from us, the constant apparent brightness means that also the true luminosity is constant from one nova to another.

Lundmark only knew that the luminosity is constant, but how large is it? Fortunately, several novae had also been observed in the Milky Way. Their distances could be estimated by methods that work in nearby space. Thus Lundmark could derive their average luminosity which he then used to infer the distance of the Andromeda novae, about 200 000 pc. This placed Andromeda far beyond the borders of the Milky Way: it really is another Island Universe, or as we have learned to say, a galaxy. Allan Sandage summarized this history: "What are galaxies? No one knew before 1900. Very few people knew in 1920. All astronomers knew after 1924." Among those few was Knut Lundmark. So what happened in 1924?

7.3 Pulsating stars light up the way to Andromeda

The Small Magellanic cloud is a patch of light in the southern celestial hemisphere in the constellation of Tucana. This faint nebula became a stepping-stone to the heavens, when in 1912 Henrietta S. Leavitt published her remarkable work on its Cepheid stars. Cepheids were known in the Milky Way as regularly variable stars with variation periods ranging from one day to several weeks. Leavitt found that there is a tight relation be-

Henrietta Leavitt 1868-1921

tween the length of the period and the average luminosity of Cepheids. Cepheids are pulsating stars, and the bigger they are, the slower they pulsate. Everyone has seen one Cepheid: the Polar Star is such, palpitating with a period of four days.

Leavitt realized that the period–luminosity relation could be used for deriving the distance to any cepheid for which one measures the length of period and the flux. True, one must first calibrate the relation, i.e. to write it between period and *absolute* luminosity. The calibration can be made using the Cepheids in the Milky Way (though it is not easy – work on Cepheids' calibration still keeps astronomers busy).

Edwin Hubble 1889-1953

In 1917 the largest telescope at the time started working at Mount Wilson Observatory not far from Los Angeles. The size of its light gathering mirror, 2.5 meters, made it ideal for a study of faint nebulae, a task undertaken by Edwin Hubble. He joined the Observatory staff in 1919, after war experience on the Western front where he was wounded. Hubble had studied law and practiced boxing at Oxford, England, worked as an advocate in Kentucky, and defended his Thesis in astronomy at Chigaco university. Having the feeling that spiral nebulae are Island Universes, he tried to search in their photos for stars for making a distance estimate.

Only a big telescope can resolve stars from among the porridge of billions of suns which make up a spiral nebula. But one also must be sure that the faint light point is a star and know what kind of a star it is. If the light is variable, then it is most probably a star, of one variable type or other.

Looking originally for novae, in 1923 Hubble found in the Andromeda nebula a variable star which was a Cepheid. With this and nine other Cepheids, he calculated that the distance of the nebula is about 285000 parsecs. This confirmed beyond doubt the large distance that Lundmark had inferred from novae. Hubble's observations of Cepheids were officially reported on the first day of 1925 in a scientific meeting in Washington.

7.4 The diversity of galactic geometries

Penetration into extragalactic space opened for Man's eyes a "Zoo" unlike anything observed before. Yet, a small number of galaxy species, hopefully reflecting the physical processes in this new realm, were apparent to experienced viewers of the sky photos. Indeed, recognition of categories of objects has been an aim of natural sciences since the time of Aristotle.

Fig. 7.2 Inhabitants of the galaxy universe in the Virgo cluster. Hubble wrote in his *The Realm of the Nebulae* that the history of astronomy is a history of receding horizons. Knowledge has spread in successive waves, when astronomers have measured the distances to the Moon, to the stars, and to the galaxies.

Consider closely similar galaxies widely separated in space. They belong to one class. Their similarity has an important message: they have been formed in a similar way under similar conditions. If galaxies may be divided into a small number of classes, one has reason to assume that galaxies are produced by a few distinct processes, occuring everywhere in the universe.

In 1926 Hubble proposed a simple sequence of galaxy classes, which has, almost unchanged, stood the test of the years:

- *elliptical galaxies*
- *spiral galaxies*
- *irregular galaxies*

Elliptical galaxies consist of old stars and contain little gas and dust. Spiral galaxies contain both old and young stars, the latter forming the conspicuous spiral arms. There is much gas and dust. Irregular galaxies are clad with much gas and dust and produce many new stars.

One may think of a visible galaxy as put together from one or more geometrical forms defined by billions of stars in complicated orbits around its center: a sphere or an ellipsoid, a disk, and a system of spirals. In this view, elliptical galaxies are rather close in form to ideal ellipsoids. Spiral galaxies are composed of a disk embedded with spiral arms and an ellipsoidal bulge. The so-called lenticular galaxies are made of a disk without spirals. Some irregular galaxies have signs of underlying regularity. But galaxies have also their invisible side, massive dark haloes about which there will be more in Chapter 10. We do not know their forms, but they might be spheres or ellipsoids.

The same year as Hubble, Lundmark published a system of galaxy classes. In some respects the classifications were rather similar, in others they differed. The similarities made Hubble think that his scheme had been copied, and he publicly attacked the Swedish astronomer. Hubble's irritation tells much about the hot race into the new realm of galaxies.

We are happy to say that sixty years later, one of us (P.T.) could show that Lundmark had worked already a few years on his nebula classes, independently of Hubble. Knut Lundmark left in his will his vast collection of old astronomical literature and papers to Tuorla Observatory. Its founder, Yrjö Väisälä was a close friend. Among these papers there is a hand-written nebula catalogue, prepared by Lundmark in 1922 when he was visiting the Lick Observatory in California. The roots of his classification system are found in this document. *

Yrjö Väisälä 1889-1971

7.5 Our home galaxy – the Milky Way

Spiral galaxies rotate, as may be guessed from their whirlpool-appearance. In 1927 the Dutch astronomer Jaan Oort showed that also our Milky Way is rotating around a point lying in the direction of the constellation of Sagittarius. The distance from our Sun to this center, obscured by a thick wall of dust, is about 9000 parsecs. Revolving with a speed of about 220 km/sec at this distance, we make one full turn in 200 million years. Rotation speeds and sizes tell us that the mass of an average galaxy is about a hundred billion solar masses (Appendix A.2). One cannot actually count how many stars there are in a galaxy. However, such a mass suggests that

Jaan Oort 1900-1992

---

*More details are in the article "Knut Lundmark's unpublished nebula classifications on Crossley plates in 1922", Journal for the History of Astronomy 20, 165 (1989) by P.T.

a good galaxy contains a hundred billion stars:

typical galaxy = 100 000 000 000 stars (+gas+dust+dark matter)

It has been established, though it is not easy from our position, within, that the rotating Milky Way is a spiral galaxy, as is our nearby neighbor Andromeda. The Sun lies in the inner edge of the Local Spiral Arm which contains such familiar stars as the blue Sirius of Canis Major, the red Betelgeuse of Orion, and the stars making the "W" of Cassiopeia.

7.6 Spectra – fingerprints of stellar matter

Joseph Fraunhofer 1787-1826

Newton saw that sunlight is dispersed into a spectrum of different colors when it passes through a prism. Joseph Fraunhofer, eleventh son of a poor family and a skilled maker of optical instruments, found that the spectrum contained, besides the colors, numerous narrow regions with no light at all. Each such *absorption line* corresponds to a certain wavelength – apparently the light with such wavelengths has been absorbed by something.

Robert Bunsen 1811-1899

Gustav Kirchoff 1824-1887

Around 1860 Robert Bunsen and Gustav Kirchoff realized that the Sun's spectral lines are fingerprints of familiar chemical elements. If one vaporizes an element in a hot flame, it begins to radiate and with a spectroscope one may see lines characteristic to this element and always having the same wavelengths. Now it became possible to get information on the chemical composition of stars. The same agent that allows us to see the stars, carries with it subtle information which eludes the naked eye. Astrophysics began.

Only quantum physics disclosed how the spectral lines are formed due to energy levels in atoms. This explanation of the spectra, from the candle flame to the distant star, belongs to the success stories of 20th century physics, and gave a firm theoretical basis for astrophysics.

Among the thousands of hot globes visible by naked eye and the billions reached by telescope, some are very massive and luminous and some are dwarfs in comparison with the Sun. Some have a blue color (indicating a high surface temperature), others are red (cooler). † Stars are different, but their properties are not haphazard: a regularity appears when one inspects luminosities, masses, and spectra for large numbers of stars. This

---

† We do not burden the reader with details on how to infer the properties of stars. Measurements of flux and colors, investigation of spectra, and other studies of stars give information on their luminosity, temperature, mass, size, and chemical composition.

pattern is strikingly seen in the *Hertzsprung-Russell diagram*, which is a most important tool in the hands of astronomers.

Ejnar Hertzsprung 1873-1967

Henry Russell 1877-1957

One way to build the HR diagram is to use temperatures and luminosities. One at once sees that most stars lie on a narrow band from cool and faint stars to hot and luminous ones. This strip is called the *Main Sequence.* Two other regions also accommodate stars: red giants above the Main Sequence and white dwarfs below it. To understand why the HR diagram has such a look as it has, has required a lot of observations and theoretical studies of the structure, energy source, and evolution of stars.

7.7 Spectral line shift – a celestial message

Christian Doppler 1803-1853

Christian Doppler from Austria discovered in 1842 that when a body emitting sound moves away, the sound is heard with a lower pitch. That is, the sound waves become longer. He suggested that this effect appears in any wave motion, including light. In case of light the shift caused in the wavelength by the *Doppler effect* is simply the ratio between the velocity of the light source and the speed of light:

Doppler shift = velocity divided by speed of light

The Doppler effect became a rich source of cosmical information, when astronomers began to apply it to measure motions of stars. Similarly as the spectrum of the Sun, the light from a star is marked by absorption lines. Even small shifts in their wavelengths can be measured and the motion of the star inferred with the aid of Doppler's formula. [‡]

William Huggins 1824-1910

Sir William Huggins, a wealthy British amateur astronomer in the still romantic "age of science", was the first to use the Doppler effect in astronomy: he constructed his own spectroscope, attached it to his telescope, and in 1868 he pointed it at Sirius, the brightest star in the sky. The line shifts indicated that Sirius is moving away from us with the velocity of 50 km/sec. But how do we know that the line shifts in the spectra of the stars are due to the Doppler effect? Are the stars really moving?

"Fixed" stars are not nailed on the firmament nor in any absolute space. Lengthy observations reveal that nearby stars slowly but surely shift in the

---

[‡] Accurately speaking, the simple Doppler formula contains only the *radial* velocity V, i.e. the component of velocity in the direction of the line-of-sight from our eye to the star. Also, the formula is valid only for speeds much smaller than the velocity of light.

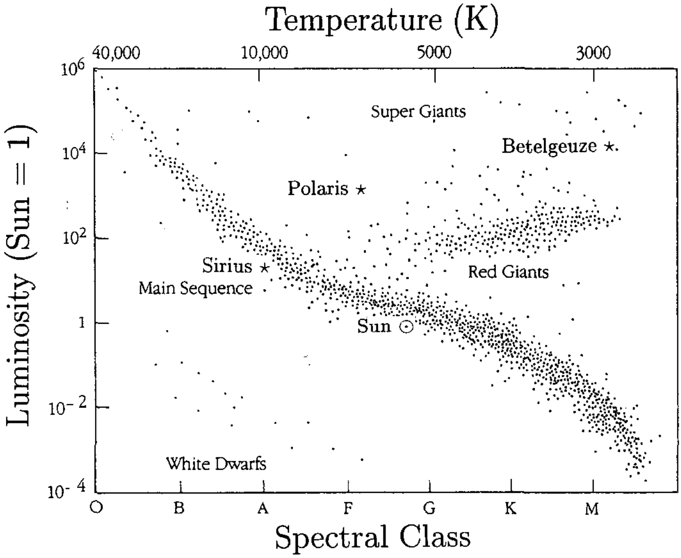


Fig. 7.3 The Hertzsprung-Russell diagram for stars. The horizontal axis gives the surface temperature and the vertical axis the luminosity in terms of the radiation power of the Sun, a Main Sequence star. Also some other candles of the sky are indicated.

sky relative to other, more distant stars. Motion is clearly seen in double stars, where two stars rotate around each other. The radial velocities of the stars, inferred from the Doppler effect, decrease and increase as expected from the orbital motions.

7.8 Discovery of extragalactic redshifts

In 1912 Vesto Slipher measured the first spectral line shift of a spiral nebula. It was a blueshift (a shift towards shorter, or bluer, wavelengths), and indicated a surprisingly high approach velocity. As often happens in science, the discovery of the high speed of 300 km/sec for the Andromeda nebula was a by-product of a plan to detect another thing, the rotation of the nebula. This was the task given to Slipher by his boss in the Flagstaff Observatory, Percival Lowell, a planetary astronomer who was interested in the theory that spiral nebulae are a stage in the formation of planetary systems. It was not yet known that they are huge stellar systems.

Vesto Slipher 1875-1969

Percival Lowell 1855-1916

The speed of 300 km/sec was quite high in comparison with the velocities normally measured for the stars. Now we know that the blueshift of the Andromeda nebula reflects largely our own motion: the rotation of the Milky Way carries the Sun with a speed of about 220 km/sec.

By 1917 Slipher had managed to measure a total of 25 radial velocities for the faint nebulae each needing exposures of tens of hours (in comparison with ten minutes or even less with modern telescopes). The line shifts were mostly redshifts (towards longer wavelengths), corresponding to large speeds of up to 1100 km/sec. Another aspect appeared in 1919 when Harlow Shapley noted that fainter nebulae tend to have larger redshifts: "...indicating a relation of speed to distance or, possibly, to mass."

7.9 The search for a relation between redshift and distance

Knut Lundmark came close to discovering the Hubble law before Hubble. In 1924 he presented a diagram in which a distance–redshift relation was discernible. However, instead of being happy with a linear relation he added in his formula also a negative quadratic term, so that at small distances the redshift first increases, then at larger distances the extra term makes the redshift curve down. Lundmark held in his hands the chance to discover the new law, had he simply thrown away the miserable quadratic term.

Lundmark used the diameter of a galaxy as an indicator of its distance: smaller angle means larger distance. But the angular diameter, without any other aid, is a quite poor distance indicator. That is why he could not see clearly any linear relation between redshift and distance.

In his epoch making study of 1929, Hubble used as the distance indicator the brightest stars in galaxies. Though generally these "stars" were gas clouds excited to emission by young giant stars, nevertheless their brightnesses provided better *relative* distances than diameters of galaxies. This was seen as a decreased scatter in the distance–redshift plot, and one could better discern any law linking redshift and distance. Hubble concluded that a linear law, redshift directly proportional to distance, described what he saw on the diagram.

Fig.7.4 shows a modern Hubble diagram, in which redshift and distance grow logarithmically. It contains three classes of distance indicators which cover the distance range 1 Mpc to 500 Mpc (we discuss distance measurements in Chapter 15). The slope of the straight line (= 5) is what one expects if the linear Hubble law is valid (see Appendix A.3).

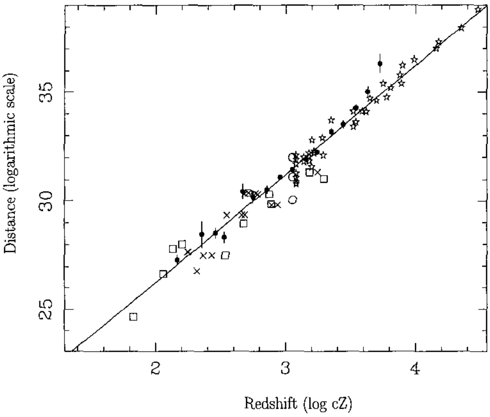


Fig. 7.4 The Hubble law (distance versus redshift) as indicated by galaxies whose distance has been measured by different methods (Cepheid variable stars, the Tully-Fisher method of rotating galaxies, and supernovae of type SNIa).

7.10 The law of redshifts: a new cosmic phenomenon

The "happy twenties" were an incredible decade for science, when the microcosm was shown to be ruled by the strange laws of quantum mechanics and the galaxy universe was discovered. And in the macrocosm a new physical phenomenon, cosmological redshift, was immediately found when one entered the previously inaccessible realm of galaxies. A remarkable feature of the redshift is that its value does not jump randomly from one galaxy to another, but is larger for more remote galaxies:

$$REDSHIFT = constant \times DISTANCE$$

A few nearby galaxies do have blueshifted spectra, but these are rare exceptions. All other galaxies all over the sky have redshifts in their spectra. In nearby space, the redshift is small. So the light from the Virgo cluster, the closest large galaxy cluster, is redshifted by an amount of one part in three thousand, or its redshift is 0.003. This means that the wavelengths of all the spectral lines are 1.003 times the wavelengths of the same lines in the laboratory. The highest observed redshifts reach values close to 7. Such large redshifts have a dramatic effect on the spectra. They transform the emitted ultraviolet light (which would be stopped by our atmosphere's ozone layer) into visible rays, allowing us to see spectral lines from the ground which we could otherwise only observe with a space telescope.

It is good to keep separate the observed Hubble law and any interpretation of it. When Hubble found his law, he was not sure about its implications. A natural first assumption is that redshift is caused by the Doppler effect, when a galaxy moves away from us. If so, then redshift is simply this recession velocity divided by the speed of light.

Hubble and Humason were careful to write in their article in 1931 that "the term 'velocity' will be used for the present in the sense of 'apparent' velocity, without prejudice as to its ultimate significance". This practice of expressing the redshift as an apparent velocity is still used today, especially when astronomers discuss relatively nearby space. For example, the Virgo cluster has a redshift of 1000 km/sec. Often astronomers use the apparent velocity in the Hubble law:

$$\textit{recession velocity} = \textit{Hubble constant} \times \textit{distance}$$

Here appears the celebrated and debated Hubble constant. [§]

Already in the 1920's theoretical anticipations of the Hubble law were in the air, based on solutions of Einstein's equations, and it did not take long before the law of redshifts was seen in the light of the expanding universe model. The expansion of the universe is the explanation given by modern cosmology and accepted by the overwhelming majority of astronomers.

But did Hubble discover straightaway the expansion? One may read "in 1929, Edwin Hubble made the landmark observation that wherever you look, distant galaxies are moving rapidly away from us". But things were not that simple. Such assertions should be taken with a pinch of salt, because Hubble never could see galaxies moving and he was open to different interpretations of his law. What astronomers see, is something more subtle. They look at the spectra of galaxies and see the familiar spectral lines at different positions on the spectrogram, i.e. at different wavelengths. As compared with the spectral line emitted by the same atom on the Earth, usually these lines are at "wrong" places, shifted towards longer wavelengths, towards the red.

Remarkably, the redshift has the same value no matter for which spectral line the shift is measured, i.e. the cause of the redshift influences all

---

[§] The Hubble law is: $Z = const. \times R$ and the apparent velocity is related to the redshift as $V_{app} = Z \times c$, combining both formulae results in $V_{app} = H \times R$ (H is the Hubble constant), which is often also called the Hubble law.

photons in the same way. Modern physics knows two experimentally verified mechanisms which produce redshifts that are independent of wavelength.

* the *Doppler effect* is caused by the mutual velocity of a light source and the observer moving in space. For example, the wavelength of a spectral line emitted from the Sun differs across the solar disk. The line shift varies between the redshift of +20 km/sec and the blueshift of −20 km/sec, exactly as expected from the solar rotation revealed by the sunspots.

* the *Gravitational redshift* appears when light is emitted by an atom sitting closer to a gravitating mass than the observer. Roughly speaking, the redshift is due to the energy loss suffered by photons which work their way away from the gravitating body. ¶ A nice example is given by the spectral lines emitted from the solar disc. They have an extra redshift equal to 10^{-6}. This apparent velocity of +300 m/sec is a measure of the solar mass and size: the Sun is not rushing away from us at this speed!

Besides these empirical and well understood redshift mechanisms, there is the theoretical phenomenon accompanying the expansion of space.

* the *Expansion of space*, causing a stretching of a photon's wavelength. This widely accepted explanation for cosmological redshift is a consequence of general relativity. Its is not the same as the Doppler effect (Ch.12).

Thus three redshift mechanisms act in Nature. Observed redshifts may contain contributions from each of them. ‖ At large redshifts expansion redshift is thought to dominate over the Doppler shifts arising from the motion of our Milky Way and from galaxy streams. The gravitational redshift is normally regarded as too tiny to detect in the light of galaxies.

7.11 Galaxies live in swarms

Clustering or non-uniformity is an important attribute of the realm of galaxies. It strikes the eye already from the old map showing the distribution of bright spiral nebulae in the sky (Fig.7.1). Indeed, the best place to find a galaxy is close to another one. Walter Baade, an eminent astronomer,

---

¶ In fact, the interpretation of the gravitational redshift depends on the theory. In general relativity the redshift is the result of curved space-time and it may be also inferred from the equivalence principle. In Feynman's field gravity it is caused by the shift of the atom's energy levels when the atom and the external gravity field interact.

‖ The observed redshift is not a simple sum of the different redshifts, but is obtained from the product of the $(1 + Z)$ terms: $(1 + Z_{obs}) = (1 + Z_{exp})(1 + Z_{Dop})(1 + Z_{grav})$

Walter Baade 1893-1960

recalled in his lectures how "Hubble and I had a long-standing bet of 20 dollars for the one who could first convince the other that a [galaxy] system which he had found was single. We never could decide the bet. ... So single galaxies may be rare." Our Milky Way faithfully follows this rule: it is a member of a group of galaxies, which Hubble named the *Local Galaxy Group.* Another major member is the Andromeda galaxy, visible even to the naked eye. In addition there are about twenty smaller member galaxies. The closest large galaxy cluster is in the direction of the constellation of Virgo. The *Virgo cluster* contains hundreds of galaxies, and is a center of the much larger *Local Supercluster.* Our Local Group lies in the outskirts of this system, about 20 million parsecs away from the central Virgo cluster.

7.12 Super energies in the galaxy universe

Galaxies were discovered with the largest telescopes at the time. Since then astronomical instruments have much advanced both in size and technique. While a century ago the largest tube had the light collecting diameter of 1.5 m, presently tens of telescopes are larger than 4 meters. Sensitive electronic detectors have replaced photographic plates. But the astronomer not only strives for many photons – they should also draw sharp images. The ever boiling atmosphere erratically refracts the light rays and spreads the points of stars into dots with sizes of a second of arc or more. For instance, distant galaxies cannot be well studied through the air. That is why the Hubble Space Telescope, orbiting the Earth at a height of 600 km, was such a big step. Beyond the air it can make sharper images than usual telescopes.

Going to space is not only for hunting up better pictures. The atmosphere is transparent only for those photons with wavelengths in the narrow optical and radio "windows". Opening new wavelength windows lets in unexpected messages, as photons of different energies originate in different cosmic phenomena. Through the optical window one sees mostly starlight (galaxies). Many high energy events are detected in the radio band. Infrared photons are mostly thermal radiation from the cosmic dust, and in microwaves the cosmic background radiation reaches us. The mysterious gamma-ray bursts, which occur in distant galaxies, are the most powerful explosions in the universe.

Karl Jansky 1905-1950

In 1933 the radio engineer Karl Jansky, who worked for the Bell telephone company, investigated the reason for disturbances in radio telephone traffic across the Atlantic Ocean. The pioneer of radio astronomy found that unknown radio sources followed the revolution of the sky.

Simple radio telescopes are handicapped by poor resolution: a point-like radio source is seen in a radio picture as a fuzzy blob, covering in the sky not one arc second, but one arc minute. ** Then a remarkable thing happened in 1951. Graham Smith, working at Cambridge University, had localized a bright radio source in the constellation of Cygnus with the mentioned arc minute accuracy. He asked Walter Baade at the Palomar observatory, to take a photo of that direction with the new 5 meter telescope. Baade was astonished to see in the photo what looked like a collision between two galaxies. And surprisingly, the redshift was so large (0.06) that the distance must be about 300 Mpc (500 times farther than Andromeda). In spite of its large distance "Cygnus A" is the second strongest radio source in the sky (the strongest source – after the Sun – is Cassiopeia A, a remnant of an old supernova explosion). *Radio galaxies* had been found.

Victor Ambartsumian 1908-1996

In the late 1950's, Victor Ambartsumian came up with an exciting explanation for such violent phenomena. The Armenian astronomer said that energy is generated in *active galactic nuclei.* And indeed, astronomers have discovered the outflow of energy in narrow channels from the nuclei.

Cygnus A is a double radio source. In radio images such objects give the vivid impression that two radiating clouds have been ejected in opposite directions from one point. Between the clouds one often finds a giant elliptical galaxy, the site of the ejection. Narrow jets start from the nucleus of the mother galaxy, cosmic umbilical cords which feed by high energy particles the shining twins. A double radio source may be 1 million parsecs in size, tens of times larger than the galaxy giving rise to it. Without the radio window, we would be unaware of such great phenomena!

Often in the center of the double source there is no galaxy, but a star-like thing that is nevertheless not a star. These are *quasars*, quasi-stellar objects, first discovered as "radio stars" in 1960 by Allan Sandage and Thomas Matthews. That quasars are extragalactic was betrayed by their spectra. Maarten Schmidt studied a "star" which coincided with a pointlike radio source called 3C273, and found in its spectrum strong emission lines – sim-

---

**Modern large radio telescopes have many movable antennas, like the American VLA (Very Large Array) system and the Russian 600-m RATAN telescope. One can also join several remote radio telescopes into one system and reach a very high angular resolution (*V*ery *L*ong *B*aseline *I*nterferometer). VLBI can study structures which have angular sizes of one thousandth of an arc second.

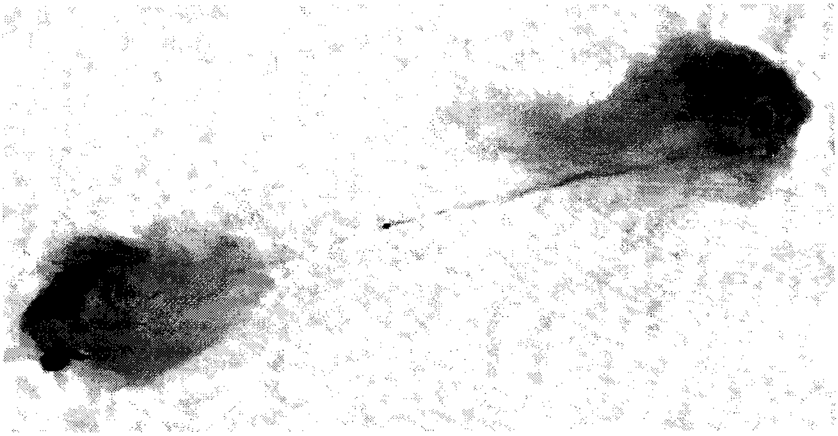


Fig. 7.5 The classical double radio source Cygnus A in the direction of the constellation Cygnus. The thin jets connect the active nucleus of the host galaxy with the hotspots in the outer extended radio components (VLA map courtesy of R.A. Perley).

ilar strange lines had been earlier found by Sandage in the star coincident with another radio source 3C48. The mystery of 3C273's emission lines was resolved when Schmidt realized that they are the familiar fingerprints of hydrogen, though shifted by 16 percent towards the red. The Hubble law says that 3C273 is far beyond the Milky Way, and not a star.

What was and still is puzzling in these compact bodies, is their huge luminosity: they can emit hundreds of times more light than a whole galaxy! The energy must come from a very small volume, because in a few days the brightness of a quasar may double. This rapidity requires that the quasar be small, about the size of the Solar System. Observations show that quasars reside in the centers of galaxies, though their dazzling brightness almost masks the light of the host galaxies. With their sizes one millionth of the host galaxy, these one billion solar mass objects fully deserve the name of active galactic nuclei.

The latest quasar catalog contains over 10 000 quasars. So the quasars, all too faint to be seen by naked eye, cover the sky more densely than the visible stars. The total number of quasars exceeds millions. This is still nothing compared with the 100 billion faint galaxies of which there are about one million per square degree in the sky.

7.13 Anomalous redshifts – the exception to the rule?

In 1966 Halton Arp, at the Palomar Observatory, inspected the photos of galaxies in his well known *Atlas of Peculiar Galaxies.* He happened to notice that radio sources, among them quasars, tended to be close to, or aligned across, some of the galaxies of the catalog. Could high-redshifted quasars be associated with galaxies having much smaller redshift? If true, these would be flagrant violations of the Hubble law: the quasars would have a large extra component in their redshift, in addition to the cosmological one. This possible extra line shift became to be called an *anomalous redshift.*

The same year, Fred Hoyle and Geoffrey Burbidge suggested that quasars may be objects ejected from rather nearby galaxies. The aim was to make obsolete the huge energies required: if the monsters were less distant than indicated by the redshifts, their power also would be diminished.

Since 1966 Arp has found many interesting cases where high redshift quasars lie in the sky very close to a low redshift galaxy. It must be emphasized that for quasars there are no easy way to tell their distances independently of the redshift. To say for sure that a quasar is at the same distance as a galaxy, one should see a physical connection between them. Arp has pointed out many galaxy–quasar pairs where such a link seems to exist in the form of a luminous "bridge" or "tail".

In 1996 Burbidge calculated the probabilities. The number of known cases when a quasar is close (less than 3 arc minutes) to a bright galaxy was 46. Is that much expected from chance alone, or does the number of pairs suggest some kind of connection? One can calculate the expected number of chance associations, if one knows how many galaxies and how many quasars there are in the sky. So, draw a circle with a radius of 3 arcmin around every inspected galaxy. Then throw randomly, eyes closed, all the "quasars" on the celestial vault. Counting the scores, how many of the quasars would be found inside some 3 arcmin circle? The answer is 1!

Burbidge concluded (as he did in a recent review in 2001) that in general quasars have an intrinsic redshift component plus a cosmological one. In order to explain the origin of the anomalous redshift, Arp & Narlikar and Hoyle & Burbidge have presented theories based on the idea that quasars are ejected from the nuclei of galaxies and are composed of new, freshly created matter. This would require new, non-standard physics which is difficult to test and thus far has not attracted much attention. We come back to this subject in Chapters 10 and 16.

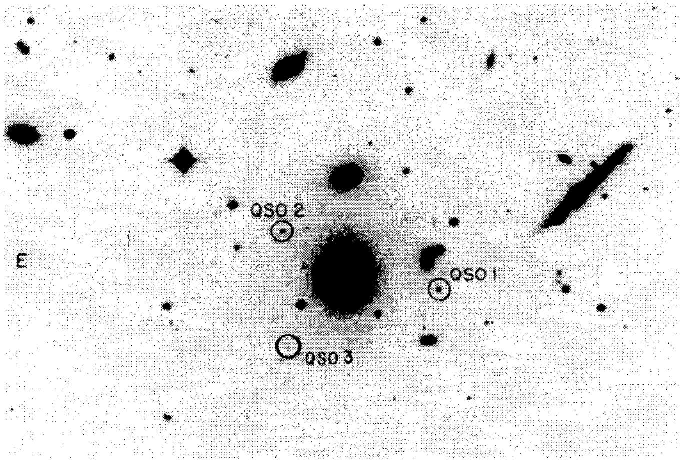


Fig. 7.6 An example of Arp's galaxy–quasar association. Three quasars (circled on the photo) are found within 1 arc minute from the center of the elliptical galaxy NGC3842 which has the redshift $Z = 0.02$ corresponding to the recession velocity of 6000 km/sec. The redshifts of the quasars 1, 2, and 3 are 0.34, 0.95, and 2.21, respectively.

7.14 Redshift quantization?

Another odd redshift effect has haunted the sky: periodic redshifts. In 1976 William Tifft studied binary galaxies. He found the totally unexpected thing that the redshift difference of the two member galaxies occurs preferentially in steps of 72 km/sec (when redshift is expressed as velocity). Similar "quantization" has been reported in the redshifts of galaxies in clusters. So Bruce Guthrie and William Napier analyzed the Local Supercluster using a hundred galaxies with accurately measured redshifts. After they subtracted from the redshift the rotation of our Milky Way, the redshifts of galaxies had a periodicity of 36 km/s, one half of Tifft's original period.

If Tifft's effect is real, it is hard to understand. Leaving aside the unknown origin of the quantization, there is the problem of "Doppler smoothing": real motions of galaxies should sweep away any redshift peaks. It seems that galaxies are "nailed" in their places in such a strange universe and the motions of double galaxies around each other are quite sluggish.

An echo of this exotic idea came from the North. Ari Lehto, a Finnish physicist, had in 1990 come up with a general formula for quantization of physical quantities, without knowing of Tifft's astronomy studies. He was then amazed to find that his formula describes well the redshift periods that Tifft has observed. If the energy of any photon is quantized, as Lehto suggests, then also the frequencies and, hence, the redshifts of light would occur at preferred values. [††] A fascinating idea, but still in the bud.

Astronomers are quite sceptical about the phenomena discussed by Arp and Tifft. For instance, redshift periods has been so far studied by a small number of "lonely wolves" only and it needs elaborate analysis of redshift data. And its incompatibility with our ordinary view of cosmology rather tends to drive away than attract. But one may also say that it is a part of cosmology to question and test our view of the redshift, one of the cosmic key observations. Then it is natural to turn one's attention to all its properties, including subtle phenomena which, if real, would easily elude detection without dedicated studies. A close look at exceptions may shed new light on the physics behind the law.

In Chapter 10 we speculate that Arp's effect possibly tells about dark mass in haloes of galaxies rather than about redshift itself. If so, this does not diminish the importance of the effect, but shows how rare phenomena, ignored by almost everybody, may be decisive keys in science.

Toivo Jaakkola 1940-1995

Toivo Jaakkola, a Finnish astronomer, was one of the pioneers in this kind of a phenomenological study ('phenomenological': such exploration of observations, which is inspired by general theoretical ideas, without an exact theory at hand.) He tried to detect the influence of matter on the redshift in different environments. So he used spiral galaxies to see whether the light emitted by the more distant edge is more redshifted than the light from the closer edge. This might be caused by some kind of interaction when photons travel a long way close to the massive and luminous galaxy.

Jaakkola concluded that there is evidence for such variations of redshift. However, these small effects, if they exist, are hard to detect from observations. Toivo Jaakkola's untimely death interrupted his new analysis of the redshifts of galaxies' edges. In a letter to one of the authors, a few days before he passed away, he was confident that the results will be convinc-

---

[††] A. Lehto had searched for a common rule for the properties of the "small" and the "large". He examined how space and time occur in physical quantities appearing in Nature, and found that the ratios of the quantities involving lengths or energies may be simply expressed as $2^{n/3}$ ($n = 1, 2, 3, \ldots$). That the "natural" ratios seem to be built on the basis of the number 2, he interprets as a phenomenon analogous to that called period doubling in chaotic systems, which brings the subject into contact with fractality. The theory of scale-relativity created by L. Nottale on the basis of fractal space-time (Ch.15) is also said to lead to quantized redshifts.

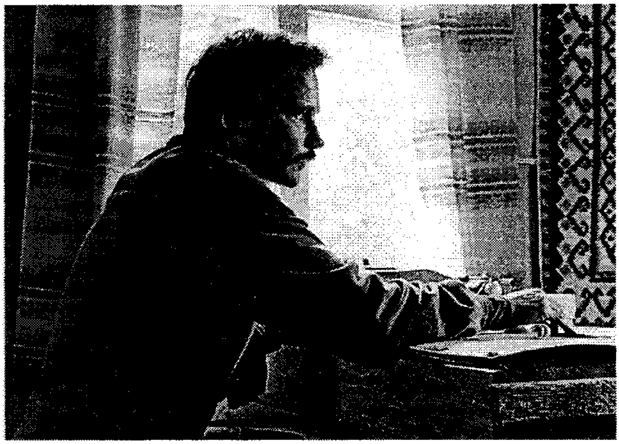


Fig. 7.7 Toivo Jaakkola in his house close to the Polar Circle. His cosmological views led him to question the standard interpretation of the cosmological redshift and to study its properties on different scales. For this original thinker cosmology was something like the air we breath, without which we cannot live, even if few of us wonder about it.

ing. This modest and humane man was an ardent treasure-seeker all his life. He believed he had seen a glimpse of a Rosetta stone buried deep in astronomical observations.

Original explorers like Arp, Jaakkola, and Tifft have an attitude which leaves no stone unturned, suspecting that the available observational data, perhaps already covered by dust in archives, contain information that waits to be seen from a fresh point of view. Indeed, at all times there have been observations of phenomena, which have not been explained by the contemporary theory of heavens, but which are natural in the next cosmology. Just recall the loops of dancing planets.

* * *

The vast kingdom of galaxies is inhabited by giants and dwarfs – a diversity of stellar systems obliged to follow the law of redshift. The nuclei of galaxies are volcanos of extremely violent ejections of matter, streaming through narrow channels nearly at the speed of light. The galaxies do not like loneliness, they gather together, forming rich structures from small groups to superclusters. The prediction that somewhere beyond the misty horizons, on very large scales, the local mountains are mere grains on a vast plain, led to a cosmological model which proved to be a great success.

Chapter 8

The triumph of uniformity in cosmology

In its ardent exploration of Nature at small and large, the 20th century gave birth to a world model in which the primordial explosion determines the history of the universe. The model rests on Einstein's cosmological principle, the idea of a uniformly dispersed matter which evolves following the laws of general relativity. The success of big bang cosmology in explaining important observations and its impressive, almost shocking world view, appeal to both reason and soul of modern man. Said Stephen Hawking: *The discovery that the universe is expanding was one of the great intellectual revolutions of the twentieth century.*

Big bang cosmology is the modern "Great Synthesis" which has united astronomy and physics. Now the work was not by one author in one book, but by generations of scientists in thousands of articles. As a paradigm it both rules and inspires science. And the ideas of big bang and evolving universe have penetrated society in general. It seems almost impossible to think about reality without the enigmatic beginning.

A pivotal aspect of big bang cosmology is the conviction that the huge cosmic structures we see around are "secondary effects": results of the evolution from tiny primordial ripples on the underlying calm uniformity.

* * *

8.1 Friedmann's discovery of expanding universes

It happened that the University of St. Petersburg, built by Peter the Great on the beautiful embankment of the river Neva, played an important role

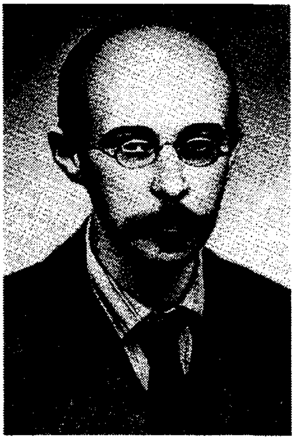


Fig. 8.1 Alexander Friedmann. He was a mathematician, specialized in meteorology, and became a founder of modern cosmology.

in the creation of big bang cosmology. Alexander Friedmann was a professor of this seat of learning, when he found the famous equations describing the expanding universe. George Gamow, who added the hot start to the expansion, was Friedmann's student. Gamow learned from him general relativity. At that time Leningrad was the Alma mater for a whole generation of creative physicists, we mention Lev Landau, Vladimir Fock, and Victor Ambartsumian. After Friedmann and Gamow, Yakov Zeldovich, who also had his scientific roots in "The City of White Nights", formed the famous Moscow school of cosmology. No wonder it is joked that the best Muscovites are Leningraders, and not only in cosmology...

Alexander Friedmann 1888-1925

A few years earlier, in 1917, Albert Einstein had applied general relativity, soon after he had invented it, to the cosmological problem. If the matter distribution is uniform, the curvature of space is the same everywhere. Furthermore, Einstein required that the cosmos should not change, or that its curvature remains constant at all times. He showed that this leads to spherical space, as if the universe had curved around on itself.

What made Einstein turn to the large scales of cosmology? One starting point for the problem of gravity is to put a mass into an empty space. This mass deforms the space around it, which possible "planets" feel as gravity. At very large distances from the mass the space is less and less deformed, and there a "planet" or a test particle will feel practically no gravity any more. However, it will still have its mass and will obey Newton's law of motion (i.e. it has inertia). From where does this inertia come? Einstein wanted to include Mach's principle into his gravity theory, so that inertia is caused by an interaction between the particle and all the mass in the rest of the universe. But if the world is just the Milky Way surrounded by an infinite emptiness, then why should there be inertia at very large distances? To cut a long story short, Einstein thought that he could avoid this unsatisfying situation by the striking idea of getting rid of infinity and letting the universe be a finite unbounded structure, a spherical space. Already in 1900 Karl Schwarzschild had speculated on the spherical geometry of the world, but Einstein could put this idea on a firmer basis with the aid of his equation connecting curvature and the mass content of the universe. Mach's principle inspired Einstein, but the origin of inertia is still in our days an enigma.

Because Einstein required that space does not change in time he used the equations of general relativity in a form more complex than originally. He had to add there the so-called lambda-term, or *cosmological constant*, which balances the otherwise collapsing universe.

An important breakthrough came in 1922. Alexander Friedmann found another cosmological solution in the article "On the curvature of space". Einstein's equation does allow changing, non-static cosmologies with uniform matter distribution. In these models space is expanding, and the curvature and the matter density change with time. For such universes, there exists a moment of creation a large, but finite time ago.

The first reaction of Einstein to Friedmann's article was not very encouraging. He published a short critical comment of five sentences, where he claimed that Friedmann actually had proved that his static model is the only possible one! Later, after having received from Friedmann a private letter, Einstein admitted in four sentences that his criticisms was unfounded – he had himself made a small error of calculation and now regarded "Mr. Friedmann's results as correct and clarifying".

Friedmann's models do not need any lambda. Later Arthur Eddington showed that Einstein's universe would be unstable even with a lambda-term. Its attempt to balance in a static state would soon fail, as had happened with Newton's standing needles. When Einstein, in view of the Hubble law, accepted that expanding models describe the world better than his static model, he sadly said that the lambda-term "was the greatest blunder in my life"(as Gamow recalled in his book *My World Line*). Ironically, today Einstein's cosmological constant is again in the limelight: certain observations suggest that it exists after all. Now it is needed in order to explain new observations and to remedy certain problems encountered by big bang cosmology.

Alexander Friedmann died just after the galaxy universe was discovered, but before the Hubble law was found. He never came to know the importance of his studies for modern cosmology. We think of it with a kind of nostalgia that parts of this book were written in the small town of Pavlovsk (close to St. Petersburg), where Friedmann headed the Geophysical Observatory. A few weeks before his untimely death he had made, for scientific purposes, a flight with an aerostat to the record height of 7400 meters.

8.2 Cosmological redshift in expanding space

The scale of a regular map is constant, but to describe an expanding universe one needs a changing scale. That is why the Friedmann models contain an important quantity called the *scale factor*. It tells how the distance between any two galaxies increases when the cosmic time flows from the birth of the expanding universe. In an earlier cosmic phase, the scale was smaller than now and the galaxies were closer to each other.

Expanding space has an important effect on light. It is redshifted due to the stretching of light's wavelength. An exact calculation shows that the wavelength changes as much as the scale factor between the moments of emission and detection: when the scale of the universe doubles, also the wavelength doubles. This leads to a very simple relation between the observed redshift and the ratio of the present scale factor to the past scale factor. So the redshift $Z = 1$ means that the light was emitted when the universe was half of the present size. * In this fashion, cosmologists use the redshift to denote different past epochs of the universe.

Redshift is also a kind of measure for distance, and, in our opinion, modern man should have a feeling of what redshift means in a manner roughly similarly to hearing in the news about the "Richter scale". In terms of redshift, one may divide the observable universe into three "layers":

---

*G. Lemaître showed that there is a simple connection between the scale factor $S = S(t)$ and the observed redshift Z: $S(now)/S(emit) = Z + 1$. Here $S(now)$ is the present value of the scale factor, $S(emit)$ is the value it had when the light was emitted.

- The *local universe*: where $Z < 0.1$
- The *intermediate universe*: where $0.1 < Z < 1$
- The *deep universe*: where $Z > 1$

This division is somewhat arbitrary, but convenient. Speaking about very large distances, astronomers usually have in mind redshifts beyond one. The distance corresponding to the redshift $Z = 1$ characterizes the observable part of the universe, about 5000 Mpc. Recall that the distance to the nearest galaxy, Andromeda, is 0.7 Mpc. The ratio of these sizes is large, but not huge – like the Earth to a village!

8.3 Uniformity gives rise to the Hubble law

After the realm of galaxies had been discovered, it soon became common to think that Man had finally entered a representative part of the universe. For instance, the Hubble law starts not far from the Milky Way, at only two times the distance of the Andromeda galaxy. And Hubble's galaxy counts in the 1930's were welcomed as suggesting a uniform distribution of galaxies, something that Newton and Einstein had intuitively expected and that became known as the Cosmological Principle. There are about equal numbers of faint galaxies in all directions (outside the dusty Zone of Avoidance). This isotropy was also taken to imply uniformity. The Hubble law and the uniformity were taken as properties of the universe as a whole.

Howard Robertson 1903-1961

Goeffrey Walker 1909-2001

Then it was realized in the 1930's by Howard Robertson and Goeffrey Walker that there is an intimate connection between the uniformity and the velocity of expansion. General geometrical reasoning shows that if a uniform space expands so that it remains uniform, then the velocity – distance relation can be seen everywhere: [†]

$$VELOCITY = constant \times DISTANCE$$

This theoretical relation is also often called the Hubble law, along with the original, observed redshift – distance law. The connection between the supposed uniformity of matter distribution and the Hubble law was the

---

[†] From Einstein's equations it follows that in expanding uniform space there is the exact velocity – distance relation $V_{exp} = H \times R$, while the redshift is a complicated function of the distance $Z = Z(R)$. Only for small redshifts is the redshift a linear function of the distance: $Z \approx (H/c)R$, $Z \ll 1$. From the last formula one also can derive an approximate value for the expansion velocity, when one knows the redshift: $V_{exp} \approx cZ$.

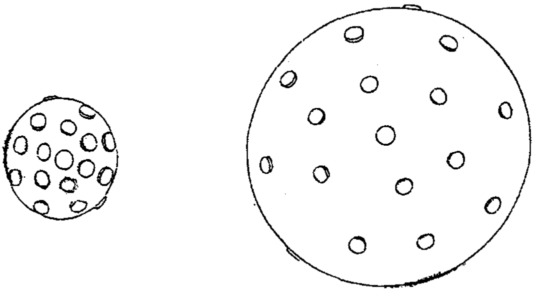


Fig. 8.2 Arthur Eddington's balloon analogy of expanding space. When the radius of the balloon grows with time, the distances between the coins glued on the balloon increase. This demonstrates how the Hubble law emerges in the expanding universe. The coins represent nonexpanding galaxies, hence space expands only between the coins.

first great success of the expanding world model. It necessarily leads to a beginning in the past, the big bang. In the form of one logical chain:

$$\textit{Uniformity} \Rightarrow \textit{Hubble Law} \Rightarrow \textit{Big Bang}$$

It is essential for the expanding universe that space itself stretches and the recession of galaxies is not ordinary motion inside space. This solves the Herculean problem (arising in classical space) of how to organize for each galaxy the requisite original velocity so that the Hubble law is always fulfilled. Everywhere swelling space takes care of the Hubble law, like the raisins in a rising pudding, without any effort of their own, move uniformly away from each other.

Of course, it is required that there is a homogeneous substratum leading to uniformly expanding space. Whatever other problems the homogeneity requirement brings about (how the universe "knows" to be everywhere homogeneous), the automatic appearance of the Hubble law when space expands uniformly, has unquestionable inherent beauty.

8.4 The Hubble constant measures the age of the universe

Alexander Friedmann discovered the model universe which has the moment of creation a finite time ago. He also discussed the concept of a periodic world. The universe may expand into a maximum size, then contract into zero size, and then again expand and so forever. Though he was careful to note that "our information is completely insufficient to carry out numerical calculations and to distinguish which world our universe is", he gave an illustrative example leading to a world period of 10 billion years. The imaginable age of the universe had entered the range of billions of years.

Friedmann opened a new direction in cosmology where the age of the universe became measurable. And because of the simple notion of expansion from the big bang, the age determination is easy in principle: the world itself is like a cosmic pendulum! A galaxy at distance R recedes from us with velocity $V = H \times R$. Here H is the Hubble constant. So the simple rule "time is distance divided by velocity" yields the moment when the galaxy and the Milky Way were close together. And no matter for which galaxy you make this calculation, the resulting time is the same. This *Hubble time* is approximately equal to the age of the Friedmann universe. ‡

If one thus may calculate the moment when the universe started, one cannot resist asking what was before the beginning? This is not a new question. Plato wrote that Time came into being together with the spatial world where the movements of matter define time. In this sense, the world was not created *in* time and it is meaningless to ask what happened *before.*

Modern cosmology greatly approves Plato's answer to the question of time. There is a zero-point on the time axis. The axis and time do not exist "before" the zero, which denotes a singularity, of which next to nothing is known, expect that the density of the universe is there infinite. However, from the zero-point up to the present moment, one may try to describe the evolution of the world with a theoretical model where time ticks much the same as in physics generally. One may admit that if the model satisfactorily describes the world from the present time backwards down to some moment close to the zero-point, even if the nature of this point is unknown, there is reason to speak about the age of the universe.

8.5 The oldest stars – almost as ancient as the universe

If one could follow how a forming star changes when it as a protostar contracts from an interstellar cloud, it would be first seen to wander in the

---

‡From the Hubble law $V = H \times R$, the Hubble time is $T_H = R/V = 1/H$. The usual unit for the Hubble constant is km sec^{-1}/Mpc, and also 1/sec. The distance $cT_H = c/H$ is called the Hubble distance. It is about 5000 Mpc for $H = 60$ km sec^{-1}/Mpc .

HR diagram (Ch.7) and, when the fusion reaction starts, finally settle down in the Main Sequence. In that narrow band it will linger for a time, before it swells into a red giant. What determines how long the star will enjoy its peaceful, sun-like life in its Main Sequence phase? The key is the mass.

The masses of the stars cannot be arbitrarily small nor large. The smaller the star, the lower is its inside temperature. If a protostar's mass is less than 0.05 times the Sun, the temperature is too low to ignite the nuclear fuel. Such "failures" are called brown dwarfs. And for the biggest stars Nature seems to have put an upper limit of about 100 solar masses.

A star like our Sun will spend about 10 billion years in this mature phase of its life. A more massive star, also hotter and more luminous, will experience sooner critical changes in its interior and has a shorter life expectation. It is sad but true that the brighter the star, the shorter the life! Let us write down actual figures on the lifetimes (mass and luminosity in units of the Sun and Main Sequence lifetime in million years):

mass	luminosity	MS lifetime
30	140 000	4.9
5	630	68
1	0.74	10 000
0.5	0.038	30 000

Clearly, stars can have widely dissimilar life times. Except for the most massive stars having a very short Main Sequence phase, the total life time (⇒ protostar ⇒ MS star ⇒ red giant ⇒ stellar relic) is almost the same as the Main Sequence time. The other stages in a star's life are shorter.

The table reveals two more things. First, the short age of very massive stars, seen in abundance in the Milky Way, should convince one even without other evidence that the birth of stars is still going on: otherwise one would not expect to see any massive stars. Second, the lifetimes for stars smaller than the Sun are quite "cosmological". This grants the possibility to use the Main Sequence to derive a lower limit to the age of the universe.

Especially good objects for age determination are globular star clusters, dense stellar systems which may contain millions of stars. They surround galaxies as relics from the ancient times when their host galaxy was formed. Our Milky Way also has its swarm of over one hundred "globulars".

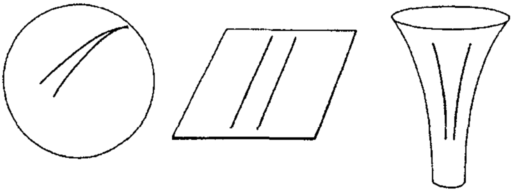


Fig. 8.3 Spherical, Euclidean, and hyperbolic geometries of the Friedmann universe. These 2-D analogies show how parallel lines intersect in spherical space and diverge in hyperbolic space.

In the HR diagrams of old globular clusters the stars which are presently escaping the Main Sequence have masses a little less than the Sun. Such stars must have landed on their maturing phase more than 10 billion years ago. The theory of stellar evolution leads to the conclusion that the oldest globulars of our Milky Way were born 13 - 18 billion years ago. This is close to what studies of the Hubble law have given for the Hubble time: between 10 and 20 billion years. The coincidence gives credence to big bang cosmology where galaxies have formed not long after the universe was born.

8.6 The geometries of Friedmann's world models

Allan Sandage once gave a definition which is hard to cut shorter: "cosmology is a search for two numbers". Indeed, Friedmann models can be characterized by only two quantities (forgetting for a moment Einstein's cosmological constant). One is the Hubble constant H. It says how fast the scale factor is growing at the present epoch. The other is the density parameter Ω (omega). It is the ratio between the present mass density of the universe and the critical density:

$$\Omega = \textit{cosmic mass density / critical density}$$

The value of the critical density depends on the value of the Hubble constant. In any case, it is a very small quantity, around 10^{-29} g/cm^3. § This density, rather "typical" for the universe around us, corresponds to mere 1/100 grams of matter in as large a volume as occupied by the Earth!

§ The density parameter $\Omega = \rho/\rho_{crit}$ contains the critical density which is $\rho_{crit} = 3H^2/8\pi G = 0.677 \times 10^{-29}$ g/cm^3 for $H = 60$ km sec^{-1}/Mpc.

Friedmann's cosmological models allow only three types of space geometry: spherical, Euclidean, and hyperbolic. These three uniform spaces have positive, zero, and negative curvature, respectively. The value of the density parameter is important, because it determines the geometry and the future fate of the expanding space as shown in the table.

geometry	curvature	density	volume	evolution
spherical	$K > 0$	$\rho > \rho_{crit}$	finite	expansion and collapse
Euclidean	$K = 0$	$\rho = \rho_{crit}$	infinite	eternal expansion
hyperbolic	$K < 0$	$\rho < \rho_{crit}$	infinite	eternal expansion

The finite model is like Einstein's spherical universe, but expanding. The expansion will eventually stop and change to contraction. The middle case for which the density is equal to the critical density has infinite Euclidean space. Forever expanding spaces with density less than the critical density also have infinite (hyperbolic) geometries.

Sandage's old definition – the search for two numbers – may sound a bit too restricted, but in fact it is concerned with fundamental issues: How old is the universe? Is its size finite or infinite? Is space Euclidean or not? What is the density of cosmic matter? What fraction of this density is in the form of ordinary matter? Today the search has come to include also Einstein's cosmological constant, but the questions are still much the same.

8.7 The cosmic density of matter in the universe

Thus there is a way to determine the geometry, if the universe is of the Friedmann type. Just measure the density of matter in all its visible and invisible forms! This is easier to say than to do. As a first step, studies of the amount of luminous matter in the form of stars in galaxies have established a firm lower limit to the density parameter in the local universe:

$$\Omega > \Omega_{lum} \approx 0.004$$

If galaxies were the only component of the universe, a density $\Omega = 0.004$ would mean that we are living in an infinite universe with curved hyperbolic space. However, astronomers can obtain information also on the *total* density parameter Ω, including dark forms of matter, from a variety of cosmological observations. Today it is popular to regard Ω as close to one. This implies that we can directly see via starlight only 0.4 percent of the substance making up the universe!

8.8 George Gamow's hot beginning

George Gamow 1904-1968

The early universe was first contemplated as a very hot place by George Gamow, discoverer of alpha-decay, the hot beginning, and the genetic code. He was a Russian physicist born in Odessa, studied at the University of St. Petersburg (Leningrad), specialized in quantum theory, and learned about the cosmology of expanding space from Friedmann. In 1928 he explained how alpha-particles can run away from their prisons, atomic nuclei, via a quantum-mechanical "tunneling-effect". This phenomenon also works the other way around and in fact makes the Sun shine! Inside the Sun two protons are supposed to be able to find each other and fuse, but the strong mutual electric repulsion would seem to prevent this happening. Even the 15 million degree temperature does not give them enough energy to cross the barrier. But tunneling helps the protons come so close that the strong nuclear force takes over and pulls them together.

In 1933 Gamow did not return to the Soviet Union from a conference trip to the West. [¶] He moved to live and work in America. In 1948 he suggested that the initial state of the universe was very hot, filled with thermal radiation. The expansion of the universe made this radiation cool down. Gamow's research group predicted that in our time its temperature should be about 5 degrees above absolute zero.

Gamow was also well known for his popular books on science. He created Mr. C.G.H. Tompkins, a modest bank official, whose adventures made relativity and quanta familiar to millions of readers. One may sense in the initials of the hero a dream of a relativistic, quantised, gravity theory...

[¶] George Gamow had earlier attempted, together with his wife, a "tunneling" away from the Soviet Union by a small rubber boat, starting from the Crimea. They set course towards Turkey, but the initially quiet Black Sea proved true Gamow's formula about the impenetrability of the Soviet border: where there are less frontier guards (A) there are more natural obstacles (B), or $A \times B =$ constant.

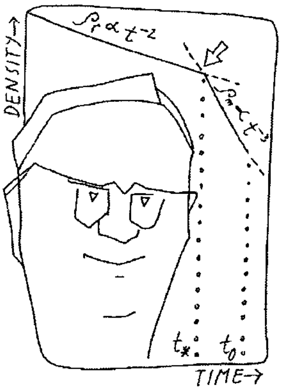


Fig. 8.4 Georgij Antonovich Gamov and his derivation of the temperature of the cosmic background radiation as visioned by A. Chernin

8.9 Discovery of the cosmic thermal radiation

In 1965 Arno Penzias and Robert Wilson, radio astronomers working for the Bell telephone company, were studying a communication hampering radio noise, as the pioneer Jansky had done 30 years earlier. They concluded that this noise originated from outside the Milky Way. The radiation had a thermal character with a temperature of about 3 degrees Kelvin. Such a low temperature radiation is most intense in microwaves (wavelength about one millimeter). For this discovery of the *cosmic background radiation* Penzias and Wilson were awarded the Nobel Prize in physics in 1978.

The radiation discovered by Penzias and Wilson was naturally understood as the relic of the hot beginning, as predicted by Gamow's team. It is currently regarded as the best evidence for the big bang, and is very difficult to explain in any other way.

Observations of the cosmic radiation by the COBE satellite in the 1990's have vastly improved our knowledge of its properties. It is perfectly thermal radiation with the temperature 2.725 ± 0.001 K. Its intensity is almost the same in every direction, excepting a temperature difference of 0.0033 K between two opposite sides of the sky, which reflects the motion of our Sun, at 400 km/sec, relative to the distant sources of the cosmic radiation.

On angular scales more than 7 degrees, COBE detected small temperature fluctuations so that the sky temperature may vary by about 0.00003

K from one direction to another. These are generally regarded as caused by primordial seeds of the large scale structure of the universe.

Instruments carried by balloons and also ground-based observations have detected temperature fluctuations on angular scales of about one degree. This allows the cosmologist to derive, using the Friedmann model, that the density parameter of the universe is close to one: $\Omega = 1.01 \pm 0.02$. The small error bars in the inferred density parameter are remarkable in cosmological research – if they really contain all sources of errors, including hidden systematic ones, then they indeed herald the coming of "precision cosmology" (the term coined by the cosmologist Michael Turner).

8.10 The 3 Kelvin glow – the cool relic of the hot bang

The universe has been filled with photons since the early phases of the big bang. One may suppose that originally this light-energy was created together with space and time and long ago light was much more important than matter! The energy of individual photons is decreased because the wavelength is stretched together with space. But the rest mass energy of a matter particle remains the same. This explains why radiation now contains much less energy than matter, even though photons are much, much more numerous than protons and electrons!

If one looks backward in time, as if watching a film about the evolution of the universe in reverse order, one sees that galaxies were in the past closer to each other and the density of the matter was higher than now. But one sees the radiation density increasing more quickly still – there was a time when the radiation had a higher energy density than the matter. The universe was *radiation dominated,* and it was much hotter than the present 2.7 K degrees. The temperature of the radiation has a simple relation to the redshift. So, the galaxies at redshift = 1 are bathed at a temperature two times higher than now, about 6 degrees K. [‖]

At large redshifts, around 1500, the temperature exceeds 4000 K which means that before that epoch all the hydrogen was ionized into a hot plasma of protons and electrons. There were no galaxies or stars at that time. The radiation interacted strongly with the electrons of the plasma, and this guaranteed that the plasma and radiation were in thermodynamic equilib-

---

[‖] How matter density, radiation density, and temperature behave as functions of redshift: $\rho_{matter} \propto (1+Z)^3$, $\rho_{rad} \propto (1+Z)^4$, $T(Z)/T_{now} = (1+Z)$.

rium, giving the radiation a perfectly thermal character.

When the expansion proceeded, the temperature decreased and electrons were captured by protons, forming ordinary hydrogen gas which offered little resistance to the radiation. The present day thermal glow is the cooled down descendant of those photons that decoupled from the matter around $Z = 1500$. This redshift also signifies the last visible barrier; from the times before that epoch we cannot receive messages carried by photons.

8.11 Cooking the light elements

The initial minutes offered a most dramatic act in the play called Big Bang – the first parts of the Table of Elements were written. This scenario was uncovered in the 1950's by the efforts of several theoreticians. Their calculations indicated that in the hot phase the two main chemical elements of the universe could have formed: the result of the nucleosynthesis would be about 25 percent of Helium leaving 75 percent of Hydrogen. These numbers are close to what is observed in stars and cosmic gas.

Helium was born about one minute after the big bang. Then the whole universe was for some time like the center of a star! Its temperature had dropped to 10^{10} K and it was now too "cool" for protons and neutrons to transform from one to another. Their number ratio was "frozen" so that there were about 15 neutrons to 100 protons. The neutrons combined with the same number of protons into helium nuclei (two protons + two neutrons) – via the same nuclear reaction that powers the Sun. The remaining protons were available for making hydrogen. Hence, the expected mass fraction of helium is about $30/(15 + 100) \approx 0.25$. Also, small amounts of Helium-3, Deuterium (H^2) and Lithium-7 were made. The heavier elements filling the entries of Mendeleyev's Table were born later, inside stars and in supernova explosions.

Dmitrij Mendeleyev 1834-1907

Observations of the abundances of light elements, together with the theory of big bang nucleosynthesis, have given still another precise cosmological number – the density parameter of baryonic matter (protons, atoms and other ordinary matter): $\Omega_{bar} = 0.060 \pm 0.006$. As the density of luminous matter is $\Omega_{lum} = 0.004$, stars and cosmic gas reveal to us a mere 10 percent of all baryonic matter. Does the bulk 90 percent of the predicted baryonic matter really exist and is its composition also 75 percent Hydrogen and 25 percent Helium? We do not know yet the answer to this crucial question.

8.12 After solving Newton's paradoxes of infinity...

Friedmann universes are not haunted by the old paradoxes of an eternal, infinite universe. The dark sky reflects the sparse forest of galaxies visible up to a maximum distance which in light years equals the age of the universe. The age of an average star is not far from that, so it is not surprising that there are so many stars shining in the bloom of their life. Also the heat death is not current in our relatively young universe.

We introduced the dark night paradox in Chapter 3 by saying that in infinite space filled with stars any line-of-sight hits upon a star. But why should the disk of a distant star be as bright as the solar disk? In fact, the brightness of a surface remains the same in a transparent Euclidean space, no matter how far away it is looked at. But not in the Friedmann universe! The surface dims quickly when one looks at more and more distant disks. Though this effect works in the right sense, it is in fact the small age of the universe which makes the sky dark, as was pondered already by Lord Kelvin. And the night sky is not totally dark, but glows the sum of light from all galaxies, which reaches us now and was emitted at different times in the past. This cosmic light is very dim and hard to detect, but still it is the real universe's pale rendition of Olbers' dazzling sky.

8.13 ...new enigmas of Friedmann's uniform world appear

Among all Friedmann universes the Euclidean space model (curvature $K = 0$ and density parameter $\Omega = 1$) is very special. If the universe was flat in the beginning, it will be flat forever. But if the density parameter in the beginning deviated from one by any minute amount, up to the present this slight deviation has grown enormously. This makes it very odd that now the observed Ω is not far from 1. ** If space initially was Euclidean then this is not surprising. But if the space had no reason to be Euclidean,

---

** Namely, looking backwards in time, Ω is seen to approach the critical value 1 arbitrarily closely. For a numerical example, we use a result from the theory of Friedmann universes: the density parameter Ω is connected with the scale factor S as $\mid \Omega - 1 \mid = const \cdot S^2$ Assume that at the present time $\mid \Omega - 1 \mid \approx 1$, i.e. Ω does not differ much from 1. When the universe had the age of one second, its scale factor was only about 10^{-8} times of the present scale. At that time, as the above formula shows, Ω had to be equal to 1 with an accuracy of 16 decimal places: it was somewhere between 1.0000000000000001 and 0.9999999999999999.

why then the parameter Ω was so exceedingly well fine tuned? This is the Euclidean paradox.

The thermal cosmic radiation is regarded as strong evidence for the big bang. However, another property, its almost perfect isotropy (equal intensity in all directions) is a riddle for the same cosmology. Why is the observed isotropy mysterious for the isotropic Friedmann model?

To see the core of the riddle, consider two parts of the universe, which are separated by a distance larger than which the light has travelled since the birth of the universe. Because the universe has existed only a finite time, these remote parts of the space cannot have "known about each other", as the speed of light is the maximum velocity for transmitting information. If the world is infinite, it was infinite immediately when it was born. Because of the finite speed of light, there are then regions which could not have been causally connected during the short existence of the universe.

One can easily point out two such regions in the sky. Consider any two opposite directions. The cosmic radiation photons coming from these directions were born about 15 billion years ago, hence – roughly speaking – their birth places are now separated by 30 billion light years. When the photons were born, the size of the universe was 1500 times smaller than now, hence these places had the distance of 20 million light years between them. But at that time the universe was only 300 000 years old!

An accurate calculation based on the Friedmann model shows that the cosmic radiation photons coming from directions which differ more than 3 deg (six full moons) have departed from regions which have never been in contact. Why then is the temperature all around the sky the same with an accuracy of 10^{-5}?! This is the isotropy paradox.

8.14 Inflation comes and resolves the paradoxes

In attempting to understand the mentioned paradoxes, one may ask what should have happened in the early universe, something not included in the ordinary Friedmann model. Thus the inflationary model was invented by the American physicist Allan Guth in 1981. It was further developed by Andrej Linde from Russia, in the form of a chaotic inflation. This model of the very early universe says that the cosmological constant was at first nonzero and the scale factor of the universe increased exceedingly fast due to the gravitational repulsion of the so-called false vacuum.

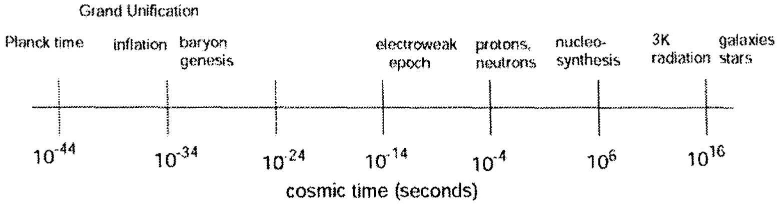


Fig. 8.5 Some milestones in the history of the big bang universe. Two epochs are especially important. First, at the time 10^{-12} seconds the electromagnetic force and the weak nuclear force became distinct. Second, 10^{13} seconds $\approx$ 300000 yr is the recombination epoch, when the cosmic background radiation ceased to interact strongly with matter. It is now observed as the 3K thermal glow uniformly filling the universe.

Inflation is an attractive idea, because it also gives an explanation of the "bang" and suggests an elegant solution for the riddle of isotropy. It can be shown that for the time before inflation, when the early universe was at the mature age of 10^{-40} seconds, all the matter inside a region with size 10^{-24} cm had sufficient time to be causally connected. Then the universe started to expand with rapidly increasing acceleration: the scale factor increased exponentially (as e^{Ht}), with the characteristic time $1/H = 10^{-34}$ sec.

During the inflation the scale grew enormously, more than 10^{25} times from its original size, and the causally connected 10^{-24} cm region achieved the size of a grapefruit around the age of about 10^{-32} sec. At that time the presently observable universe was compressed well inside such a volume where the matter and radiation had interacted, and so the temperature now can be the same all around the sky.

The inflation model also explains why the observed universe is not far from Euclidean. The inflationary expansion phase "flattens" the space as it were blowing up a balloon into extremely large size. The resulting space would be closely Euclidean across very long distances, somewhat similarly as Wright's huge sphere (Ch.4) gave for the local viewer the impression of a flat layer of stars producing the Milky Way.

8.15 The age of the inflationary universe

Let us examine the age of the universe with the critical matter density. When we see a galaxy with redshift Z, the age of the universe T_{age} at the time when the light was emitted, may be calculated by a simple formula if one knows the Hubble constant H. The present age of the inflationary universe is simply the Hubble time ($1/H$) multiplied by 2/3.

The observations put H somewhere in the range 50 - 70 km sec^{-1}/Mpc (Ch.16). In the table below we give "look-back time" T_{back} and "elapsed from the big bang time" T_{age} for a few redshifts, using $H = 60$. The ages are given in millions of years. ††.

Z	T_{back}	T_{age}
0	0	11000
1	7100	3900
3	9600	1400
5	10250	750
7	10490	510
10	10700	300

In this model, the present age of the universe is 11 billion years. Clearly, no celestial body can be older than the universe itself, notably if the universe started from the big bang singularity which could not contain even any elementary particles, not to speak of stars or galaxies.

8.16 When were the galaxies and their clusters born?

When the big bang universe was younger than about 100 000 years, photons carried most of the energy density. All ordinary matter was disrupted into electrons and protons and this porridge of charged particles was strongly influenced by photons which quickly dispersed any attempts by the matter to start forming dense blobs.

The expanding universe cooled down and when the temperature reached 3000 K (at the age of some 300 000 years), electrons and protons could unite and form neutral, non-ionized atoms. Then the matter started to evolve "on its own", under its own gravity in the expanding space. Radiation hardly interacted further with matter, while gradually getting cooler.

In terms of redshift, this revolutionary phase happened around $Z \approx$

---

††$T_{age} = (2/3) \times T_H/(1+Z)^{3/2}$. It is easy to adjust these ages to any other value of the Hubble constant – just multiply them by $60/H$

1000. After this, though before the epoch when we already can see galaxies ($Z \approx 7$), denser regions of the primordial gas started condensing into young galaxies producing stars. This long (about 1 billion years) period before the "first light" appeared has been aptly called the *dark age*, of which very little is presently known. It is difficult to penetrate the dark age and it is not known at what time the first true galaxies were born.

And what about larger observed organizations, like clusters and filaments formed by galaxies? There is a great urge to understand how such giant structures have been formed. Modern models of structure formation start from the primordial density fluctuations. One calculates how these grow into larger blobs of matter, gathered by gravity from the small seeds floating in the expanding smooth substance. The end result should resemble the galaxy universe as we see it now. To explain the lumps of galaxies, one inevitably requires large amounts of invisible dark matter in the form of unusual particles. The most popular models are CDM (Cold Dark Matter) and HDM (Hot Dark Matter). Some models mix cold and hot dark matter, in attempts to make more realistic structures. Dark matter is very important for modern cosmology: this mysterious substance will appear again and again in our story.

8.17 The big bang triumph – its logic and components

Summarizing our excursion through the big bang we list here the first principles and the main observations which make the triumph of this cosmological model. Two cornerstones of big bang cosmology are:

- *The Cosmological Principle of Uniformity*
- *General Relativity, valid for the whole universe*

The homogeneous matter distribution and Einstein's equations of general relativity lead to Friedmann's cosmological model and to the following interpretations for the key observations in cosmology:

- *The universe is expanding, producing the Hubble law*
- *The hot dense beginning gives rise to the thermal cosmic radiation and primordial seeds*
- *The universe is cooling, leading to the present low temperature*
- *Primordial seeds grow by gravity, forming large scale structures*

But this is not enough. Important ingredients must be added in order to provide the true triumph of the big bang:

- *Inflation – explains the isotropy of the cosmic radiation*
- *Dark matter – in effect makes the large scale structure and reduces the ripples of the cosmic radiation*
- *Dark energy – explains the accelerating expansion*
- *The cosmic recipe: about 70 percent of dark energy, 25 percent of cold dark matter, 2.5 percent of HDM, 2.5 percent of baryonic matter including 0.5 percent of shining stars and gas*

Dark matter builds structures without leaving a strong imprint in the cosmic background radiation (Ch.18), as inevitably would happen if the baryonic matter would cluster itself, without good help from dark matter. This list includes the very recent development of the big bang, the strange substance "dark energy". It is needed to explain the observations of very distant supernovae. More about that in Part III!

* * *

The tests of big bang cosmology, so far performed, seem to drive to completion the impressive triumph of the big bang paradigm which allows one to understand the cosmological redshift, the thermal background radiation, and the abundances of the light elements.

We might be tempted to finish the text at this point. However, this is not the end of the story, and the most dramatic turns are still forthcoming. Indeed, any theory of the heavens is not expected to be the ultimate truth. When the exploration goes deeper, new unexpected phenomena always emerge which make the respected old theories – old and respected.

PART III

THE ELUSIVE SIMPLICITY OF UNIFORM SPACE AND MATTER

Chapter 9

The mysterious singularity

The Friedmann model, based on uniformly distributed matter, tells that the universe expands and in the remote past had a unique event, the beginning, a state with zero extension and infinite density. From this enigmatic entity – the *singularity* – space, time, and matter were created.

In a singularity all ordinary physical laws break down and physics stands silent before this piece of space-time. John Wheeler, from Princeton University, has called this situation, when a fundamental physical theory inevitably leads to singularities, the "greatest crisis in physics". Have we really found the border of knowledge or is the singularity an alarm signal that the mathematical theory has been pushed too far, beyond its limits of applicability? We may be on the verge of the answer. Light on the nature of singularity is expected from gravitational waves emitted by exploding stars. And quasars, the energetic nuclei of galaxies, give information on both the singularity and the large scale arrangement of matter.

* * *

9.1 A uniform matter distribution leads to a singularity

In Newton's cosmology the stars were scattered uniformly through infinite space. This natural, innocent-looking assumption, leads to severe problems as we saw when we visited the paradoxical universe of Sir Isaac and were left without the beauties of the starry sky on dark nights.

The Friedmann world model also adopts a uniform matter distribution, but it is free from the blazing sky, because of the small age of the universe.

However, a new paradox appears. Galaxies are floating away from each other, hence there was in the past a moment when they all were at "one point", in the singularity. The unavoidable existence of the singularity in general relativity, was proven as a theorem by Roger Penrose and Stephen Hawking in the 1960's. The infinite radiation intensity scorching an infinite classical universe is like us being pushed into an instant in the early history of the Friedmann universe.

But how to probe the big bang singularity, as it is the only one and impossible to reach? In fact, the singularity is a strong prediction, not only for the whole universe, but also for very compact massive objects: extraordinary regions of space appear where an infinite space curvature develops. As such a black hole singularity is of the same nature as that of the big bang, the discovery of real black holes would serve as a proof for their cosmological big brother. No wonder that black hole candidates in our Milky Way and in distant quasars attract ample attention. They have opened new prospects for studying the physics near singularities and testing the nature of gravitation.

9.2 What is a black hole singularity?

Einstein's equations have as an exact solution unusual mathematical objects, first termed black holes by John Wheeler in 1967. The black hole is like a one-way gate, a trapdoor, through which matter can leave our universe but never return. Gravity on the edge of a black hole is so great that nothing, not even light, can escape. Small black holes (masses less than that of a mountain, one billion tons) do suffer from quantum evaporation. Predicted by Stephen Hawking in 1974, this process has not yet been experimentally tested. For black holes with stellar masses, the evaporation would take much longer than the age of the universe.

Karl Schwarzschild 1873-1916

The distance at which the black hole border – the *horizon* – lies is called the Schwarzschild or *gravitational radius*. Its value depends only on the mass of the matter that has collapsed into the black hole. The formula was derived by the German astronomer Karl Schwarzschild, who solved Einstein's equation for a single point mass at rest in empty space. This was in the last year of his life, when he served as a volunteer on the Eastern front. His surname means "black shield", an apt description of the border of the black hole...

If one compresses a body to a size smaller than its gravitational radius, then the escape velocity from the surface is greater than the speed of light. For example, if someone could manage to squeeze the Sun into a ball having a radius less than 3 km, the result would be a black hole. *

Anything that falls into a black hole reaches the singularity and is forever lost to the rest of the universe. However, there is an interesting difference between what an external observer sees and a brave explorer falling together with the matter will experience. The observer sees the explorer approaching the horizon of the black hole forever, but never reaching it. He also feels its gravity, as before. In this sense, black holes are never-ending processes.

The gravitational redshift caused by being close to the black hole increases quickly so that after a short time any signal coming from the explorer becomes too weak to detect. But what about the explorer himself? The theory says that he will cross the Schwarzschild radius after some finite time as shown by his watch, and he will not even notice anything special at that moment. Though, after this step of no return his fate is absolutely determined. In the wink of an eye huge tidal forces will tear the explorer apart and his remains will be swallowed by the ultimate monster.

9.3 Einstein objects to the physical reality of the singularity

It was discovered theoretically in 1939 by Robert Oppenheimer and his student George Volkoff, a Russian emigré, that a neutron star does not remain stable if it is more massive than the Sun. Such objects, says general relativity, will inevitably collapse into a black hole. Later it was shown that the Oppenheimer-Volkoff limit for the mass of a neutron star may achieve 3 solar masses. It is an oft used criterion that if one finds a compact dark object more massive than three suns, this object is a black hole.

Robert Oppenheimer 1904-1967

George Volkoff 1914-2000

It is interesting to realize that Einstein himself, the father of general relativity, wrote in 1939 an article where he attempted to show that singularities cannot exist. He had a simple argument, which practically reversed John Michell's reasoning that black holes may exist. In 1784 Michell noted that the gravity force of a mass may be so large that even light cannot flow out of it, but after a futile attempt falls back.

John Michell 1724-1793

---

*The gravitational radius is given by $R_g = 2GM/c^2$ where M is the mass of the body. For a solar mass, $R_g = 2.95$ km. The velocity for a particle to escape the body is: $v_{esc} = (2GM/R)^{1/2}$. If the radius of the body $R < R_g = 2GM/c^2$, then $v_{esc} > c$.

Fig. 9.1 Albert Einstein and Robert Oppenheimer in Princeton. In 1939 they published two opposite views on black holes. Oppenheimer calculated how black holes are formed. Einstein wanted to prove that black holes cannot exist in physical reality.

Einstein inspected the case where a particle is initially placed outside a compact body with a size less than the Schwarzschild radius. In essence, he noted that its free-fall velocity will exceed the speed of light when the particle reaches the surface of the body, because the final collision velocity is equal to the escape velocity. As superluminal motion is forbidden, Einstein concluded: "The essential result of this investigation is a clear understanding as to why 'Schwarzschild singularities' do not exist in physical reality."

Modern students of Einstein's heritage think that "the reluctant father of black holes" made an interesting conceptual mistake (even his mistakes – if they were such – were always interesting!). He considered compact stationary bodies, but appeared to overlook the non-stationary nature of space itself, within which a black hole is a continuous process of collapsing. In general relativity the space inside the horizon flows into the singularity and the velocity of the falling body relative to the space stream is found not exceed the speed of light. However, what was regarded as an oversight within geometrical gravity, was a brilliant intuitive guess in the frame of quantum field gravity in which there is no flowing space!

Einstein was not alone with his worries. One of the founders of modern astrophysics, Sir Arthur Eddington, was deeply disturbed by the predicted collapse into a singularity. Eddington's intuition told him that something essential was missing from the theory. To the end of his life he believed that there must be a physical law which prevents the ultimate collapse of massive compact stars.

Arthur Eddington 1882-1944

In 1979 at a conference dedicated to Einstein, the Danish physicist Claus Møller again raised the question about the reality of the singularity. *Mathematical* singularities do exist in general relativity and in other physical theories as solutions of equations. However, in physics one usually finds some physical reason to stop at a finite distance and ignore the ghost of the singularity. Møller suggested we should be looking for a gravity theory which reproduces all the positive features of general relativity in a weak gravity field, but which does not permit the singularity.

9.4 Are there alternatives to singularity?

"Has the reign of black holes come to an end?". These words opened the January 2002 issue of *New Scientist*, a prestigious weekly review of developments in science. They reflect new possibilities with which to struggle with the black hole singularities, found by Pawel Mazur and Emil Mottola. Early black hole theories ignored quantum effects, while these physicists consider the quantum structure of the physical vacuum and conclude that instead of black holes the universe may contain "gravastars". These could be extraordinary celestial bodies which look like black holes, but are free of a singularity and are supported by the negative pressure of the vacuum. Perhaps this is the way to avoid the singularity in general relativity, but the debate is only beginning.

A quite different reasoning against the singularity is offered by quantum field gravity. In his lectures Richard Feynman enthusiastically ascertained that gravitation is not "somehow mysterious", but may be understood as a field containing energy. Gravity is a genuine force, the concept of energy is well defined and conserved according to Noether's theorem. The energy of the field is important for the quantum interpretation of gravity and "every radiated graviton carries away an amount of energy". What about the singularity? Let us see how the conservation of energy determines a size limit for a particle with mass.

But first, let us look at a good example from electricity. The electric field implies that an electron cannot be arbitrarily small. Around an electron there is an electric field whose total energy *outside* a given radius has a definite positive value. This energy must be less than the rest mass energy of the electron, because the rest mass includes everything in the electron, also its electric field. The smaller we take the size of the electron, the larger is the field energy. Thus there is a critical size, called the *classical electron radius*, when the electric field energy is the same as the rest mass energy. If the electron were smaller, the field would possess more energy than the rest mass, which is contradictory (Appendix A.4). †

It is easy to repeat the above chain of arguments when a gravitating mass takes the role of the electric charge. A star is surrounded by a gravity field. Its energy can be calculated from a similar formula as for the electric field. And as the field energy cannot exceed the rest mass energy, it again follows that the star must be larger than a minimum size. This is one fourth of the Schwarzschild radius of a black hole (Appendix A.4)

Thus quantum field gravity contains a physical reason which prevents the singularity. Energy conservation puts a size limit for any star: to collapse into a size less than the gravitational radius, would require more energy than the initial rest mass energy of the star. Quantum field gravity does not require the existence of black holes! In his book Feynman did not consider this inevitable consequence of the positive field energy, but regarded it as a fundamental physical quantity.

9.5 Gravastars, eternally collapsing objects, dark stars...

The gravastar of Mazur and Mottola is made of the physical vacuum embedded in a non-vacuum shell. Matter falling into a gravastar strikes the shell and radiates much more energy than would be produced if it were falling into a black hole. A similar external appearance is expected from the eternally collapsing objects predicted by the Indian physicist Abhas Mitra. He has analyzed the physical meaning of the space–time inside the event horizon and made the surprising claim that "spherical collapse of a physical fluid in general relativity does not permit formation of trapped surfaces". This would mean that the falling explorer never crosses any kind

---

† In quantum physics the Compton radius of the electron $R_C = h/m_e c = 2.4 \times 10^{-10}$ cm appears, which is larger than the classical radius $R_e = e^2/m_e c^2 = 2.8 \times 10^{-13}$ cm.

of trapdoor, contrary to the current understanding of the black hole theory. Hence the radius of any collapsed body will exceed the gravitational radius. These two examples show that it is not excluded that even geometrical gravity might build relativistic objects with radii close to the event horizon, but which are not genuine black holes.

Furthermore, to make the world interesting even without black holes, the field theory has a vision of *relativistic dark stars* which are stable and compact, with radii not far from the Schwarzschild radius. Their existence may be anticipated from general physical arguments, but their properties are still unknown in detail. If real, such dark objects would be a strange class of celestial bodies, though not so outlandish as black holes.

One distinctive property of a dark star is that the gravity force at its surface not only remains finite, but decreases with increasing mass (the maximal acceleration $g_{max} < c^4/GM$, where M is the mass of the star). In black holes of any mass the gravity force at the Schwarzschild radius is always infinite. Because of the finite gravity force, dark stars may in principle have masses larger than 3 solar masses, the Oppenheimer-Volkoff border between neutron stars and black holes.

Another difference between a dark star and a black hole is that the former has no one-way horizon, but a material surface, and the escape velocity is less than, though close to the speed of light. The light oozing from a dark star loses almost all of its energy due to the strong gravitational redshift, and the star is practically invisible. But if such stars accrete surrounding gas, a large part of the rest mass energy of the gas may be converted into radiation, and they could be very bright. This makes distant cousins of "gravastars", "eternally collapsing objects", and "dark stars". They are competitors to black holes when high energy observational phenomena are interpreted. Theoreticians may debate the existence of such entities, but only observations can decide what Nature prefers to have.

9.6 Relativistic astrophysics probes strong gravity

Relativistic astrophysics is an exciting new branch of astronomy, which investigates celestial bodies where relativistic gravity effects are much stronger than in the Solar System. Such relativistic phenomena as the precession of rotating compact stars and the wobbling of the discs around them are identical for geometrical and field gravities, but there are other phenomena which can distinguish between the two gravity theories. The most promising tests of strong gravity come from studies of the following astrophysical phenomena.

- *the orbital motion of pulsars in binary systems*
- *gravitational radiation generated by relativistic collapse*
- *high energy radiation from X-ray binaries*
- *X-ray spectroscopy of supermassive objects in galactic nuclei*
- *multiwavelength radiation of violently variable quasars*

Our understanding of gravitation and the singularity will crucially depend on these observations, which have already given some amazing results.

9.7 A binary pulsar – an ideal gravity laboratory

The large radio telescope at Arecibo, with a dish of 300 meters, allows astronomers to measure with high accuracy the arrival times of weak pulses coming from distant pulsars, rapidly rotating neutron stars with strong magnetic fields. Especially interesting are pulsars revolving around another star in binary systems, which offer conditions for testing gravity theories not achievable in the solar system.

The famous binary pulsar PSR1913+16 (the discovery of which brought the Nobel Prize to Russell Hulse and Joseph Taylor in 1993) contains two neutron stars, each having the mass of the Sun and orbiting close to each other with the high speed of 500 km/sec. The relativistic gravity effects we know around our Sun, have now been measured with a much higher accuracy in the orbital motion of this binary system. The small perihelion shift of Mercury (43 arcsec/century) here becomes 4 degrees per year!

The superbly accurate observations by the Arecibo telescope have revealed that the size of the orbit is slowly decreasing with time. The cause is the gravitational radiation generated by the huge acceleration of the orbiting neutron stars. The shrinking rate is well predicted by general relativity. Still, a slightly "too" rapid contraction of the orbit has been observed. It seems the binary is emitting gravity waves one percent over that predicted by general relativity. Future observations will show the true excess, if any. Incidentally, quantum field gravity predicts an excess of 0.7 percent in the form of scalar gravity waves. A part of the observed excess may be explained by the rotation of the Milky Way, but the poorly known distance to the pulsar (somewhere between 3 to 8 kpc) does not allow a precise estimate of this source of error. Astronomical distances are a permanent problem! We are looking forward to an accurate distance in a decade when the astrometric satellite GAIA will be launched.

9.8 The search for gravity waves from collapsing stars

In his paper on special relativity in 1905 Henri Poincaré noted that if gravity is a relativistic phenomenon like electromagnetism, then there should be gravitational waves traveling with the speed of light. A decade later Einstein showed that general relativity predicts such moving ripples of space.

Making electromagnetic waves is easy. If one shakes an electron, the variations of the electric field around it will propagate through the space as electromagnetic waves. Making gravitational waves is just as easy. Accelerate a massive body and it will radiate waves as disturbances in the gravity field. The problem is that the waves are very weak. Only violent cosmic phenomena, such as exploding stars, can produce so much gravity wave energy that their detection becomes possible.

Joseph Weber 1919-2000

The idea of detecting gravitational waves with large, massive cylinders was proposed by Joseph Weber in 1960, who for long years was a lonely and persistent pioneer of gravity wave searches. The method is simple in principle. A gravity wave changes the distance between any two material points which it passes by, making a metal bar shiver or ring like a bell. In practice, the expected shifts are so small (say 10^{-13} cm) that even very small disturbances (e.g. distant traffic) easily hide the tiny signal. [‡]

The new generation of gravity antennae are now coming in two types. One is the metallic bar, like Weber's original detector, but kept at a chilly temperature of 1 K degree. For a period of a few weeks the temperature can be lowered down to a few thousandths of degree K. Highly sensitive bar detectors *Allegro* in Louisiana and *Explorer* in Geneva have a 1.5 ton bar made of aluminium or niobium, and placed in a cryostat.

In another antenna type a laser measures the distance between sus-

---

[‡] The dimensionless amplitude of a gravitational wave is $h = \Delta l/L$, where Δl is the change in the length L of the antenna affected by the wave. The signal h detected on the Earth depends on the distance r to the exploding star, the radiated energy E_{gw}, the frequency of the wave ν_0, and the duration of the signal τ_g so that $h = 1.4 \times 10^{-20}(1\ Mpc/r)(E_{gw}/1\ M_{\odot}c^2)^{1/2}(1\ kHz/\nu_0)(1\ sec/\tau_g)^{1/2}$

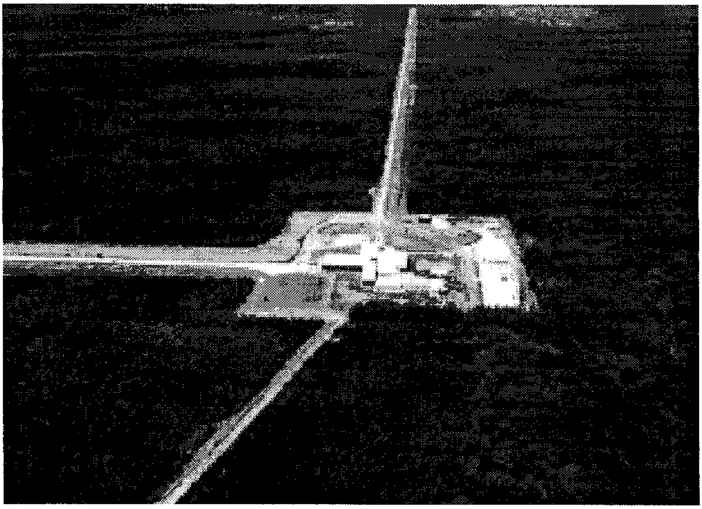


Fig. 9.2 An aerial view of the LIGO gravitational observatory in Livingston. The antenna of the observatory consists of two long vacuum tubes (left and up in the picture), each extending over four kilometers from the central corner station. (Provided by LIGO)

pended test masses (mirrors). The LIGO antenna (Laser Interferometric Gravitational Wave Observatory) in the USA consists of two detectors of this kind, separated by 1000 kilometers – a genuine gravity wave passing through the Earth should be seen at both sites. Lasers measure the shifts of the mirrors set in motion by gravity waves. A similar gravity observatory VIRGO will operate in Italy.

Recent studies of quantum field theory predict many new particles, not yet observed in laboratory. In particular, gravity interaction could be mediated by two kinds of particle (or field): tensor particles corresponding to the usual gravity waves, and scalar particles which actually are mediators of the repulsive force. Orbital motion in compact binary stars gives rise to the ordinary tensor waves predicted by both general relativity and field gravity. In this case the scalar waves are only a small addition to the radiation. However, for spherically pulsating stars general relativity forbids the gravitational radiation, while field gravity predicts the emission of scalar waves. The new antennae are sensitive to both tensor and scalar waves, and may revolutionize our understanding of gravitation; for example, scalar radiation is forbidden in classical general relativity.

It is necessary to have highly sensitive instruments. During the 40 year history of gravity wave searches there have been only two *possible* recordings which may have been related to gravitational waves. These events were the explosions of the supernovae SN1987A and SN1993J. With the new detectors one expects to see the signals from supernova explosions about once a month emerging from nearby big clusters of galaxies, such as Virgo. If they are there to be seen, we should soon start getting routine detections.

9.9 Two closest supernovae – signs of gravity waves?

Our satellite galaxy the Large Magellanic Cloud became in 1987 the host of a supernova called SN1987A. At a distance of 50 kpc, this was the closest supernova explosion in modern times. Luckily a gravity antenna *Geograv* was working at that time in Rome. A 2300 kg aluminium bar at room temperature, it was constructed by Edoardo Amaldi, a student of Enrico Fermi. It detected a strong signal simultaneously with the arrival of neutrinos at the underground Mount Blanc neutrino observatory. Neutrinos are produced when a supernova explodes, and are expected to arrive at the Earth side by side with the gravity pulse.

But something was strange about the signal. If it was a short (millisecond) gravity wave pulse, then the mass of the exploding star had to be fantastically large, more than 1000 solar masses. Old astronomical photos of the Large Magellanic Cloud show that at the explosion site there was a star with the mass of about 20 suns. This discrepancy has led astronomers to reserve judgement on the signal in Geograv as an accidental, unexplained event not related to the supernova explosion.

Then in 1993 a supernova flashed in the galaxy M81 at a distance of 3 Mpc, the next closest supernova since SN1987A. At that time two bar detectors, Allegro and Explorer, were working at a low temperature of a few Kelvin. The data suggested that there may have been a signal from SN1993J. But its strength, if calculated for a short pulse, again demanded that the exploded star should be as big as 1000 solar masses. Were these incidents with the two brightest supernovae in the last twenty years just coincidences or did the signals tell something about gravity physics? [§]

[§] At a meeting on gravity waves, held in 1994 in Frascatti (Italy), one of us (Yu.B.) discussed with Joseph Weber the event recorded by Amaldi's Geograv. Weber expressed his opinion that in general relativity there are two major problems: the theoretical prediction of the singularity and the inability to explain the signal from SN1987A.

Let us play with a non-standard view of supernova explosions. Perhaps the pulse of gravitational radiation from a collapsing massive star is made of a long oscillating signal, having a total duration of one second, rather than milliseconds, and comparable with the duration of the neutrino signal. Then the recordings by Geograv, Allegro, and Explorer could be true detections of gravity waves from SN1987A and SN1993J. An exploding star of 20 solar masses would lose to gravity waves 10 percent of its mass, sufficient to explain the observed signals. ¶ Future detectors will clarify the properties of the gravity pulse, the last chirp of a dying star.

9.10 X-rays betray black holes in binary stars

When you look at the night sky, you hardly suspect that many of the stars are actually double, with their two components revolving around each other. But for astronomers who can, with their telescopes and spectroscopes, see and measure them, the double stars are very real indeed – they are for us practically the only source of information about the masses of stars! And they are also crucial in the hunt of black holes.

Many double stars contain an invisible compact massive object plus a normal star. These systems are detected thanks to the strong X-rays emitted by the *accretion disk* rotating around the dark body. The disk is formed when the compact body attracts gas from its companion star. The gas disk gets so hot from the strong gravity of its central object that it starts radiating X-ray photons.

How do astronomers study X-ray binary stars? For example, a binary in the constellation of Cygnus consists of a normal star of about 20 solar masses and a companion. The latter is covered by dense gas clouds which emit X-rays (hence the designation Cygnus X-1), and cannot be observed directly. But the revolution of these stars around each other is betrayed by spectral lines shifting periodically due to the Doppler effect. From such observations, one may infer the masses of the stars using good old Newtonian mechanics, because the stars are moving rather slowly, not relativistically.

Observations of X-ray binaries have established almost beyond doubt that there are dark, compact objects with masses exceeding three solar

[¶] The frequency band of the long oscillating signal ($\tau = 1$ sec) is about 1000 times smaller than for the short signal ($\tau = 0.001$ sec). This reduces the required mass 1000 $M_\odot$ for the short signal to about 1 $M_\odot$ for the long signal.

masses, the Oppenheimer-Volkoff limit. In our Milky Way galaxy more than a dozen such *black hole candidates* have been discovered, ranging from four to twelve solar masses. In other galaxies a few compact relativistic stars have been found up to 100 solar masses. The Chandra X-ray space observatory even detected in the central region of the unusual galaxy M82 an object 700 times the mass of the Sun. Dark compact objects seem to exist – but are they really black holes? How to authenticate the candidates?

In 1974 Stephen Hawking and Kip Thorne, both well known for their works on the theory of black holes, made a bet about whether or not Cygnus X-1 contains a black hole. If this binary star contains a black hole, then Thorne is the winner. In 1990 Hawking finally admitted victory to Thorne. But perhaps the bet should still be open! After all, even if the compact star has a large mass, it is not necessarily a black hole. To measure the mass is only half the battle.

To show that this object is a black hole and not a dark star, one needs a clear proof that it is really a "one-way gate", has no hard surface (as the falling observer could testify...), and only swallows ambient matter, in other words, it has a horizon. The most direct way to prove that singularities exist is to observe the process of formation of the black hole horizon. This is impatiently anticipated from the new gravitational observatories. According to the current theory of the collapse leading to a black hole, there must be a short single gravity wave pulse (one millisecond or so). In the case of a dark star (as formed in field theory) the gravity signal should consist of many pulses over a much longer time, comparable with the duration of the neutrino signal (about one second).

9.11 The best candidate sits at the center of the Milky Way

Besides X-ray binaries containing relativistic stars with masses in the range from one to 100 suns and the more massive compact X-ray objects up to 1000 Solar masses, there is evidence that the nuclei of galaxies harbor much more massive relativistic objects.

The closest galactic nucleus is the center of our Milky Way, in the direction of the constellation Sagittarius. At a distance of about 8 kpc, it is hiding behind a thick layer of dust. But careful studies of this region of space, both with ground based and space observatories, have revealed extremely interesting phenomena in the inner few parsecs at the center.

Stellar motions have betrayed a dark compact massive object of about 2.6 million solar masses. It is now regarded as the best supermassive black hole candidate. Coincident with a compact radio source Sagittarius A, the super mass is surrounded by a cluster of stars, a dusty ring, and hot gas.

Many astrophysical phenomena such as the streaming and radiation of gas are connected with the central mass. They may reveal effects of strong gravity and even the black hole horizon and thus may permit one to test this chief prediction of general relativity (see below).

9.12 Supermassive objects in the nuclei of other galaxies

The Milky Way is not alone. Thanks to the high resolution of the Hubble Space Telescope, astronomers have been able to measure the motions of stars in the very center of tens of other galaxies. In order to explain the observed velocities, one has to assume that both elliptical and spiral galaxies harbor in their centers very compact supermassive objects, speeding up the stars. These huge things of up to billions of solar masses are now viewed as important structure components of galaxies. They could be black holes.

When a host galaxy feeds its supermassive black hole by gas and stars, the black hole becomes a very efficient energy machine which appears to us as an active galactic nucleus, such as a quasar. Its power corresponds to the conversion of the rest mass of the Sun into photons every year, impressively equivalent to the light radiated by 1000 normal galaxies (each with their 100 billion stars...). Quasars show rapid variability of their light, which is possible only if they have small sizes. Brighter than 1000 galaxies, having millions of solar masses within a small volume less than the Solar System, the quasars are astounding even for the hard boiled astronomer.

It is even thought that some nuclei contain two or more supermassive objects dancing around each other. Why does one expect more than one super mass in a galaxy nucleus? Galaxies form groups and clusters inside which they are moving. Occasionally, a galaxy comes close to another one, and as a result of such near-collisions galaxies tend eventually to merge. Mergers form a new generation galaxy which then carries in its core the supermassive objects of its parents. Thus multiple super masses in galactic nuclei should be a common phenomenon, though they usually do not make as much noise of themselves as happens in active nuclei like quasars.

Because such objects have sizes comparable to the Schwarzschild radius, they are promising black hole candidates. With a mass of one billion solar masses, their radius would be 3 billion kilometers, about the Sun – Uranus distance. However, as is the case with the stellar mass black hole candidates, one needs independent information on the nature of their surfaces and other properties, before they can be said to be genuine black holes.

9.13 Approaching the horizon...

Observations of accretion disks permit the study of the region close to the putative horizon. Space observatories record the time variability of radiation coming from the inner parts of disks. To estimate how close to the center this emission originates, one can utilize causality. The size is less than the time needed for the intensity to double multiplied by the speed of light, because more widely separated parts would not be able to "march in step". In order to reach the inner edge of the accretion disk – the closest stable orbit around the black hole at three Schwarzschild radia – one must be able to measure time intervals shorter than one millisecond.

In order to see the horizon of the black hole in the center of the Milky Way, astronomers have proposed observations which are achievable in near future. One should make a radio map of the radio source Sagittarius A, using VLBI imaging techniques at submillimeter wavelengths, with an angular resolution of about one microarcsecond. [‖] The surrounding gas is transparent at such short radio wavelengths and one may view the emitting gas all the way down to the black hole horizon. Calculations by Heino Falcke, Fulvio Melia and Eric Agol have demonstrated that it will be possible to see the "shadow" of the black hole horizon in front of the glowing background gas. It should be a really black spot in the center of the radio image. The angular size of the shadow is predicted to be about five gravitational radia. This is so much larger than the radius of the black hole just because of the strong relativistic bending of light.

This experiment could test various alternatives as to the nature of the dark mass at the center of our galaxy. For example, astronomers have proposed that such a compact object could consist of fermions or bosons or

---

[‖] For a mass $M_{bh} = 2.6\ 10^6\ M_{sun}$ the gravitational radius of the black hole is $r_g = 2GM/c^2 = 7.8\ 10^{11}$ cm, which corresponds to 10 Solar radia (about one tenth of the orbital radius of Mercury). At a distance of 8 kpc the angular size of the Schwarzschild radius is $(r_g/d)206285 = 6$ microarcsecond $= 6\ 10^{-6}$ arc second.

cold dark matter. In general relativity the radia and hence the shadows of such masses should be greater than the gravitational radius. However, in the case of a dark star it is expected that its shadow will be smaller than for a black hole of the same mass. This is because now the radius may be as small as 1/4 of the black hole radius (but not smaller).

9.14 ... may offer unexpected surprises

Whenever there is available gas in the vicinity of a black hole (e.g. from a companion star), there should be an accretion disc and strong X-ray radiation. However, there are unusual binary systems, termed X-ray novae. Their bigger component is a dark compact object with a mass of ten suns, while the smaller, revolving around the dark primary, is a normal shining star, about similar to the Sun. In these systems strong X-ray flashes are followed by quiescent periods. If during the feeble phase the radiation is coming from the accretion disk, then its faintness is very difficult to understand – unless the energy goes into a black hole! A theory of *advection*, proposed by Ramesh Narayan and his team at the Center for Astrophysics in Cambridge (the USA), attempts to describe what happens. Briefly stated, it assumes that the thermal energy stored in protons cannot be transferred to (the lighter) electrons from which the observed X-rays come. So the energy is not radiated away, but is swallowed by the black hole. The paradox – why does the system with a black hole *not* radiate! – would thus seem to lead to a proof of the trapdoor, the horizon.

However, a severe problem for the advection flow arises when a magnetic field exists in the accretion disk. Gennadij Bisnovatyj-Kogan from the Astro Space Center in Moscow and Richard Lovelace from Cornell University have shown that the radiation efficiency of any flow of a hot gas with a magnetic field cannot be less than one fourth of the value predicted by the usual accretion disk theory. Hence, it seems that other explanations than advection are needed for understanding the quiescent phase of X-ray novae. Indeed, Stanley Robertson and Darryl Leiter have suggested that these objects have a surface and magnetic fields which explain the observed properties of the black hole candidates.

When we look beyond the Milky Way, it happens that the Seyfert galaxy MCG-6-30-15 plays an important role in the observations of strong gravity effects close to the putative horizon of a supermassive object. The European

X-ray satellite XMM-Newton ("MM" means multiple mirror) has observed, thanks to its fine spectral resolution, the Iron spectral line originating in the inner edge of the accretion disk around the supermassive object in the nucleus of this active galaxy. To explain the observed form of the Iron line, the inner edge of the accretion disk should lie at an unexpectedly close distance of 0.6 R_g from the central object.

This impressive observation was even featured on the front page of the international Herald Tribune in October 2001, the same day that the results were presented to the scientific community. The headline was fittingly "Bright glow may change the dark reputation of black holes"... Theoretically, the minimum distance from which the X-ray radiation can come, is 3 R_g (the last stable orbit of the matter around the black hole). At smaller distances the matter would be immediately swallowed without emiting radiation. It is alarming that the observed inner edge of the disk is inside the black hole horizon. This is impossible for static black holes. But there is a loophole. Extremely rapidly rotating (close to the speed of light) "Kerr black holes" may have a radius of 0.5 R_g.

But still another problem raises it head. The concentration of the X-ray radiation towards small radii is too high to be explained by any accretion disk models. How can there be such powerful energy generation so close to the horizon even for an extremely rapidly rotating black hole? Whatever the resolution of this enigma, astronomers are now knocking the (trap?)doors leading to the interiors of the supermassive compact objects.

9.15 The rapid variability of quasars as probe of gravity

A famous variable quasar is OJ287, a faint 15th magnitude stellar-like object in the constellation of Cancer. Its regular variability, with a period of about 12 years, was verified in 1996 by an international effort coordinated by Leo Takalo at Tuorla Observatory. Such a rhythm had been inferred from older data, analyzed by Aimo Sillanpää and collaborators.

The regular variability in the light suggests that something very big is revolving. A model of the OJ287 system is shown in Figure 9.3. It consists of a heavy and a light black hole, although both are supermassive. The lighter black hole is in orbit around the heavier, and on its 12 year orbit punches through the heavier black hole's accretion disk, generating the flaring outburst we observe at the Earth. Mauri Valtonen and Harry Lehto

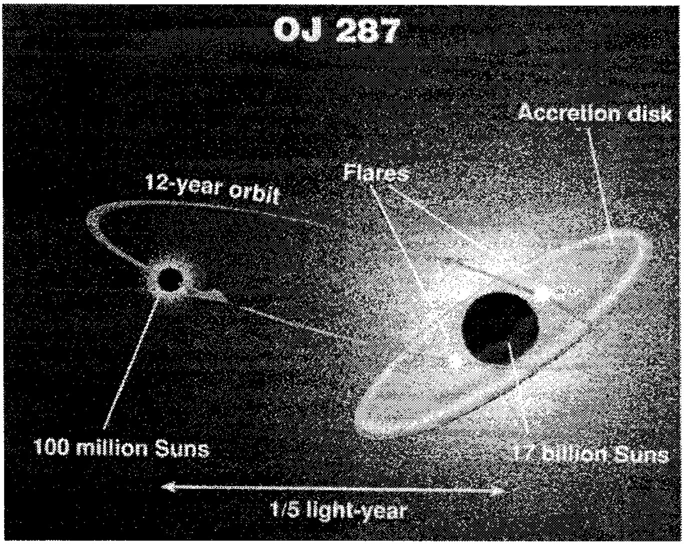


Fig. 9.3 The binary black hole model for the highly active quasar OJ287, as developed at Tuorla Observatory, explains the observed periodic variability of its light.

have even proposed that long term observations of such systems might be used for testing the predictions of general relativity about the evolution of the orbit of a massive binary system. **

9.16 Cosmology requires relativistic and quantum gravity

A viable cosmological model must meet many requirements, two of the most important being that it should be based on relativistic and quantum gravity theory. Let us briefly elucidate this point.

In terrestrial laboratories it is natural to effectively regard the space outside the Earth as an empty cosmos. Physicists may even forget that there is an external universe around the lab. For stellar astronomy the void begins just outside the Milky Way galaxy. But the luxury of ignoring surrounding space is over in extragalactic astronomy: there is no empty

---

**OJ287 is not the only regularly variable active galactic nucleus. Astronomers from St. Petersburg, Michail Babadzhanyants and Elena Belokon, a husband and wife team, have studied active galaxies since the 1960's. They have found three other objects which brighten about every ten years: 3C273 ("the first quasar"), 3C120, and 3C390.3.

space beyond the galaxy universe. The realm of galaxies has no edge.

This has an important consequence for the physics of the gravity of the universe. We recall that ordinary celestial bodies (e.g. galaxies) have weak gravitational fields – i.e. the value of the gravitational potential energy is much less than the rest mass energy of the gravitating body. But imagine that galaxies fill space uniformly. Then the mass of all galaxies contained in a sphere around our Milky Way increases with the distance cubed. Hence, the potential energy will at a certain distance inevitably reach the large relativistic value of Mc^2. This distance defines a spherical portion of the universe, which may be regarded as a new kind of object – *Hubbloid*, for which as a whole Newton's gravity is no longer valid. Its radius is not far from the Hubble distance. If the cosmic density is equal to the critical density of the Friedmann universe, then these two measures of remoteness are the same and designate the beginning of the "deep universe". [††]

The size of the Hubbloid is equal to its own gravitational radius. Thus it is a relativistic object, and this is why one must base homogeneous cosmological models on relativistic gravity. It is a pity to say that this makes cosmological physics a difficult science – as the Finnish cosmologist Tapani Perko put it, after Richard Bentley, "most clergy and laymen were forced out of cosmology through the power of tensor calculus".

But what if the mass of a sphere does not increase as the cube of the radius, but slower, directly proportional to the radius? The size of the Hubbloid then goes to infinity and only then Newton's gravity would be good for the whole universe. But any uniform background, such as photons, neutrinos, or the energy of the physical vacuum, would make such a classical world model fail. In any case, Newton's theory cannot explain the relativistic gravity effects even in our solar system and cannot be considered as the basis of modern cosmology.

General relativity is the main factor behind the very succesful Friedmann models and the triumph of the big bang. However, the bang itself is an enigma, with its singularity. The requirement for a quantum gravity theory appears when one looks back in the very deep past, when the size of the whole presently observed universe was less than so called Planck length, 10^{-33} cm. This length is a part of quantum Nature and hence general rel-

---

[††] For uniform matter the mass within the radius R is $M(R) \propto R^3$. Hence, the gravitational potential will be $\phi_N \propto M/R \propto R^2$. The distance where the potential achieves the value of c^2 is the gravitational radius of the Hubbloid $R_g \sim GM/c^2$.

ativity should be superseded by a quantum gravity theory, before one can say anything about the things closer to the singularity.

* * *

It is difficult to imagine a more intriguing thing than the singularity out of which the whole universe cropped up and into which all the matter may disappear. The singularities are no science fiction, but serious business at the frontier of science. Nevertheless, their reality is still an open question. There is optimism in the air, fed by new technology, that soon the existence and properties of the singularity can be determined by real observations at observatories. Gravitational waves detected by the next generation gravity antennae, together with the study of light from binary stars and quasars using radio-, optical and X-ray telescopes, will bring us closer to the puzzling nature of gravity – the force creating the cosmic order.

Chapter 10

Dark matter – the grey eminence

The discovery of huge amounts of invisible matter in the universe is an absorbing thriller in which the last page is still to be written. Astronomers see through their telescopes only the luminous matter, in the form of ordinary stars and gas. But the large majority of cosmic substance hides in the deep shadows. This *dark matter* discloses itself by its gravity effect on light rays and on the motion of nearby matter.

Dark matter exists, but its nature and composition is a deep riddle of modern cosmology. Its spatial distribution is currently one hot topic in astronomy. Is it a faithful companion of luminous matter or is it more smoothly dispersed, filling the holes between the galaxies? Gravitational lensing has now betrayed the presence of dark matter and shown its 3-D distribution appears to be very lumpy.

* * *

10.1 Early signs of dark matter

In 1925, Knut Lundmark attempted to weigh the Milky Way. Using sparse information on stellar velocities and the size of our galaxy, he inferred that 10^{12} solar masses would explain the motions. The estimated number of visible stars was a hundred times less, about 10^{10}. Lundmark concluded that dark stars and dark matter exist, introducing the latter term.

Then Fritz Zwicky made a remarkable finding about clusters of galaxies. He used a new method for weighing the clusters, based on the virial theorem of classical mechanics. * In 1937 he reported the result for the Coma cluster, a swarm of thousands of galaxies in the constellation Coma Berenices. Its total mass was 500 times larger than what one would expect from all the light emitted by the galaxies, if these are made of stars like our Sun!

10.2 Invisible matter makes galaxies revolve rapidly

When one looks at the beautiful photographs of spiral galaxies, nothing makes one suspect that they are made of anything else than bright stars, gas clouds, and some dust.

This natural view was shattered in the 1970's, when Vera Rubin and Kent Ford started extensive studies of the rotation of galaxies. From the Doppler shifts of the spectral lines of light from the gas clouds, they inferred how fast spiral galaxies rotate at different distances from the center. It was expected that the rotation would greatly slow down at the visible edge of a galaxy. But not so! The galaxies continue to spin at a high rate even at large radii where stars are no longer seen: the rotation is not ruled by visible matter, but by something else.

The rotation of our own galaxy is difficult to determine, because of our position within. Nevertheless, one may conclude that the Milky Way revolves as rapidly as other similar spiral galaxies, having a rotating velocity of about 220 km/sec at radii more distant than where the Sun is orbiting.

Such observations are interpreted as evidence for *dark haloes* around galaxies. The halo matter, exceeding by several times the mass of the visible stars, makes itself known only by its gravity force. It does not emit any observable light and we know very little about its composition.

But does the dark matter really exist? There has been a habit to ascribe mysterious visible phenomena in astronomy to something invisible, sometimes succesfully (recall the planet Neptune!), but not always. To take a very old example, Anaxagoras made dark bodies circling the Earth below the Moon, responsible, along with the Earth's shadow, for eclipses. Naturally enough, astronomers have attempted to explain away modern dark

---

*The virial theorem states that the potential energy of a system of gravitating particles in equilibrium is twice its kinetic energy: $U = -2E_k$, where $U \approx -GM^2/R$ and $E_k \approx Mv^2/2$. M is the mass of the system, v is the velocity dispersion of the particles, and R is a characteristic size of the cluster. Hence, the observed size and velocity dispersion give the total mass of a system, including the dark mass: $M \approx Rv^2/G$.

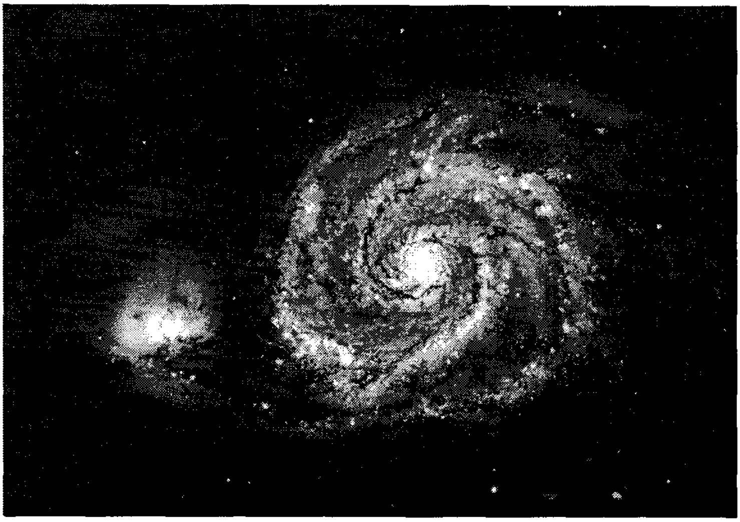


Fig. 10.1 The nearby spiral galaxy M51 in the constellation of Canes Venatici is rather similar to the Milky Way. Its rotation betrays the presence of an invisible massive halo in which the galaxy is embedded. Johan Holmberg and Chris Flynn have shown, using observations made by the HIPPARCOS satellite, that in our Milky Way the dark matter is not detectably concentrated into the flattened luminous disk.

matter, as a result of observational uncertainties, inadequate methods or even new properties of the Newtonian gravity force. But different kinds of independent evidence have surfaced, making an increasingly strong case for the gravitating dark matter.

10.3 Gravity lenses probe the dark matter

A novel way to probe the dark matter came from *gravitational lenses.* The essence of the phenomenon is simple: if on the line of sight between the observer and a distant object (a star or quasar) there happens to sit a massive body (the lens), its gravity field deflects the light rays and may focus them toward the observer. The role of the lens can be played by stars, galaxies, and also any dark celestial bodies. †

---

† Gravitational lensing is an effect of weak gravity, similar as the bending of light by the Sun. The resulting bending angle is $\theta = 4GM/Rc^2 = 1.75''(M/M_\odot)/(R/R_\odot)$, where M is the mass of the body, and R is the distance from the center of the mass, at which the light ray passes by. $M_\odot$ and $R_\odot$ are the mass and radius of the Sun.

Orest Chwolson 1852-1934

The lensing effect was first calculated in 1924 by Orest Chwolson, at the University of St. Petersburg. If the line of sight between the observer and a distant star happens to intersect another star (the lens), then the observer sees a luminous ring around the star. The ring is the lensed image of the more distant star. Chwolson also predicted that if the lens is not strictly on the line of sight, then the observer sees a double image instead of a ring. In 1936, Einstein independently made the same calculations, and pointed out that this effect must be hard to observe. If the lens is like our Sun, the angular radius of the ring will be hopelessly small, about one thousandth of an arc second for distances inside the Milky Way, and much too hard to observe.

After Einstein's calculations, Fritz Zwicky pointed out that galaxies, billions of times more massive than stars, could produce rings with sizes of a few arc seconds. But it took almost four decades before the first gravitational lens was detected. In this case, a galaxy formed a double image of a distant quasar. The two images, slightly separated by 6 arc seconds in the sky, have identical spectra, demonstrating they are two images of one, single object. Nowadays, tens of lensed images are known, among them several Chwolson–Einstein rings. In the literature these phenomena are called "Einstein rings", confirming a proverb that "The biggest cat gets all the milk"...

Astronomers recognize gravitational lenses in the sky as mirage-like effects. These are the splitting of images of remote objects and, what is very helpful, the photon flux coming from the object is much magnified, even a hundred times. Gravitational lenses are gigantic natural telescopes which allow one to see very faint and distant galaxies. They are a good tool for studying dark matter, giving information on the total gravitating masses and sizes of the lenses themselves.

Another way to shed light on dark matter is to observe the orientations of elongated galaxies – fingerprints of "weak" lensing – in a field of distant galaxies. The result is a map of the clustered component of dark matter. For this purpose a team of astronomers from the Bell Laboratories, headed by Anthony Tyson, has proposed the building of the Large-aperture Synoptic Survey Telescope (LSST) with three 8 meter mirrors, dedicated to study the images of faint galaxies (as faint as 29 mag) in fields of 7 square-degrees. Such a telescope would map out dark matter in tremendous detail, but still leave us wondering what all that stuff is.

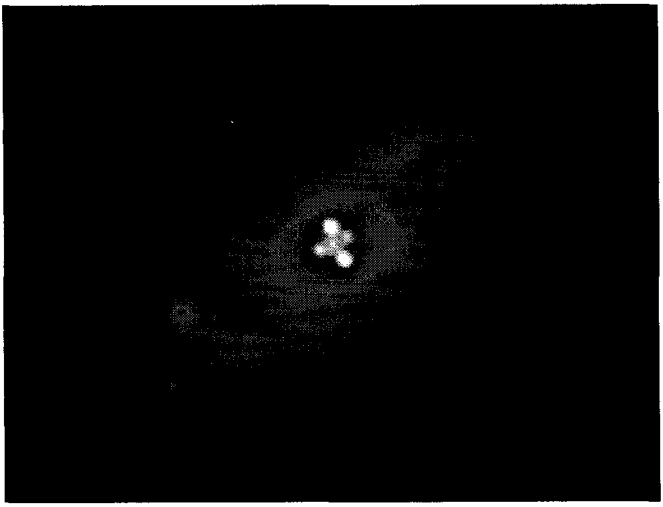


Fig. 10.2 A famous multiple image caused by gravitational lensing: the Einstein cross, as observed by A. Jaunsen and M. Jablonski at the Nordic Optical Telescope on the island of La Palma. The gravity of the galaxy's matter works as giant lens and splits the image of the distant quasar into four separate points. This effect helps astronomers to measure the dark matter in the lensing galaxy.

10.4 MACHOs in the halo of the Milky Way

Dark matter in galaxy halos could be in "Massive Compact Halo Objects", something like planets and dead stars. A clever method to detect such MACHO's in our Milky Way was proposed in 1986 by the Princeton astronomer Bohdan Paczynski. If one monitors the stars of a nearby galaxy, one should see now and then a sudden blazing of a star, when an invisible, moving MACHO crosses the line of sight. Currently, astronomers are keeping a steady watch on 8 million stars in our companion galaxy, the Large Magellanic Cloud, waiting for brightenings. After several years' effort, a dozen lensing events have been detected.

The duration of the brightening depends on the mass of the dark body, on its distance from us, and on its velocity. If the MACHO is one tenth of a solar mass and is moving with the speed of 100 km/sec at a distance of 10 kpc, then the star first grows brighter for about one month, after which it takes another month to fade back to its usual light. This symmetric behavior distinguishes lense events from more mundane variable stars.

Lensing events have very likely revealed MACHOs. What about their masses? The observations point at masses of a few tenths of a solar mass. Because very brief brightenings are hard to detect, this method cannot currently tell about very small bodies, with less than one millionth of a solar mass (something like the Earth). It is still unknown how much of the halo mass is made of MACHOs, but it may be as much as one half.

Using surprisingly small telescopes (diameters from 40 cm to 1 m), though state-of-the-art computer systems, one may have already caught a first glimpse at the mysterious matter component.

10.5 Do Arp's quasars reveal dark matter in galaxy haloes?

In Chapter 7 we discussed the evidence that high-redshift quasars and low-redshift galaxies are associated in the sky more often than one would expect from chance coincidences. But does this mean that there is an anomalous, extra component in cosmological redshifts? Perhaps, but it is also possible that gravitational lenses in the halos of the foreground galaxies brighten the quasars and hence produce apparent associations.

Arp's quasars lie typically at a projected distance of 60 – 100 kpc from the center of their "host" galaxy. This means that if the quasars are very distant objects, their light has passed through the massive halo of a nearby galaxy. If the halo contains suitably grainy dark matter, then Arp's objects might be mirages produced by the gravitational lensing.

However, it has been calculated that the microlensing effect by faint stars in the halo cannot produce the observed number of quasar–galaxy associations. This is because the stars move and cross the line of sight rather quickly: the quasar should fade in half a year, while Arp's objects are still in the sky, shining as before.

One of us (Yu.B.), together with Julia Bukhmastova, has considered another explanation of Arp's effect as gravitational lensing. This would require large clumps of dark matter in the halo, with masses similar to globular star clusters. Microlensing by stars or macrolensing by galaxies cannot produce this phenomenon. Only "mesolensing" by globular clusters and other massive dark matter conglomerations, in galactic haloes, might make itself known in the garments of Arp's effect. Halton Arp's discovery has perhaps opened a new view on the nature of dark matter in the halos of galaxies, complementing the MACHO studies in our own halo.

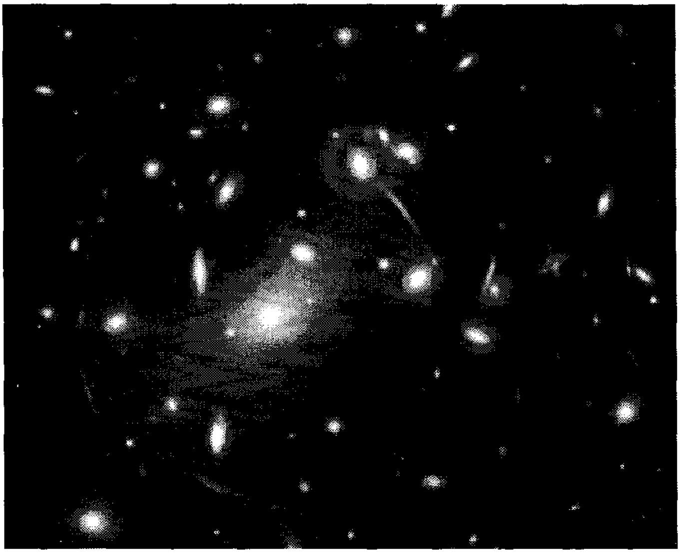


Fig. 10.3 The dark matter in the cluster Abell 2218 works as a gigantic lens, producing arc-like images of very distant galaxies behind the cluster. Hubble Space Telescope images like this one allow astronomers to weigh the invisible, mysterious substance.

10.6 Much more in a cluster of galaxies than the eye sees

After Zwicky's pioneering work, astronomers have studied all kinds of galaxy systems, from small groups to large clusters. As a rule, the virial mass needed to keep them from breaking apart, has been found to be tens of times larger than the mass of all member galaxies. [‡]

There is another way of measuring the total mass of a large galaxy cluster: The cluster mass continuously attracts surrounding galaxies towards its center. This should be seen as a small deviation from the linear Hubble law. Indeed, the galaxies around the close galaxy cluster of Virgo, reveal such a deflection. The mass inferred from this effect is about the same as the virially determined mass of Virgo, again revealing the dark matter.

X-ray satellite observatories, launched into orbit around the Earth, have detected intensive X-rays coming from many galaxy clusters. These are

---

[‡] The method of virial mass is based on a number of assumptions, such as the cluster being stable, neither expanding or contracting. Also, one must distinguish the true members of the cluster from foreground and background galaxies. If one of these assumptions fails, the derived mass may be too large. So one also needs other independent methods.

emitted by the very hot (up to 100 million degrees K) gas filling the cluster. And remarkably, the total mass of the cluster, which keeps this hot "atmosphere" from escaping, is also found to be very large. It is roughly the same as has been derived using the virial theorem. The mass of the hot gas itself, even together with galaxies, is not large enough to make up the total cluster mass.

Stars and galaxies deflect light, and so do galaxy clusters which act like giant gravitational lenses. Large optical telescopes have revealed strange arcs around clusters. The sizes of these rings may reach a few tens of arcseconds. Such large mirages can form only if the mass of the cluster again is similar to what X-ray observations and the virial theorem suggest.

10.7 The total amount of dark matter in the universe

For big bang cosmology, the dark matter is a part of life. *First*, the inflation model insists that the dark matter makes 99 percent of the total mass of the universe. *Second*, the theory of primordial nucleosynthesis demands that ordinary matter, such as atoms and molecules, can form only a few percent of the total mass. The overwhelming majority of the dark mass must be in some unknown form. Hence the unexpected question: Is there enough dark matter and not too much ordinary matter?

According to the inflation model, the cosmic mass density is equal to the critical value, i.e. the density parameter $\Omega = 1$. The observed density of luminous matter (stars in galaxies) is much less, and one has to postulate that there ia a lot of dark matter. But its constitution cannot be arbitrary. Ordinary, baryonic matter can form only a small part of it.

Visible matter constitutes only 0.4 percent of the critical density:

$$\Omega_{lum} \approx 0.004 \text{ \textit{(luminous matter)}}$$

This value depends on the Hubble constant, on the mass-to-luminosity ratio for galaxies, and also on the volume where the galaxies are counted (due to the fractality; Ch.17). Though there is strong evidence for dark mass around galaxies and in clusters, it is still not clear from observations whether there is enough dark matter to fill the gap between $\Omega_{lum} \approx 0.004$ and $\Omega = 1$.

One can calculate theoretically how much of the light elements (Helium, Deuterium, Lithium) were produced from Hydrogen during the first minutes of the universe. The abundance of Helium is especially well known from observations. Its mass fraction in gaseous nebulae and stars in which nuclear reactions have not changed its primordial value converge to the value of 23 per cent. This observed amount of Helium implies very little normal matter in the form of *baryons*, i.e. protons and neutrons, the building blocks for chemical elements: [§]

$$\Omega_{bar} \approx 0.06 \textit{ (big bang prediction)}$$

Hence, if the inflation model is correct and $\Omega = 1$, only 6 per cent of the cosmic mass can be in the form of ordinary matter such as stars, gas, and dust. The remaining 94 percent must be some exotic form of *nonbaryonic matter* which has never been detected in the laboratory. Known nonbaryonic matter, such as electrons and massless neutrinos, can contribute only a very small part of the needed dark matter.

In any case, if one assigns to all clusters of galaxies the high masses derived from X-ray studies and gravitational lensing, this alone makes the density parameter of the Friedmann model to be around 0.2, i.e. a few times larger than the predicted density of baryonic matter. Also various dynamical estimations of the density of dark matter converge to around

$$\Omega_{dark} \approx 0.3 \textit{ (dark matter)}$$

Hence there *must* be some kind of nonbaryonic matter, if one accepts the big bang model. The true nature of this "extra" dark matter is a crucial test for the big bang – it cannot be any ordinary stuff.

10.8 An ocean of massive neutrinos?

The big bang predicts that *neutrinos* should fill space as relics from the first second of the universe – they should be about as numerous as the photons of the cosmic radiation ($\sim 1000/\mathrm{cm}^3$) outnumbering electrons and protons by hundreds of millions to one. But standard elementary particle theory tells that neutrinos are massless like photons. However, it is hard to measure whether indeed their mass is zero. Neutrinos interact with other matter very weakly, they just zoom around almost unhindered, at, or nearly

---

[§] The value of the baryonic density parameter predicted by big bang nucleosynthesis depends on the Hubble constant: $\Omega_{bar} = 0.02(100/H_0)^2$, $H_0 = 60$ km sec^{-1}/Mpc.

at the speed of light. The neutrinos, if massive, would thus form a medium called "hot dark matter", because they move so fast.

The tiny or zero neutrino mass is a big question for the particle physics and the problem of dark matter. Their huge number means that even if the mass is, say, one millionth of the mass of electron, the total mass of neutrinos would be several times that of the visible matter. In fact, in the 1980's the first experiments on the neutrino mass hinted at such large values that the cosmic mass density due to neutrinos could equal the critical density. ¶ However, this result was not confirmed, and now new experiments show that the neutrino mass is less than 2.5 eV.

Recently, new evidence for a non-zero neutrino mass has been gathered at the Super-Kamiokande Laboratory in Japan. One detects neutrinos from the interior of the Sun, produced in nuclear reactions. It should be noted that the cosmological, primordial neutrinos, if they exist, have such small energies (because of redshift) that they cannot yet be detected with our neutrino telescopes. So one does not yet have direct evidence on this prediction of the big bang.

The heart of the neutrino observatory is a tank of 50 000 tons of water, sited in an old zinc mine. Neutrinos come in three kinds: physicists speak about electron- , muon- , and tau-neutrinos. When electron- or muon-neutrinos, very rarely, collide with a water molecule, the resulting faint flash is recorded by sensitive photomultiplier detectors covering the roof and walls. Analyzing the recordings, the team of 120 Japanese and American physicists concluded that too few neutrinos are arriving from below, through the Earth. This asymmetry they ascribe to the different path lengths traveled by the neutrinos created in cosmic-ray collisions in our atmosphere before reaching the water tank.

It is theorized that each neutrino is a mixture of three "mass states", which determines whether it is of electron- , muon-, or tau-type. The mixture is not stable, but can change, leading to another neutrino type. Those neutrinos which go the long way through the Earth have time to change their type, which explains the asymmetrical arrivals of detected muon-neutrinos. Physicists are rather convinced that the Super-Kamiokande results indicate a non-zero neutrino mass, perhaps for all types. The mass

---

¶ In order to estimate the contribution of the neutrino mass to the cosmological density parameter Ω, one may use the formula: $\Omega_{neutrino} = (m_e + m_\mu + m_\tau)/30eV$ where the masses of three neutrino-types are expressed in eV and $H_0 = 60$ km sec^{-1}/Mpc.

Fig. 10.4 Two madams from Lyon gossip about hot news on the neutrino mass (eavesdropped by Georges Paturel).

could be as "large" as five millionths of the electron mass, or equal to 0.1 eV. If so, then the primordial neutrinos could carry as much mass as all the stars in the universe, i.e. $\Omega_\nu \approx \Omega_{lum}$. Thus the neutrinos as a candidate for hot dark matter still give only a small fraction of all dark matter. In fact, the latest news tell that an experiment in the Sudbury Neutrino Observatory in Canada has confirmed that the sum of the masses of the three neutrino types lies in the range from 0.1 to 8 eV.

10.9 The search for dark matter goes on

It is not yet known what the physical carriers of dark matter are, apart from the fact that they interact gravitationally with visible matter. In the search for dark matter candidates, it is good to keep in mind that they can be divided neatly into two classes: ordinary baryonic matter such as in planets and stars and exotic non-baryonic matter, such as the neutrino relics from the early universe.

As to the composition of baryonic dark matter, astronomers follow the strategy, reminiscient of what Sherlock Holmes recommended, that they first determine which types of objects *cannot* be good candidates. One can exclude ordinary stars and smoothly distributed gas and dust, since these would be visible with existing optical or radio telescopes. Very faint stars, such as white dwarfs and neutron stars, or some as yet undreamt of dark star, are still quite difficult to exlude. And various conglomerations in galaxy haloes, something like comets, asteroids, planets, remnants of dead stars, and very cold molecular clouds still may have slipped unnoticed through the astronomer's web.

The Super-Kamiokande discovery that the mass of the neutrino is about 0.1 eV opens a new era of experimental study of nonbaryonic dark matter. Neutrinos are "hot", while the standard model of galaxy formation assumes that it mainly consists of "cold" particles, capable of taking part in gravitational clustering. The theories of elementary particles offer a plethora of candidates for such substance. Two are very promising and can be searched for by high energy experiments. These particles are the axion and the neutralino. The mass of the axion is expected to be 10^{-6} eV to 10^{-4} eV, while the neutralino is much heavier, 50 GeV to 500 GeV. The Lawrence Livermore National Laboratory in the USA is now testing the preferred axion mass range. The DAMA experiment in Cran Sasso and the CDMS experiment in the Stanford underground facility are probing the mass range in which neutralinos might be detected.

* * *

The observations show that dark matter exists in abundance. But it is still uncertain how it is scattered in space – it appears to be lumpy and to follow galaxies. And the more dark matter candidates we have, the less we know about its real constitution! As observations do not give the chemical composition of the bulk 90 percent of the predicted baryonic matter, this means that the abundances of the light elements is no longer such a pillar of the big bang as it used to be. And thinking about all that non-baryonic matter, it may come as a shock to realize that we do not know what 99 percent of the universe is made of. But one thing is certain. It is no longer possible to understand the universe without knowing its dark side.

Chapter 11

Dark energy – the new emperor

Recent measurements of the expansion of the universe have brought to light another odd substance, even stranger than dark matter. It has negative pressure, its energy is larger than that of dark matter, and it manages the dynamics of the whole universe. Because of its unusual physics it is called "dark energy" to distinguish it from dark matter, which is also mysterious, but more familiar with its gravitational effects. It has long been thought that the gravity of matter, both dark and luminous, is slowing down the universal expansion. So it came like a bolt from the blue, when astronomers studying distant supernovae inferred that the expansion is *accelerating*. What is speeding up the universe? Theorists cannot find other way out than to ascribe the acceleration to the "antigravity" of dark energy, also called quintessence. Dark matter and dark energy – these mysterious components of the universe seem to be deciding our fate.

* * *

11.1 Revolution in cosmology – Einstein's lambda returns!

The January 1999 issue of Scientific American opened the new year with a special report: "Revolution in Cosmology". This echoed a surprising result by two groups of astronomers who for some years have studied very distant supernova stars. Their aim was to measure the expansion rate for a good fraction of the visible universe using so called supernova Type Ia. Such a supernova is thought to occur, when a white dwarf with a mass of about 0.6 solar mass explodes as a result of mass overflow from its companion star. The two famous supernovae in our Milky Way, which celebrated the revolutionary times for cosmology in 1572 and 1604 were probably of Type Ia. The first one was carefully observed by Tycho Brahe * and was also seen by Galileo as a small boy. The second, "Kepler's nova", much impressed Galileo – the heavens are not inalterable, after all!

The maximum luminosities of these explosions are almost the same from one supernova to another, and may be used as good standard candles. Typically they become as bright as a whole galaxy – rather an achievement from a star less massive than the Sun!

Using large telescopes on the Earth and the Hubble Space Telescope, astronomers have been able to detect and measure such supernovae at large distances, up to the gates of the deep universe, at redshifts around 1. Different Friedmann models predict different relations between magnitude and redshift. The curve for the standard inflation model had been popular for two decades, though there was no direct observational evidence for its superiority. When the new observations were plotted on the Hubble diagram, it came as a small shock that the supernovae did not follow the prediction of the inflation model (they were fainter at large redshifts). To explain the observed Hubble diagram, one had to put Einstein's cosmological constant back into the Friedmann models. We show in Fig.11.1 results by the High-Z Supernova Search Team headed by Brian Schmidt. Another important team is the Supernova Cosmology Project headed by Saul Perlmutter.

The hopes to prove the inflation model seemed to be crushed. Inflation predicted that the critical density is wholly due to matter. The cosmologist Lawrence Krauss wrote at the time: "One thing is already certain. The standard cosmology of the 1980's, postulating a flat universe dominated by matter, is dead." The notorious cosmological constant, Einstein's lambda-term, was now warmly welcomed like a prodigal son, and started to play a central role in the interpretations of the things seen in the universe.

One important consequence of Einstein's gravity equations is that the positive lambda-term corresponds to a repulsive force – "antigravity". Hence the expansion of the universe is speeding up, even though the gravity of matter tries to resist.

---

*The appearance of this very bright star (brighter than any other star or planet) in the constellation Cassiopeia was decisive for Tycho's career, it "put him on the path he was to follow for the rest of his life", as Victor Thoren wrote in his Tycho biography. By the way, "Kepler's nova" appeared in the constellation Ophiuchus.

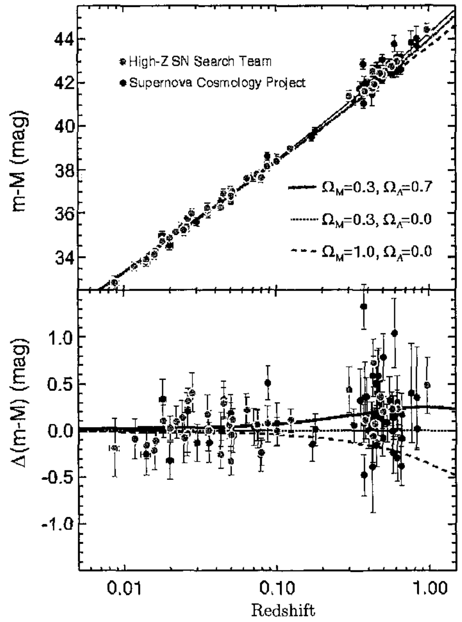


Fig. 11.1 The redshift–magnitude diagram for distant supernovae of Type Ia reveals the acceleration of the universe. The dashed curve is the expected relation between the magnitude and redshift in the "old" inflation model where the lambda term is zero. The observed supernovae are fainter and are clustered around the (thick) curve corresponding to the model where the dark energy makes 70 percent of all mass in the universe. These diagrams were made by the High-Z Supernova Search Team.

The discovery of the universal acceleration required a remarkable collaboration of astronomers working with space and ground-based telescopes. This effort and its result, a modest deflection of the data points in the Hubble diagram, reminds us of Kepler's momentous discovery of the elliptic orbit of Mars from the data produced by Tycho Brahe's project. Consider also the precession of the vernal equinox in the sky, discovered by Hipparchus in the second century B.C. (Ch.2). When visiting Alexandria from his native Rhodes, he learned of old observations of stars, made one and half centuries earlier. When he compared his own measurements of stellar positions with those old ones, the precession of equinox appeared (later explained as reflecting a wobbling of the Earth's axis). To detect such a slowly advancing effect needs long term and accurate observations. Not all important and unexpected phenomena are very spectacular!

11.2 A short course in the physics of "nothing"

The lambda-term in Einstein's equations may be interpreted as the contribution of the cosmological vacuum to the cosmic density. If one wants to save the inflation idea of zero curvature, then the *sum* of the matter and vacuum densities should be equal to the critical density. In this case the supernova observations can be explained if the vacuum density is about two times the matter density. But then the true ruler is no longer the dark matter, but another mysterious stuff, the cosmological vacuum!

The cosmological vacuum, as it is pictured in the theory, is a very unusual kind of "matter". It has positive energy but negative pressure. In Friedmann's equation the cosmological vacuum appears as the sum of a positive energy density and three times a negative pressure, making in the end a negative *gravitational* mass density for the vacuum. A negative gravity mass produces a repulsive force, or "antigravity". Between any two galaxies in expanding space the negative vacuum mass is increasing with time, which leads to an increasingly high acceleration of the universe. †

In modern physics the vacuum is not just empty space, but an ocean of continuously created and annihilated particles. Another strange aspect of the vacuum is the zero-point energy of all the quantum fields existing in the universe. This gives rise to a major puzzle. Why is the density of the cosmological vacuum found to be so small, close to the critical density, about 10^{-29} g/cm^3? A direct calculation of the vacuum density from theoretical physics leads to an incredibly huge value of about 10^{+94} g/cm^3! This value is also called Planck density, and it may be expressed via the three fundamental constants G, h, and c ($\varrho_{Planck} \approx c^5/G^2h$). Recall that we already encountered the Planck length in the early universe (Ch.9). The Planck time, length, and density should naturally appear in a relativistic quantum gravity theory. As there is no complete "Ghc-theory", the vacuum is still a fundamental problem of modern physics. ‡

---

† The Friedmann model is described by the following exact equation of motion: $\ddot{r} = -GM_{\text{eff}}/r^2$, where M_{eff} is the gravitating mass of the matter and the Λ substance inside the radius r. Due to homogeneity, $M_{\text{eff}} = \frac{4\pi}{3}(\rho + 3p/c^2)r^3$ where $\rho = \rho_m + \rho_\Lambda$ and $p = p_m + p_\Lambda$ are the total density and pressure. The effective mass is negative if $-3p > \rho c^2$. In terms of the density parameter, the total density of the universe is the sum of the Λ and matter densities: $\Omega = \Omega_\Lambda + \Omega_m$.

‡ The name "*Ghc*-theory" has its roots in an early article published in 1927 by the young Russian physicists George Gamow, Dimitrij Ivanenko and Lev Landau, who classified physical theories according to the fundamental constants they contain. Among

Why then are the physicists in their laboratories not so worried about such a huge vacuum density? This is because in the physics of elementary particles one usually can ignore gravity, which in the microcosm is in any case small. The Planck density is regarded as a zero-level, like the calm surface of the sea, relative to which energies are measured.

11.3 Dark energy, quintessence, spintessence...

The theory of elementary particles predicts that not only the vacuum, but also certain quantum fields can be described as substances with positive energy densities and negative pressure. Hence these theoretical entities may be considered as candidates for the role of dark energy. Hypothetical scalar fields with such properties are called quintessence. (A 'scalar field' describes a material as an array of numerical values at various points in space. Less exotic than cosmology's antigravitating scalar field would be a temperature scalar field giving the temperature at every spot in, say, a crowded concert-hall.)

Einstein's equations for uniform dark energy, having a simple equation of state between pressure and density, give an amazing, but strict result: the acceleration of the universe is described by the exact Newtonian equation for the gravity force, in which the mass inside a sphere around any point is negative! [§] A negative gravitating mass means that instead of deceleration, as caused by ordinary matter, quintessence produces acceleration. This strange behavior, the antigravitation of dark energy, is a new invention of cosmological physics and has not yet been tested in the laboratory.

When Pandora's Box is opened, weird things start to pop up in physics. Recently a new hypothetical entity – spintessence – was proposed as a "top candidate" for dark energy. This spintessence is not a real (in the mathematical sense...) but a *complex* scalar field.

the physicists in Leningrad they were known as "three musketeers". Later Matvei Bronstein (1906-1938) joined this joyful team. In 1936 he published pioneering work which suggested gravitons as the particles mediating the force of gravity. His short life ended two years later as a victim of Stalin's terror.

[§] $M_{DE}(r)$ is the gravitating mass of dark energy within the radius r and with the equation of state $p_Q = w\epsilon_Q$, $w \in [-1, 0)$: $M_{DE} = \frac{4\pi}{3}(1 + 3w)\rho_Q r^3$. This is negative for $w < -1/3$. The parameter w is a new parameter of physics and cosmology (along with H, Ω, Λ). For the vacuum $w = -1$, for the "dust" $w = 0$ and for the photon gas $w = 1/3$. The shell of a gravastar (Ch.9) is made of a super stiff substance with $w = 1$.

11.4 A bit of history: redshift and de Sitter's effect

Willem de Sitter 1872-1934

In 1917 Albert Einstein derived from his equations of general relativity a cosmological solution: he found the spherical static universe, finite but with no edge. Quite soon the Dutch astronomer Willem de Sitter published another solution, resulting in a world no less strange. De Sitter considered what would happen if the universe contained so little mass that its density could be taken to be zero as a first approximation (this is typical in science – you start with a simplification in order to be able to predict at least something). He added the cosmological constant to Einstein's equations, or the vacuum as we can now say, as Einstein had done with his universe. When he inspected what kind of phenomena could occur in his empty universe, de Sitter found a surprising thing. The light received by observers would have a spectral line shift towards the red, and the redshift would be larger for light coming from more distant regions.

De Sitter's redshift phenomenon is not caused by the Doppler effect of stars moving away. It is a property of space-time, which appears when these are forced into the bitter conditions of the empty universe with the lambda-term. Remarkably, de Sitter was the first to suggest, in 1917 when galaxies were not yet known, that one should try to find a redshift–distance relation for very remote celestial bodies. This prediction was made for a static universe. Even Friedmann, who five years later demonstrated the possibility of an expanding universe, failed to point out the redshift phenomenon as a property of his own model. ¶

In his discovery paper of 1929, Edwin Hubble concluded that he had possibly found de Sitter's effect (i.e. the influence of the cosmological constant). It is also striking to read Arthur Eddington's *The Expanding Universe*, published in 1933. This nice popular book by a pioneer of modern astrophysics is permeated with the idea of the importance of the cosmological constant. Eddington was excited by the freshly discovered Hubble law and thought that the expansion of the universe was driven by the cosmical repulsion (he did not like the alternative explanation by Lemaître who proposed the "fireworks theory", a forerunner of the big bang, in which the expansion started with a violent projection from a primordial atom).

¶ Dimitrij Ivanenko has related to Arthur Chernin that Friedmann, at a seminar in 1924, discussed the high redshifts as discovered by Vesto Slipher, and considered these to be direct evidence for an expanding universe.

With the recent observations of distant supernovae, a kind of de Sitter's effect seems to have come back after 80 years!

11.5 The age of an accelerating universe

The first estimates of the Hubble constant seven decades ago brought about an odd thing: the universe seemed to be younger than the Earth! Indeed, when a cosmological model tells us that the universe has a finite, measurable age, the model has to live under a constant threat. The age problem has followed the big bang model all its life, more recently with the oldest globular clusters. And a new aspect appeared when observers started to find very distant galaxies.

The age of the universe with a critical density and made of matter, 11 billion years (Ch.8 and the Table below), is by a wide margin less than the age of the globular clusters. This apparent discordance between the ages of globular clusters and the model universe has been called the *age paradox,* a situation as perplexing as if a father is younger than his son...

But a universe accelerated by the cosmological constant can be older than the ordinary inflationary universe. This is easy to understand: the accelerating universe expanded in the old times more slowly than now, hence required more time to reach its present size.

We give a table of ages as in Chapter 8, but now also for an accelerating universe where the cosmological constant dominates (here $\Omega_{matter} = 0.3$, $\Omega_{Lambda} = 0.7$, $H = 60$). The age of such a universe is 16 billion years, and it conveniently accommodates the oldest globular clusters.

Z	T_{back}	T_{age}	T_{back}^{accel}	T_{age}^{accel}
0	0	11000	0	16000
1	7100	3900	9200	6800
3	9600	1400	13500	2500
5	10250	750	14600	1400
7	10490	510	15100	900
10	10700	300	15450	550

Moreover, the age problem with the globular clusters is for the present epoch: the paradox may worsen as one goes back in time toward the big bang, for example by observing very distant galaxies. From the table of ages in the *non-accelerating* inflation universe, one sees that at $Z = 5$ the universe was "only" 750 million years old. Observations have revealed galaxies and quasars at such high redshifts. One starts to wonder whether there was enough time for their formation so soon after the big bang. [‖]

A good world model should yield enough place and time for the galaxies to form and mature, and avoid the riddle of B^4 (BBBB = "Born Before the Big Bang"). When galaxies are found at higher and higher redshifts, this type of age problem becomes increasingly interesting. Presently, the record redshift is 6.7 which corresponds to a time of 550 million years after the big bang (or about 1000 million years in an accelerating universe).

The Keck 10m telescope at Hawaii revealed a red radio galaxy 53W091 which at redshift = 1.55 is more than 3.5 billion years old. The age of the non-accelerating model at this redshift is only 2.7 billion years. This situation illustrates why it is so important to observe distant galaxies. They give us a glimpse of the end of the Dark Age when galaxies were formed, and in so doing give us evidence for the cosmological constant. The accelerating universe offers more time for the galaxies to form and mature.

11.6 The fifth element may rule in your backyard

As the effect of the cosmological vacuum was detected from observations of very distant supernovae (with redshifts close to 1), one might think that it is of no concern whatsoever in our local galaxy neighborhood. And what could its effect be? The vacuum has negative gravitational mass, and its antigravity tends to compensate the attractive gravity between masses floating in the vacuum. At a certain distance from a mass the compensation is complete (gravity = antigravity). At smaller distances gravity is superior, while at larger distances the vacuum repulsive force gains the lead.

It is not hard to calculate for our Local Group of galaxies where the border between the gravity and vacuum dominated regions lies, if the vacuum density is about twice the matter density. The border turns out to be quite close, only at a distance of $1 - 2$ Mpc, at the outskirts of the

---

[‖] Recently Yurij Parijskij from Russia has pointed out that there are indications of an age problem for distant radio galaxies which on the basis of their red colors contain stars which appear to be much older than 750 million years.

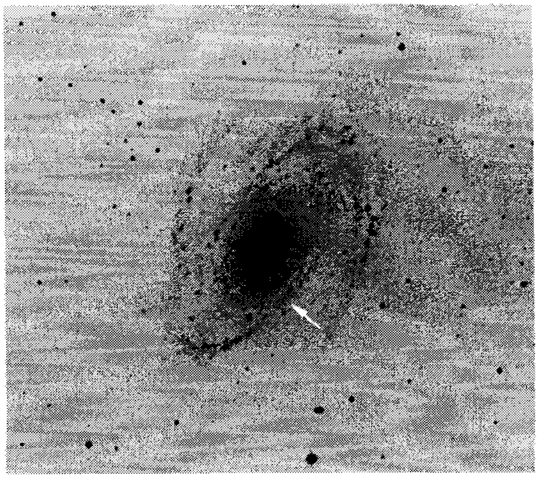


Fig. 11.2 In 1993 a supernova exploded (the site shown by the arrow) in a spiral arm of the nearby galaxy M81 in the constellation of Ursa Major. Studies of much more distant supernovae have led astronomers to conclude that the universal expansion is being sped up by some antigravitating substance. At the distance of M81 (3 Mpc) the antigravity of the space between it and the Milky Way exceeds the normal gravity between them.

Local Group! ** Inside our small galaxy swarm the gravitation keeps all together, but when we look at slightly more remote galaxies, it is strange to think that between us and "them" there is a repulsive force, arising from the "nothingness" of the vacuum!

But new discoveries in cosmology have the habit of bringing with them uninvited visitors in the guise of disturbing questions. We already mentioned that the vacuum density accelerating the universe is incredibly smaller (by the factor 10^{120} !) than the quantum theory would predict. Another riddle is why the vacuum density is rather close to the matter density in the present epoch of the universe? If the vacuum density is really constant, then in the dense past it was much smaller than the matter density. Why the coincidence now? Or perhaps it is not a coincidence? Such questions have produced some brain-storms among cosmologists.

A few of them seriously ponder the possibility that the "vacuum" has its

---

**Arthur Chernin was the first to point out, in discussions with the authors, that the border between gravity and vacuum lies so surprisingly nearby. This follows from the exact equation of motion: $\ddot{r} = -GM_{\text{eff}}/r^2$, where $M_{\text{eff}} = M_m(r) + M_\Lambda(r)$ is the sum of the gravitating masses of the matter $M_m(r)$ and the vacuum $M_\Lambda(r)$ inside the radius r, $M_m(r) = 4\pi \int_0^r \rho_m(r) r^2 dr$ is the matter mass within the radius r.

present density because it never was very far away from the matter density. But then the universe is not accelerated by the genuine vacuum, Eintein's lambda-term, but some time-evolving energy with negative pressure. Others have even suggested that the coincidence may be a case of the *anthropic principle* which states that a human being, as he/she is, can exist only in the universe as it is! If the universe were very different from what it is, then perhaps it would not offer the physical conditions (e.g. sufficient time) for life to emerge, and we would not be here wondering about all that! This intriguing possibility is much discussed also in other cosmological contexts, but is still far from being a physical explanation.

Such a novel form of energy is often called dark energy, being if possible still stranger than the other unknown stuff, dark matter. As "a dear baby has many names" (says a Finnish proverb), the theoreticians also like to speak about the quintessence or the fifth element. We recall how Plato put order in the microcosm, relating four regular solids to the four traditional elements, while the remaining fifth, the dodecahedron, was ascribed to the heavens. Aristotle called the fifth element the "aether" while in the Middle Ages the Latin "quintessence" denoted the sublime perfect substance. It is amusing that this classical term appears now when a new substance is needed to make the large scale geometry of the universe Euclidean!

* * *

It is a nice side of dark energy that it is expected to be more smoothly dispersed than galaxies. In particular the vacuum may be the truly uniform substance, the regular visitor in the dreams of cosmologists. The quantum physics of the vacuum and gravity, still to be developed, has become crucially important for the understanding of the universe.

Though the value of the energy density of the vacuum is still poorly understood in cosmological physics, the observations make one agree with the old words of Georges Lemaître, who was the first to realize that the cosmological constant in Einstein's equations represents the vacuum with negative pressure and thus cosmological repulsion: "Everything happens as though the energy *in vacuo* were different from zero."

Chapter 12

Expansion and curvature of space

The expansion and cooling of the universe are the vital processes in the big bang cosmos. The expansion of space looks familiar when compared with an inflating balloon where everyone can see how any two points on the rubber membrane are separating. However, this process hides its share of riddles. Galaxies are rushing away from each other, at velocities greater than the speed of light, but do this without actual motion! The cosmological redshift in Friedmann space is not caused by the Doppler effect, and gas and radiation cool during the expansion without performing work!

The enigmas of expansion, curvature and energy loss shed light on the old debate on the relation between physics and geometry. What happens with the meter stick in expanding space, is a subtle question.

Remembering the surprising microcosm with its quantum laws, one may expect that the macrocosm on its largest scales offers things as difficult to picture in terms of ordinary physics. Some of the queer things in the classical big bang model may remind one of the situation a century ago when classical physics was coming to an end.

* * *

12.1 The nature of redshift – Allan Sandage's 15th problem

Allan Sandage, who was born three years before his early mentor Edwin Hubble found the redshift law, has his whole life devoted to cosmology. When in 1991 he was awarded the prestigious Crafoord Prize, Sandage emphasized that cosmology is an experimental science which deals with testable propositions. * There should also be a way to test whether the cosmological redshift is caused by expansion or something else.

In big bang cosmology the redshift is caused by space expansion. When you measure a redshift, you also measure the expansion. But it is logically possible that the redshift could have another cause in some other cosmological model. Hence, one needs to verify the expansion.

Hubble emphasized in his discovery article that what was measured was "displacements of the spectra". Later he envisioned it as an important task for the large telescopes to decide what mechanism actually causes the cosmological redshift. In 1995, at the conference on *Key Problems in Astronomy and Astrophysics* held at the Canary Islands, Allan Sandage presented a list of 23 astronomy problems for the next three decades, in a form analogous to Hilbert's famous 23 problems in mathematics. The first problem in cosmology was: *Is the expansion real?*

Surprising as it may sound, it is not easy to prove in some direct way that redshifts are due to the expansion of space rather than some "tired-light" effect. Tired light is a common name for certain theoretical phenomena proposed for explaining the redshift, instead of expansion. Suppose that a photon gradually loses its energy when it traverses space. As a photon's energy is inversely proportional to the wavelength, a loss of energy increases the wavelength (redshift).

The theory of tired light is presently just an exotic possibility, whereas there are three known phenomena which really produce redshifts. These are ordinary motion and gravitation, both experimentally verified, and space expansion:

- *motion in space – the Doppler effect*
- *gravitation – the Einstein effect*
- *expansion of space – the Lemaître effect*

Consider a *Gedanken experiment.* Suppose we can attach one end of a strong but light cable to a distant galaxy, while in the Milky Way it can roll out freely from a gigantic wheel. One can measure two things: the redshift

---

*The Anna-Greta and Holger Crafoord Fund was established in 1980 with the purpose of promoting basic scientific research in mathematics and astronomy, the geosciences, the biosciences and rheumatoid arthritis in Sweden and other parts of the world. The fund is held in trust by the Royal Swedish Academy of Sciences which annually awards prizes and grants from it. One may regard the Crafoord Prize as complementing the Nobel Prize which was not intended for mathematicians or astronomers.

of the galaxy and the velocity with which the cable leaves the wheel. Both space expansion and ordinary motion predict that the wheel will rotate. Simply by comparing these two measurements with the two formulae (the Lemaître effect and the Doppler effect), we can decide which mechanism is correct. The wheel might also behave in an unexpected manner, it might not rotate at all. This could happen if gravitation or some tired-light effect were the cause of the redshift.

Richard Tolman 1881-1948

A more practical, though less direct test for the nature of the cosmological redshift was proposed by Richard Tolman and Edwin Hubble in the 1930's. In a static Euclidean universe the surface brightness of a source does not depend on its distance from the observer, while in the expanding Friedmann universe the surface is much fainter at large redshifts Z:

- *classical space without redshift*: surface brightness is constant
- *tired light effect*: surface brightness decreases as $(1 + Z)^{-1}$
- *expanding space*: surface brightness decreases as $(1 + Z)^{-4}$

The tired light effect causes a decrease in the surface brightness, because each photon loses energy by the redshift factor $1 + Z$. In expanding space, two extra effects cause a still quicker drop in the brightness.

Allan Sandage and Lori Lubin made the surface brightness test in 2001 using high-redshift elliptical galaxies measured by the Hubble Space Telescope. They could extend the test to redshift 0.92, so that if the expansion is real, one should see a strong decreasing trend in the surface brightness. The result was rather close to that expected in an expanding space, while the tired light effect has serious difficulties. †

This remarkable test probed deep space. But the Hubble law starts already close to the Local Group of galaxies, but here the surface brightness dimming would be unobservably small. So to which galaxy should we attach the cable?! Fortunately, there are realistic ways to test the expansion even at such small distances, such as the linearity of the local Hubble law (Chapter 18).

As the Doppler effect also passes the surface brightness test, one may still wonder if the recession of galaxies could be described as a motion within space, avoiding all that talk about "expanding space". And one

---

†The same $(1 + z)^{-4}$ behavior is predicted also for the Doppler redshifts of receding objects and for gravitational redshifts caused by mass distributions. Hence, strictly speaking the test cannot separate space expansion, gravitation, and Doppler redshifts.

may ask if Newton could have foreseen an expanding universe. Assume that in his infinite world the substratum of stars is expanding relative to absolute space, so that we see around us the Hubble law. The problem is that the universal expansion of the stellar content defines a privileged center in absolute space. (Perhaps this is why Newton did not consider the expansion or contraction of the whole universe, for which he has been critisized by modern commentators.) So the observed expansion appears to tell us rather loudly that static absolute space has dissolved long ago, which we of course already knew from relativity theory. Though Edward Milne attempted to show in the 1930's in his kinematical cosmology that global expansion of matter within the space of special relativity may exist without preferred points, modern cosmology views space itself as expanding.

12.2 Understanding the expansion of space

Speaking about space expansion, one would like to know what space itself is. Is it just a relation between bodies or perhaps a kind of substance? This fundamental open issue in physics was discussed by Henri Poincaré who analyzed in his books *Science and Hypothesis* and *Science and Method* the relation between geometry and experimental physics. Experiments are concerned with bodies, not with space. One cannot move a piece of space and compare it with another bit of space. One can only move and compare one body with another. The size of a body is always relative to a material meter stick, the unit.

To illustrate space relativity, Poincaré gives a variant of Kafka's *Metamorphosis*: Suppose that during the night, all sizes of the universe were increased one thousand times. When I open my eyes in the morning, what feelings would I have? Surprisingly, I would notice nothing. My bed had increased by the same factor as my body, and in general, every size measurement would give exactly the same result as before: The meter stick had increased by the same amount. The lesson is that one can measure an expansion only if there is a reference unit which does not change. [‡]

As the cosmological redshift suggests that we can be aware of space expansion, what then doesn't change in our world? Eddington wrote in

---

[‡] Strictly speaking Poincaré's example should be complemented by a statement about the exact behavior of other physical quantities, such as mass, electric charge and velocity of light. For the expansion to remain undetectable, such constants should change suitably.

The Expanding Universe that such unchanging things exist as an ordinary meter stick, the Earth, and our Solar System. "...only the intergalactic distances expand. The galaxies themselves are unaffected; and all lesser systems – star clusters, stars, human observers and their apparatus, atoms – are entirely free from expansion." Later, to this list have been added galaxy clusters. The existence of such a ladder of rigid objects is usually ascribed to binding forces due to electricity or gravitation.

We have learnt to look at space expansion through the familiar analogue of the expanding balloon. Dots drawn on the balloon recede from each other following the Hubble law, and the surface (space) increases during expansion. One may also speak about "swelling space", having in mind the rising pudding analogue.

However, if one desires to understand space expansion more deeply, such analogues fall short. Swelling space may create the misunderstanding that space itself is somehow streaming – one cannot move a piece of space as Poincaré said. And in the rubber balloon the growth of surface happens at the cost of the thinning membrane. But there is no such store of extra space inside space itself: to say that space expands is close to saying that space is created. To the space within the Hubble radius a volume like that of our Local Group of galaxies is added every second. The continuous birth of space is almost a metaphor and the term "creation of space" is sometimes regarded as a bad way of speaking about space expansion. To our thinking, it helps one to understand some important features of expanding space.

Consider the good old balloon, now with pennies affixed to it. The stiff coins pinpoint the places in the universe where there is no expansion of space. Between the pennies in heaven, in the realm of the Hubble law, there is the physical phenomenon of increasing volume, no matter which word, 'expansion' or 'creation', is used. If the distance between two galaxies increases, but the galaxies do not move *inside* space, then a natural way to understand this is that space emerges in the region between them.

The metaphor of space creation becomes more substantial, when one realizes that, according to the standard theory, space appears together with a real substance, the vacuum, a boiling ocean of particles and antiparticles. This would, by the way, a little warm up the relations between the big bang and its old rival the steady-state cosmology. Now a continuous creation process would exist in both models. In the steady-state the matter is born together with space, keeping the matter density constant, while in the big bang the vacuum remains the same.

In his *The creation of the universe* George Gamow explained how an infinite physical space may expand and offer more and more space to the galaxies. Infinite space in itself is an endless store of space (and energy, we might add), similarly as in "Hilbert's hotel" with its infinite number of rooms, even if all reserved, in which the hospitable host can always find rooms for new guests. It is tempting to play with the idea that expansion in some deep manner implies infinity, even though mathematical universes may perfectly well be finite. In any case, it is interesting that our expanding world, or the model best describing it, indeed appears to be infinite.

12.3 The Lemaître phenomenon versus the Doppler effect

In the classical paper of 1927, where he first introduced the cosmological redshift, Georges Lemaître was careful to emphasize that the cause for the redshift is the increasing radius of the universe. He called it an *apparent* Doppler effect, because an ordinary motion of the source would imitate a similar shift in the spectrum. What is the crucial difference between Lemaître's space expansion redshift and the Doppler effect caused by motion? In the case of space expansion the wavelength of a photon is continuously stretching all along its path from the light source to the observer. But when a receding body in static space emits a photon, its wavelength does not change along the path and has a constant shift determined at the moment of emission.

The space expansion redshift differs from the Doppler mechanism, and it is not quite correct to explain the cosmological redshift as a Doppler effect: "We can determine the present rate of expansion by measuring the velocities at which other galaxies are moving away from us, using the Doppler effect", or "The absorption lines are progressively redshifted as the distance to the [galaxy] cluster increases. This redshift is due to the Doppler effect."

Edward Harrison has pointed out clearly that even the mathematical expressions for these two mechanisms essentially differ, as they should if the phenomena are different. [§] If one measures the redshift of a distant

---

[§] Even in the theoretical derivation of an approximation of the Hubble law for nearby space, valid in any Friedmann model, $Z \approx (H_o/c) \times \mathit{distance}$, it is not necessary to refer to Doppler. Lemaître did not use the Doppler effect when he arrived at his redshift law. His formula for small redshifts, in terms of the scale factor $S(t)$, is $(S_2 - S_1)/S_1 = dS/S = [(dS/dt)/S]dt = (\dot{S}/cS)r$, where $dt = r/c$ is the time needed for light to go the small distance r from the source to the observer. $\dot{S}/S$ is the Hubble constant.

galaxy and wants to know its velocity of recession, the answer will depend on the nature of the redshift. A galaxy with redshift of 5 flies away with a velocity of $0.95c$ according to the Doppler effect. But in an expanding space (e.g. in the Friedmann space with $\Omega = 0$), its recession velocity, at the present cosmic time, must be calculated from Mattig's relation between distance and redshift. The result is three times the speed of light. ¶

In order to calculate the velocity of recession for an object with an observed redshift, one must know the exact relation between distance and redshift. This turns out to be a pretty complex mathematical function. The important function was first found in 1958 by Wolfgang Mattig, many years after the discovery of the expanding space models. In a letter to us, Prof. Mattig kindly recollected the history of his famous formula:

"In connection with my doctoral thesis (1957) I had to give a lecture entitled 'The cosmological consequences of the general theory of relativity'. The time for preparation was two weeks. When preparing the lecture I also studied Heckmann's book on 'Theorien der Kosmologie' and I found it extremely insufficient that the relations $z(m)$ and $N(m)$ are given in series expansions. This procedure was too obscure for me. I tried to find a closed form for the simplest case, zero cosmological constant and flat space. I succeeded within the two weeks and got through the examination with a good result. [Then] I looked for a more general solution, and the results $z(m)$ and $N(m)$ were published in Astronomische Nachrichten in 1958.

"At that time this solution was only of academic interest, the largest Z values had been around 0.2. Quasars were not yet known. The reaction of the cosmological community was nearly zero, only Allan Sandage discussed my relations in ApJ 133 (1961). He has rendered accessible my results because my papers were published in German, in an East-German Journal.

"From the beginning of my activity in astronomy (1952 in Potsdam) I have worked predominantly in Solar Physics. I have never worked in extragalactic research, cosmology is my hobby. 1961 I left Potsdam ... and came to Freiburg, an institute for Solar Physics only."

¶The Doppler formula relates the observed redshift Z and the ordinary receding velocity V of a light source: $1 + Z_{Dop} = (c + V)^{1/2}/(c - V)^{1/2}$, hence the velocity is $V = c(Z^2 + 2Z)/(Z^2 + 2Z + 2)$. Note that always $V < c$. For $\Omega = 0$, Mattig's formula (derived for the case $\Lambda = 0$) gives the expansion velocity $V_{exp} = cZ(1 + Z/2)/(1 + Z)$, which may be much larger than c for large Z.

12.4 What is the fate of energy in expanding space?

The conservation of energy has always raised questions in cosmology and given rise to paradoxes. Even Cosmas Indicopleustes, whose indefatigable criticism of the "pagan" cosmology of spheres we mentioned in Chapter 1, wondered about the motion of the outer stellar sphere, and perhaps also from where all that power comes ("Since beyond this sphere neither place nor element nor any of their parts anywhere exists, how do ye say it is moved?").

In big bang cosmology the energy paradox of Newton's eternal stars was nicely solved: the finite age of the whole world explains the presence of shining stars. But a new energy enigma crops up. In his *Principles of Physical Cosmology* James Peebles summarizes the development of 20th century cosmology. In the chapter on the thermal background radiation he discusses the riddle of the cooling cosmic gas of photons. Where does the lost energy go? He concludes that "The resolution of this apparent paradox is that while energy conservation is a good local concept, ..., there is not a general global energy conservation law in general relativity theory."

The conservation law asserts that the energy of a physical system can be changed only by two ways: doing work on the system (e.g. compressing it) or adding or removing heat. As a formula this is

$$\textit{change of energy} = \textit{transfer of heat} - \textit{work done}$$

For instance, when a pressure is applied to a piston, making the volume of the gas in the cylinder decrease, the work is equal to *pressure* × *change of volume*. If heat is not added or taken away in this process, the energy of the gas must increase by exactly this same amount, which is seen as an increase of temperature of the gas. In this case the above formula tells:

$$\textit{change of energy} + \textit{pressure} \times \textit{change of volume} = 0$$

The zero on the right side means that in any such process the total energy (energy of substance + work) is conserved.

In the laboratory the increase of volume is caused by motions of a container's walls in a static space. There are three different cases depending on the sign of the pressure. For *positive* pressure (e.g. usual gas and radiation) the internal energy decreases with increasing volume. For *zero* pressure (dust-like matter) the energy does not change. In the case of *negative* pressure (vacuum or dark energy) the internal energy must increase.

In expanding space things are marvelously different. If we apply the above formula, the change of volume now corresponds to the amount of space created. However, because on both sides of the walls the pressure remains the same (the gas is everywhere uniform), there is no work done on the gas when the walls move away from each other, i.e. "pressure × change of volume" must be equal to zero, hence the internal energy should not change. But nevertheless radiation and gas cool down during space expansion and so the energy decreases with time. [‖] *Where has the energy gone* if the uniform pressure does not perform work? This also means that the cosmological big bang is not a bang at all, because in any explosion in a physics lab there is a difference between the high pressure in the center and the low pressure outside the bang. The homogeneous universe has a uniform pressure everywhere and there is no outside space.

Edward Harrison wrote in his classic book for a wide readership *Cosmology*: "To the questions where the energy goes in an expanding universe and where it comes from in a collapsing universe the answer is – nowhere, because in this one case energy is not conserved." This energy paradox inspired Harrison in 1995 in an article in the Astrophysical Journal to point out that the expansion of the universe could be harnessed to provide the legendary "free lunch". He presents a Gedanken experiment for mining energy in an expanding universe and in this way reveals that there are unresolved issues concerning the conservation of energy.

We may see examples of the non-conservation of energy in the behavior of gravitating mass in expanding space. Take a simple universe containing dust-like particles. Consider a sphere whose radius increases together with space so that the number of particles within it does not change (this is a so-called comoving sphere). The gravitating mass of this sphere, which determines the dynamics of space expansion, also remains constant in time. However, the gravitating mass of the radiation in the comoving sphere is decreasing, and the absolute value of the vacuum mass is increasing!

[‖] This change of energy directly follows from the Bianci identity for Einstein's equations, which gives $T^{ik}_{(m);i} = 0$. This "continuity equation" implies $dE = -pdV$. It is important that the continuity equation does not express the conservation of energy-momentum, as emphasized by Landau & Lifschitz in *The classical theory of fields* (1971), (sect.101: The energy-momentum pseudotensor). This is because in general relativity the EM tensor $T^{ik}_{(m)}$ does not contain the part corresponding to the gravitational field, while the conservation law should include both matter and gravity.

12.5 Superluminal recession of remote galaxies

Relativity theory states that no particle or signal can move faster than the speed of light. All the fundamental theories (electromagnetic, weak, and strong interactions) are called relativistic for this very reason. But in the big bang model very distant galaxies have recession velocities larger than the velocity of light, as follows from the exact formula for the expansion velocity in a homogeneous space. This is related to a fact pointed out by Edward Milne, a British astronomer, in 1934: big bang expansion is described by exactly the same formulae which determine the behavior of the uniform Newtonian dust cloud. The kinematics of the cloud completely coincides with that of the big bang model. As in Newton's theory there is no maximum velocity, this freedom goes over to the expanding space. **

Edward Milne 1896-1950

Allan Guth, the father of inflation, wrote in his article "The Big Bang and Cosmic Inflation": "Although this violates the premises of special relativity, it is completely acceptable in the context of general relativity ... There is nothing in General Relativity that places any limit on the speed with which such stretching can take place."

An essential aspect of the model of an expanding universe is that the space itself expands and galaxies do not move within it. In this sense the space of general relativity reminds one of the absolute space of Newton's cosmology. In general relativity, absolute mathematical space can expand, while Newton's absolute space is static.

The laws of special relativity are concerned with motion *within* space. Hence one might explain away the superluminal motions because the galaxies are not actually going anywhere – they do not carry any information from one place to another. Hence their huge receding speeds are not the kind of velocities on which relativity puts restrictions. The increase of their distance is simply due to the emergence of space between galaxies.

However, the problem of superluminal expansion still haunts the mind and the coherent stretching of the space even at points separated by distances much more than the Hubble distance makes one wonder how the universe "knows" to create space everywhere with the same rate.

**The exact relativistic Robertson's formula is: $V_{exp} = H \times R = c \times \frac{R}{R_H}$, where $R_H = c/H$. For a distance $R > R_H$ the expansion velocity V_{exp} exceeds the speed of light.

12.6 Geometry and physics: views of Poincaré and Einstein

Henri Poincaré reasoned that it will never be possible to experimentally determine whether the geometry of space truly deviates from Euclidean. Einstein was convinced that this is possible. As modern cosmology has now measured a specific, rather accurate value for space curvature (zero, i.e. Euclidean space), it is interesting to cast a glimpse at the old debate.

If one asks a mathematician: what is geometry?, the answer may be frustrating. Take the definition given by Oswald Veblen and Alfred Whitehead: a part of mathematics is called geometry because this name seems good for a sufficient number of competent people. For mathematics itself it is not interesting how geometrical concepts are related to real space, while for physics it is the question of life. With its ideal entities, mathematical space cannot be a fully adequate picture of real physical space. *Mathematical* concepts "distance" and "curvature" are not equivalent to ***physical*** "distance" and "curvature". As Einstein said, the geometrical "singularity" may not be a part of physical reality, but may only mean the inadequacy of the mathematical scheme which leads to the singularity.

Oswald
Veblen
1880-1960

Alfred
Whitehead
1861-1947

Poincaré says that geometry does not deal with real things because its notions are elements of the ideal world. Only geometry together with physics, the *geometry–physics* unity, is subject to experimental study. You may first choose geometry and then find the physical laws, so that there will be no contradiction with experiments. Or you may go the other way round and start from physical laws and find the geometry. Poincaré's conclusion was that it is convenient to change physical laws and not touch the simple Euclidean geometry. For the modern physicist, it is not only simplest, but also carries the deep result of Noether's theorem: the symmetry of Euclidean space guarantees the conservation laws, so fundamental for physics. An example is quantum field gravity operating in Minkowski space–time.

Einstein admitted, in his 1921 lecture on geometry and experience, that in principle Poincaré was right, because geometry is based on the notion of a rigid rod, whereas in the real world it may be hard to find such a thing. Nevertheless, Einstein thought that the existence of "practically-rigid" rods permits that geometry can be the subject of empirical study. The question of whether the geometry of the universe is Euclidean or not has a clear meaning and the answer can be furnished by experiment, which he calls practical geometry. It is natural to choose Riemann's geometry and formulate physical laws within it. This is the way of general relativity.

12.7 Absolutely soft and hard meter sticks

Non-Euclidean geometries – as mathematical models – do not contain internal contradictions. Hence, physical space might be non-Euclidean. Is the curvature physically measurable, when one uses real units and procedures of length measurement? As Allan Sandage has emphasized "if space curvature is real, it must make a difference in something we can measure".

Hans Reichenbach 1891-1953

For grasping 3-D curved space, one may imagine 2-D beings on the surface of a sphere. This idea was carefully discussed by Hans Reichenbach in his *Philosophy of Space and Time.* The flat inhabitants can "walk" around their world with a number of steps and establish its finite size. They can also determine the curvature by measuring the sum of the angles of a triangle. The sides of the triangle are drawn by finding the shortest paths between its corners. [††] In principle, such a measurement can also be made in our 3-D world.

The above description of how to measure the curvature seems obvious, but it depends on a crucial assumption: There should be in a curved space an unchangeable unit of length which can be transferred from one place to another either by free (undistorted) motion or by the information inferred from light rays as in astronomy.

In Euclidean physical space, one can define and understand the unit length as the distance between two freely moving particles which were put into motion with equal velocities, e.g. perpendicular to the line through them. In this way the length unit may be transferred into any point in space. Such an "absolutely soft" (free motion) meter stick gives the same results of measurement as an "absolutely hard" rigid stick. The absolutely hard rod is something which resists all forces trying to change the distance between its end points when it moves in space.

Things are otherwise in curved space. There the Euclidean straight lines are replaced by *geodesics*, the shortest routes connecting two points. Then the two procedures (rigid stick or free motion) give different results.

During its motion on the sphere, the end points of a rigid stick do not move along the geodesics. Then the measurement of curvature is possible

---

[††] For a triangle ABC on a sphere the sum of the angles A,B, and C is $\Sigma = A + B + C = 180^o + \sigma$. Here the angular excess σ is $\sigma = \frac{S}{R^2}$, where S is the area of the triangle and R is the radius of the sphere. Hence, using measurements of the angles A, B, and C, and the area S of the triangle, the 2-dimensional beings, operating only inside their space, can determine the curvature of their space $\mathcal{K} = 1/R^2$.

and its radius may be expressed in the units of this rigid meter. Or vise versa, the local lengths may be expressed with the fundamental radius of curvature as a unit. Perhaps having this in mind, in a letter to Taurinus, dated November 8, 1824, Gauss wrote "*... I sometimes joke that it would be good, if Euclidean geometry were not true, because then we would have an a priori absolute measure of length...*" Eddington preferred for this very reason the closed spherical universe – its radius of curvature gave him a ubiquitous comparison length; this he needed for his complex cosmological theory which linked the microcosm of atoms and the macrocosm of galaxies.

One may also attempt to transfer the unit length from one place to another using two free particles. Free motion is motion along geodesics. In this case, the distance between two freely moving particles which have been put into motion with equal velocities perpendicular to the line through them, is by definition equal to a unit. If a constant length of the unit is defined in this manner, intrinsic geometrical measurements (distances, angles) cannot make a difference between Euclidean and curved geometry.

The above emphasizes the role of truly rigid stick in practical geometry whose main aim is to measure the global curvature of space by observations of the galaxy universe. Cosmological models with regular geometries are based on the assumption of absolutely hard meter sticks. Thus practical geometry is not only aimed at deciding which is the geometry of our universe, but also to determine whether the basic concepts applied to space geometry (such as rigid sticks) are valid on cosmological scales.

12.8 Geometry of space in the local galaxy universe

Euclid is said to have lived in Alexandria, in the court of King Ptolemy I, where he taught mathematics, possibly after having studied Plato's philosophy in Athens. He is supposed to have remarked to the King that there is no royal road to geometry. So little is known about his personal life! But he will always be remembered for his *Elements* which is said to be the book most reproduced and studied, next to the Bible. (It is now claimed that the popularity of Euclid's triangles has been surpassed by certain rings about which prof. Tolkien has written.) Euclid collected together all the geometric knowledge of the time and presented it as axioms, known as Euclidean geometry. It still serves as an excellent model for an axiomatic system where from a few basic premises a rich collection of results can be inferred.

Euclid
c.300 B.C.

Astronomers measure distances to galaxies using the familiar rule: a distant galaxy looks smaller than its nearby twin. Alternatively, they measure the flux of the light and utilize the $1/R^2$-law to derive the distance. (Euclid mentions that of two equal segments that which is more distant makes a smaller angle. It seems that Kepler first pointed out that the flux of light propagating in every direction from a (point) source diminishes according to the inverse square law.) Within about 100 Mpc both kinds of distance measurements lead to identical results, as expected if Euclidean geometry is valid. For example, the value of the Hubble constant is the same from angular size and flux distance indicators.

But what does one mean by "distance" in general? The simple answer is that the distance between two bodies is the length of the rope stretched between them, measured by a unit stick. In fact, this is what is called the *metric distance* in geometry, which is also the fundamental distance in our cosmological models based on rigid meter sticks. However, such a tape measure is not practical for large distances even on the Earth. Other methods for measuring distances give rise to even other concepts of distance.

We mentioned above two popular workhorses of the astronomer: *angular size distance* and *flux distance.* In a transparent Euclidean space angular size and flux distances are equal, and furthermore, they are equal to the fundamental metric distance. This cannot be directly measured across astronomical vastness. In a non-Euclidean space, all these distances may be different. The cosmologist attempts to utilize this diversity in order to restore the true geometry by comparing the different distances.

12.9 The classical cosmological tests of space geometry

There are several methods which are used in modern cosmology to detect space curvature. In particular, the classical cosmological tests are:

- *counts of galaxies*
- *fluxes and redshifts of standard candles*
- *angular sizes and redshifts of standard rods*

Everyone knows that in Euclidean space the volume around an observer increases like the cubic of distance. This is not so when the space is non-Euclidean. In spherical space, the volume increases more slowly and in a hyperbolic space more quickly. Then, the correct geometry could be derived

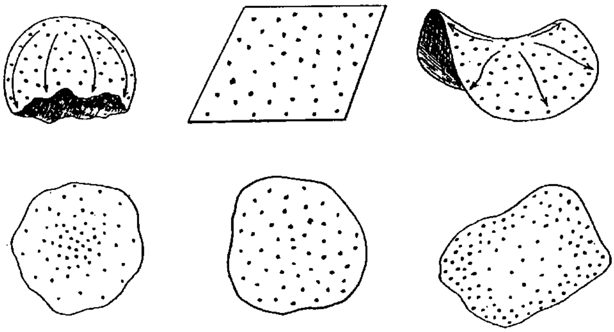


Fig. 12.1 Counts of galaxies in spherical (on the left), Euclidean, and hyperbolic (on the right) spaces. Compared with Euclidean space, in spherical space the number of galaxies at large distances is smaller, while hyperbolic space has more volume for distant galaxies.

by a simple counting of galaxies up to different distances.

In practice, however, this idea cannot be easily applied, as astronomers do not know distances to all the galaxies. And if one, instead of distances, uses the magnitudes of galaxies, as Hubble first did in the 1930's, then the counting test is seriously complicated by the unknown evolution of the luminosity of galaxies. This has been the Achilles's heel of the test. Furthermore, the counting test is only valid if galaxies fill space uniformly – an assumption which is one subject of our book.

In the methods of standard candles and rods one tries to see how apparent magnitudes and angular sizes of celestial bodies depend on their distance (i.e. redshift). Friedmann models predict different behavior of these in differently curved spaces. We have already encountered the most successful standard candle class for large distances, the supernovae of type Ia. Instead of revealing space curvature, as expected, they unexpectedly added the cosmological constant to the Friedmann model. Now the sum of matter and dark energy densities determine space geometry, whereas the behavior of the apparent magnitude does not depend on their sum, but on the two types of densities separately. This means that several different tests will be needed in order to measure the geometry of space.

In non-expanding Euclidean space the angular size of a standard rod is smaller the larger the distance, as we well know from everyday life. But something dramatic occurs in Friedmann models. They predict, as was

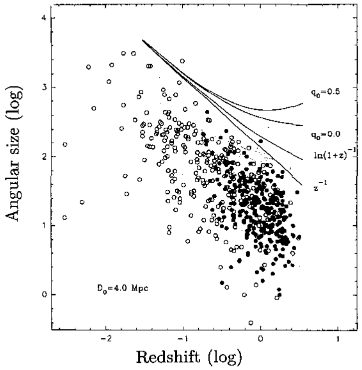


Fig. 12.2 The dependence of the angular size of double radio sources on redshift for radio galaxies (circles) and quasars (dots). These data do not show the expected minimum in the angular size around redshift 1 (i.e. at $\log z \approx 0$). The diagram made by K. Nilsson.

pointed out by Fred Hoyle in 1959, that first the angular size decreases up to a redshift of about 1 (the exact value depends on the density parameter) and then, surprisingly, starts to increase! A rough way to understand such a behavior is to remember that the universe was smaller at the time when the light left the two ends of the standard rod. But the rod itself was of the same size as a nearby standard. Hence it required, relatively speaking, more space and appears to the observer large. Note the assumption of the rigid meter stick!

Since Hoyle's suggestion, astronomers have attempted to measure the geometry of space from angular sizes. As promising standard rods were regarded double radio sources around radio galaxies and quasars. With their intrinsic sizes reaching hundreds of kiloparsecs, double sources are seen in the sky with angular sizes of tens of arc seconds even at redshifts exceeding one. To the embarrassment of astronomers the observations up to most distant double sources show a continuous decrease in angular size and no expected turn-up at large redshifts. In order to understand this behavior of the standard rods within Friedmann models, one has to think about evolution and other effects that would make the double sources smaller in the past. One simply cannot measure the curvature, even though there is in hand a rather reasonable standard rod.

Another type of standard was suggested by Kenneth Kellerman in 1993 who used the length of jet-like structures in the very centers of radio galaxies and quasars. However, the result remains uncertain because the visibility of the internal component depends on many technical parameters. ‡‡

12.10 The patchy microwave sky brings Euclid back

The cosmic background radiation has remarkable properties that provide new types of cosmological ("post-classical") tests. An example is the simple dependence of its temperature on the redshift as $T = T_o(1 + Z)$, predicted by big bang cosmology, and offering a direct test of the cooling of the universe. Atoms in distant gas clouds have been used as a thermometer to measure the cosmic temperature. For example, at redshift 2, the reading of the thermometer was found to be about 9 degrees K, in agreement with the predicted warming. The temperature can be measured accurately, but a problem is how to take into account the cool dust usually accompanying the gas and having a temperature comparable to that of the cosmic photons.

Nowadays the sensitivity and angular resolution of radio astronomy instruments have reached such a precision that one may use the measurements of the small sky patches with slightly different temperatures as a test of geometry. The long-duration balloon experiments Boomerang and Maxima, and the ground-based interferometric observations by DASI have detected a typical angular size of the patches, about one degree across (about the size of two full moons).

Such an angular size, in the frame of big bang cosmology, is strictly related to the total density parameter of the universe. The size of one degree leads to $\Omega = 1.01 \pm 0.02$. In Friedmann models the density parameter equal to one implies that the curvature of the expanding space is exactly zero. The accuracy of 2 percent means that the old dream to measure the curvature of space has turned into a reality. And the result is exciting: Euclid is back!

There is a story, which might even be true, that Newton laughed only once in his life, and is supposed to have happened when someone asked him whether Euclidean geometry was already obsolete. For some reason this

---

[‡‡] A more promising standard rod was proposed by Kaj Wiik and Esko Valtaoja in 2001, the size of "knots" (shock fronts) appearing in high resolution radio maps. The linear size of a knot may be calculated from the observed flux and variability and all the sources may be put on one angular size – redshift curve. There are still few sources with accurately measured knot sizes and one cannot yet determine the geometry, but in future this new method may deliver more strict constraints on the curvature.

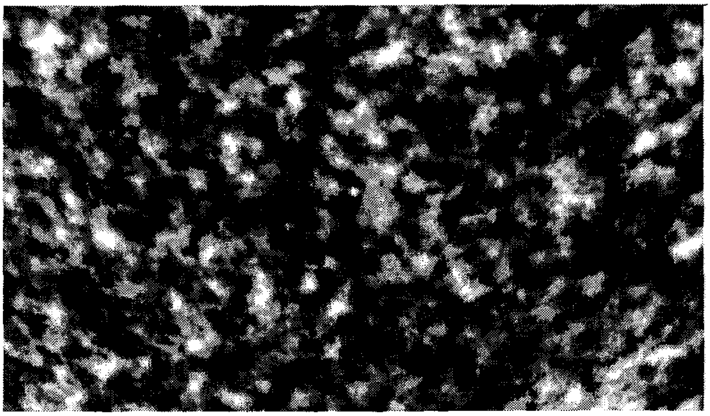


Fig. 12.3 The map of the tiny temperature fluctuations of the cosmic background radiation measured by the Boomerang experiment inside a 10 × 20 square degree region of the sky. From the typical angular size of the structures, which is 1 degree, one may infer that the curvature of space of the universe is zero!

anecdote creeps into our mind now... However, one needs further study of intervening dust and the recently discovered numerous and distant galaxies emitting submillimeter radio waves, in order to be sure that the cosmological message has been reliably extracted from the observed radiation.

12.11 The enigmatic unity of space, matter, and energy

The big bang universe is the next major step after Newton's vision that embraced the whole cosmos. Summarizing the above, the new cosmological picture of the universe again hides enigmas and puzzling features:

- *the singularity*
- *the continuous creation of space and energy*
- *the common cosmic time for observers carried by space*
- *the preferred state of rest relative to the cosmic radiation*
- *the superluminal expansion speed of galaxies*
- *the Newtonian equation of expansion dynamics*
- *the Euclidean geometry of expanding space*

The inevitable existence of the singularity in big bang universe is a great enigma. The physical law of geometrical gravity predicts singularities, the end of all physical laws. Perhaps this is our best indication that we need a theory of quantum gravity, before we can make weird extrapolations into the singularity and beyond.

The continuous creation of space and energy is a paradoxal consequence of the homogeneous expanding universe as a whole. Are we here facing, as in the case of the quantum–mechanical nebulous particle, radically new physics, a novel concept of global energy?

The common cosmic time for observers rushing away from each other also looks paradoxical, because in laboratory physics moving observers have different observed times and it is a chief tenet of relativity theory that there is no universal absolute time. But for the universe as a whole there is a kind of preferred time. The cosmologist Michael Rowan-Robinson has pointed out that, remarkably, the time in modern cosmology reminds one of absolute time as envisioned by Aristotle! And which, of course, was at the heart of classical physics.

A related puzzling thing about space has been recently emphasized by Raimo Lehti: if one can point to some one time as a preferred one, then there should also be a preferred space, as space and time are intermingled in relativistic space-time. Indeed, the cosmic radiation that fills the universe gives the possibility of determining the preferred state of rest for an observer. It seems that in spite of the principle of relativity, which is in excellent agreement with everything that the physicist observes in his laboratory, the universe has been so structured that there is a space relative to which one may measure our velocity as about 400 km/sec. But one may also argue that this is our speed relative to a material component of the universe, the electrons which scattered the photons towards us just before they combined with protons, some 300 000 years after the big bang.

In the Friedmann universe the dynamics of expansion is ruled by the relativistic equation, in which the acceleration is GM/R^2. This is unexpected because the equation is exactly the same as in Newton's theory which is valid only locally and in restricted conditions. Though here the mass includes the contribution from pressure, this does not change the Newtonian properties of the equation. For example, the relativistic formula for expansion velocity $V_{exp} = HR$ is the exact solution for the classical expanding dust cloud. Thus there is no limit, such as the speed of light, for the recession velocity of a galaxy. To borrow Edward Harrison: "In all its applications, Newtonian theory is only approximatively true, and yet in this most unlikely of all instances it yields the correct answer."

Cosmological tests have found that the curvature of space is zero within an accuracy of two percent. If this result holds, it means that the spatial geometry of the universe at any cosmic time is very close to Euclidean. This was unexpected after so many years of search for *curved* space, though the inflation model had prepared us for a Euclidean come-back! But one should note that the flat space of the Friedmann universe is not quite the same as the classical Euclidean space. Though a snapshot picture of this space at any cosmical moment makes it appear Euclidean, it is still a dynamic, expanding space, with an intimate link to the density of matter and energy. And, of course, the universe is no object of classical physics, even if the circle seems to be closed, when one encounters something like "absolute time", "absolute space" and "velocities greater than the speed of light", but now on a higher level of reality.

In a sense, Poincaré with his desirable Euclidean geometry and Einstein with his belief in practical geometry, were both right, even if in an unexpected manner. When they were pondering geometry they did not yet know of the expanding space which was to give an element of surprise to the entire affair.

For the infinite Friedmann universe there are two extreme geometries – that corresponding to no mass and that of the critical density ($\Omega = 1$). Our world, clearly containing matter, appears to have been forged according to the latter mould. Personally, we feel at home in such a universe. Euclidean geometry is the only one which is self-similar for all scales.

* * *

Intriguing conceptual enigmas hide in the foundations of the homogeneous world model. Furthermore, the model cannot live without things of a completely unknown nature, dark matter and dark energy, which determine the fate of the universe. Clearly, cosmology based on the principle of uniformity is still far from the ultimate picture in which the only interesting task is to adjust the values of the cosmological parameters. In contrast to this way of thinking, it is exciting that modern observations reveal strong non-uniformities in the spatial distribution of galaxies, gathering storms over the uniform world model which oversimplifies reality. The last Part of our book takes the reader into the deep realm of cosmic structures, which not so long ago was Terra Incognita.

PART IV

THE FRACTAL ARCHITECTURE OF THE UNIVERSE

Chapter 13

Cosmic hierarchies: from dream to science

The concept of hierarchy originally referred to "sacred power" (in Greek *hieros* = sacred, *arkhe* = power) or the class of clergymen who were at a level in the world above ordinary mortals and acted as an intermediary from God to people. The hierarchy, with its strict order of precedence, was extended also above the heads of human beings, to the hierarchy of angels, to which Dante ascribed the motive powers of the celestial spheres.

Nowadays the hierarchy is best known in its more mundane meaning, as any system of persons or things ranked one above another. Often this concept is used to characterize systems with a *special kind of structure* made of subsystems, and these having subsystems at a still lower level *etc.* Such hierarchical organizations abound in human society, the animal kingdom and also in the realm of atoms. And the linguist may attempt to grasp the grammar of a language in terms of levels according to which it is organized.

The hierarchical cosmologies of eighteenth-century thinkers were partly inspired by the vision of other solar systems (Ch.4). Their simple structures were not based on detailed observations of the sky. But these "protofractals" in the path towards fractal Nature have not lost their historical luster.

* * *

13.1 Searching the heavens for nebulae

The study of nebulae was started by William Herschel. He had emigrated from Germany to England, when he was nineteen years old, and earned there his living as a musician. He entered astronomy at the age of 35, having been inspired by a book "to see heavens and planets" with his own eyes. Herschel became a skilled builder of telescopes and he used his instruments, the largest ones in the world at the time, for a careful inspection of the sky. His largest tube had a mirror with a diameter of 120 cm. This giant could be used only with several assistants and its operation was not without risks. More than one handy-man were injured in accidents. Herschel discovered in 1781 an object which he first thought to be a comet. The Finnish astronomer Anders Lexell showed that its orbit was almost circular, hence it was not a comet, but a planet. * A great event for astronomy – the first planet discovered since ancient times! Naming it was not easy, but finally Uranus was adopted. This achievement brought Herschel the position of court astronomer.

William Herschel 1738-1822

Anders Lexell 1740-1784

Charles Messier sent to Herschel a copy of his nebula catalogue. This inspired Herschel to start a new survey of the heavens. His "sweeps" of the sky utilized the daily rotation of the Earth: the telescope was held in a fixed position and Herschel looked through it as a strip of the sky together with its stars, star clusters, and nebulae marched across the field of view. In this manner he, working together with his sister Caroline, discovered 2500 new nebulae and star clusters, initiating the building of all-sky databanks, which still continues.

Caroline Herschel 1750-1848

William Herschel's son John became an eminent explorer of the sky, too. He utilized not only the rotation of the Earth, but also its spherical shape, in order to map the sky. Namely, he took one of his father's telescopes down to South Africa, to the Cape of Good Hope, where he could study the southern hemisphere of the sky. There he found 1700 new nebulae and clusters.

John Herschel 1792-1871

John Herschel was an eager experimenter with the new art of making photos by Daguerre's method. In fact, in 1839 he was the first to use the term "photography". He first suggested daily photographs of the Sun to record the positions and sizes of its spots. However, during Herschel's lifetime photographic plates were too insensitive to make images of nebulae. This only became possible with the introduction of silver bromide dry plates. In 1880 Henry Draper succeeded in photographing the Orion nebula. The following development of nebula photography was rapid and within a few decades revealed the huge realm of extragalactic space.

---

*Lexell, by the way, made his career in those two places where this book was written, in the towns of Turku and St. Petersburg.

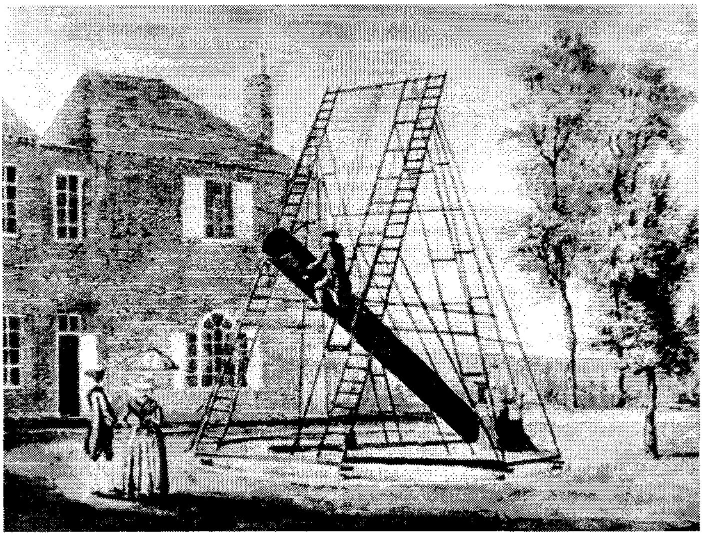


Fig. 13.1 William Herschel's 47 cm telescope which he used for a systematic survey of the sky. His devoted and able assistant was his sister Caroline. In her diary she wrote about practical problems: "My brother began his series of sweeps when the instrument was yet in a very unfinished state, ...every moment I was alarmed by a crack or fall, knowing him to be elevated fifteen feet or more on a temporary cross-beam, instead of a safe gallery ..., and one night, in a very high wind, he had hardly touched the ground before the whole apparatus came down. Some laboring men were called up to help in extricating the mirror, which was, fortunately, uninjured, but much work was cut out for carpenters next day."

13.2 John Herschel's principle of subordinate grouping

The cosmologist Edward Harrison has pointed out that John Herschel was, among many others, intrigued by the enigma of the darkness of the night sky (Ch.3). Herschel outlined an entirely new solution, which was rediscovered in the early years of the 20th century. In a private letter he wrote:

··· it is easy to imagine a constitution of a universe literally infinite which would allow of any amount of such directions of penetration as not to encounter a star. Granting that it consists of systems subdivided according to the law that every higher order of bodies in it should be immensely more distant from the center than those of the next inferior order – this would happen.

Clearly Herschel had in mind some kind of a hierarchical system. In another text he gives as examples the satellites of the planets of the solar system and the large distances between stars and asserts that "the principle of subordinate grouping" assumes "the character and importance of a cosmical law". This principle fascinated him: "While it affords another and most striking indication of the unity of plan which pervades the universe, it may lead us to believe that, if other systems yet exist in the immensity of space, they may be separated from our own by intervals so immense as to appear only as dim and nebulous specks, or utterly, and for ever, to elude our sight". He speaks here about the next higher order of bodies, of which even the nearest one would be a very feeble sight on the sky.

It is also interesting that he did not accept the explanation that light extinction by a cosmic medium may make the sky dark. John Herschel correctly reasoned that the absorbing medium itself gets heated and radiant, and gives "out from every point at every instant as much heat as it receives".

13.3 Fournier d'Albe's brave new worlds

Edmund Fournier d'Albe 1868-1933

After a century of sleep, though somewhat stirred by John Herschel, the idea of a hierarchic structure of the universe was blown into life by a London based free-lance science writer and inventor, Edmund Fournier d'Albe, He wrote a book on electricity and magnetism which was intended to be popular, but was praised by leading science journals as "the best possible introduction to modern views of electricity". In the 1910's he worked as Assistant Lecture in Physics in Birmingham University. He also appears in the history of television, having transmitted in 1923 the first television picture from London, a portrait of King George V. In 1912 he invented the optophone, enabling the blind to recognize and locate light by means of the ear. The optophone created various tones depending on the amount of light falling on Selenium cells. [†] Two years later he improved the instrument so as to make possible slow reading of ordinary text. It is curious coincidence that Swedenborg, the old advocate of self-similarity and hierarchy, was credited for the invention of the ear trumpet.

In 1907, Fournier d'Albe published the small, but remarkable book *Two new worlds* in which one finds the first mathematical description of a pos-

---

[†] The chemical element Selenium appears in different forms – in its metal-like form it has the remarkable property that its electrical conductivity is greater in light than in darkness. The modern optophone was invented by Peter Meijer and has been patented by Philips in the Netherlands. It scans an image directly from a video camera.

sible hierarchical distribution of stars. In Fournier's world the stars are distributed in an infinite space, but the mass inside any sphere increases directly proportionally to its radius. Remember that in a uniform distribution, this mass would increase like *cube* of the radius. ‡ Why did Fournier invent such an unusual model? Actually his book is concerned both with the microscopic Infra–World and the large scale Supra–World and he suggested that "a universe constructed on a pattern not widely different from ours is encountered on a definite and measurable scale of smallness, and another on a correspondingly larger scale". He was excited by the possibility that such a universe could resolve the two chief paradoxes of infinite space, namely the blazing sky and infinite gravity.

Fournier's book has sometimes been characterized as science fiction, and certainly he bravely speculates on the existence of life in his Infra– and Supra–Worlds, and he admits that it will be very difficult ever to observe the exceedingly distant elements of the Supra–World, not to speak about the inhabitants of the Infra–World. But these ideas sprang from his rather sound philosophy: "Why should we draw the line before Nature draws it for us?" A couple of years earlier the Japanese physicist Hantaro Nagaoka had proposed that atoms are like miniature planetary systems, where negatively charged electrons revolve around a positively charged nucleus. Later, the discovery of quantum laws showed that in the case of Bohr's planetary model of the atom this simple picture of repeated hierarchic elements has essential limitations.

Fournier's chapters on the large scale world contain many interesting insights which we wish to describe in some detail. He defines carefully which assumptions the dark night paradox rests upon:

- *That the luminiferous ether pervades all space*
- *That the number of dark bodies is comparatively small*
- *That the stars are irregularly distributed [in infinite space]*
- *That luminous stars have an eternal existence*

The first assumption would be expressed today as saying light propagates throughout empty space. He points out that if the stars were created some finite time ago, then we cannot see farther than the distance which light has traveled after the moment of creation, i.e. speed of light multiplied by the

---

‡Fournier's world has $M(R) \propto R$, while in a homogeneous universe $M(R) \propto R^3$. In modern terms Fournier's large scale structure has the fractal dimension $\mathcal{D} = 1$, as we'll see in the next chapter.

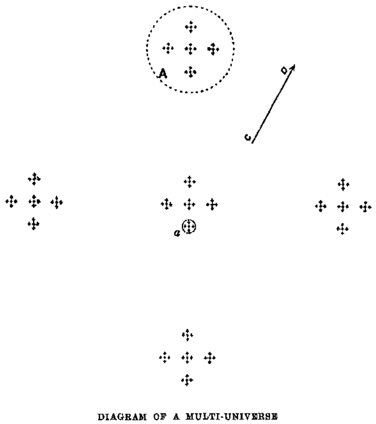


Fig. 13.2 The hierarchic model as sketched by Fournier in his book.

age of the universe. So, if the creation happened 100 000 years ago, to use Fournier's example, the sky is presently illuminated only by stars at most 100 000 light years from us. This solves the paradox of the dark night sky, as we remember from Ch. 3. However, Fournier favors an eternal universe, according to his "root-hypothesis" that

> *...this world of ours is a good average sample of the universe, as it always has existed, and always will continue to exist...I am disposed to believe that this place in which I am, and this moment in which I write, are as significant, as sacred, and as important as any I have ever had, or am ever likely to get. There may be variety, improvement, progress, or decay, but the essential elements of all these I believe to be permanent, and not confined to the Here and Now.*

Fournier takes it for granted that there are a lot of "dark stars" mixed with the luminous stars. They could form an opaque screen for distant stars. He also comments on the possible existence of a substance in space, which could absorb light (nowadays such interstellar dust is a fact, while at the time the evidence for absorption was scarce). Fournier, as John Herschel had before him, points out that the absorbing medium would be heated when catching the light, and would then itself radiate the heat "inwards upon our devoted heads". Fournier recognizes that there may be a similar problem with the dark stars collecting the light rays, though he believes that if dark stars greatly outnumber luminous ones, the illumination and temperature of an infinite universe would be quite comfortable.

13.4 Gravity within Fournier's hierarchy

But it is not Fournier's aim to save the infinite universe uniformly filled with stars. He wants to extend the hierarchic system, which seemed to work so well when one goes towards the atoms, also to larger scales. Then the visible stellar world, the Milky Way, would be only one element of the higher hierarchy level, the other members of which were beyond the available observing power, as Fournier calculated. But can one make any deductions on the structure of the super-systems without direct optical observations? Yes, he answers, gravitation is the key.

It is an interesting fact that stars have relatively small velocities, much smaller than the speed of light. From this Fournier concludes that stars "falling" into our level of the hierarchy from large distances have not reached high speeds, because the density decreases rapidly outwards from us. More exactly, he points out that the gravitational potential at the surface of a sphere of stars would be the same no matter how large the "world-sphere" is, if *... the mass comprised within a world-sphere increases as its radius, and not as its volume.*

To his satisfaction such a hierarchy resolved Olbers' blazing sky paradox and allowed one to understand the observed small speeds of stars.

Incidentally, the question of small local speeds was already touched upon by Giordano Bruno, when he pondered an infinite matter distribution. His explanation was to propose a finite range for the gravity force (Ch.3): "Because the force-spheres and mutual distances of the heavenly bodies are finite, so are also their motions finite."

13.5 Carl Charlier wrestles with infinities

Fournier wrote in his small book about entering "upon the virgin field where, I believe, the science of the future will blossom forth". And indeed, his pioneering work soon found an excited follower in the country that has been a spring of fresh ideas on the cosmic arrangement of matter. The concept of cosmic hierarchy was further studied by the Swedish astronomer Carl Charlier in his article "How an infinite world may be built up", pub-

Carl Charlier 1862-1934

lished in 1908. After a few weeks' enthusiastic inspection of Fournier's book, he developed more general models of stellar distributions which also solve Olbers' paradox and Seeliger's riddle of infinite gravity.

Charlier's speciality was the application of statistical methods to the study of the Milky Way. His interest in cosmology may be traced back to the years when he worked as Observer at the Uppsala Observatory. Karl Lundmark wrote in Charlier's obituary notice that "during his period at Upsala Charlier lived amongst a congenial circle of friends, generally known as the 'Verdandists', after a society founded for the preservation and development of freedom of thought and its manifestations. This society advanced liberal and radical ideas, and from it many of the leading politicians of Sweden obtained their first impressions of public life... Charlier did not take any active part in political life in Sweden, but he was true to his ideals of his youth and always kept an open-minded attitude towards contemporary movements for advancing new ideas and ideals."

It was in 1896 in Uppsala when Charlier wrote his first article on the size of the universe, a year before he took the position of director of the Lund Observatory. The article, written in German, was titled "Is the world finite or infinite in space and time". His conclusion was that the stellar world is finite in space, while infinite in time. The main reason for preferring infinite time was the apparent violation of the conservation of matter if there was some first moment. Also, Charlier thought that an infinite duration of existence does not contain logical paradoxes, contrary to what Kant had argued. And even if infinite time may be hard to grasp, the first moment is a still more puzzling thing conceptually, Charlier reasoned.

His primary reasons for rejecting an infinite universe filled by luminous celestial bodies, were Olbers' and Seeliger's paradoxes. The latter's article on gravity had just been published. At that time Charlier was clearly against the idea of Island Universes, stellar systems outside our Milky Way, holding with the majority of opinion. He concludes by asking "Is there any possibility that the world could be considered as having infinite spatial extent? When we leave aside fruitless speculations about invisible Milky Ways and other similar things, and insist upon the presently available scientific results, then one has to answer No to the question."

It is interesting that Charlier mentions in his 1896 article that previously he had held the opinion that the nebulae are mostly Milky Way -like systems at very large distances. But to his own surprise, he was forced to change his mind, because new studies seemed to favor the view that the nebulae

Fig. 13.3 Carl Vilhelm Ludvig Charlier opened new perspectives for the infinity universe, when he brought Fournier d'Albe's hierarchic model into astronomy.

are inside the Milky Way, amongst the stars. He refers to the symmetric distribution of nebulae relative to the Milky Way (which we discussed in Ch. 7). He also says that many nebulae are found connected with the stars of our stellar system (we now know this is indeed true for "genuine" nebulae of gas and dust).

13.6 Charlier's criteria for infinite worlds

Charlier's classical article of 1908 started with a note that reveals how his inner feelings had not been quite in accord with the conclusions that he had to make: "I did not see at that time any possibility to ignore the arguments against an infinity extension of the world, even though such an infinity of the world (matter) seemed to me almost self-evident from the philosophical standpoint." No wonder he was excited when Fournier's book opened the prospects for an infinite universe.

Fournier d'Albe included a single diagram in his book. It represented an example of how an infinite series of similar successive universes may exist. That example was very regular and simple and was not intended to be a real plan of his Supra-World. Five members or galaxies form one element of a hierarchy level. Then five such elements are combined to form an element of the next level and so on. Fournier noted that in such hierarchic worlds the average matter density decreases when one climbs upwards through levels. This decrease makes it possible to construct an infinite world where the total amount of matter is infinite, but the average density of matter is zero, *and which does not suffer from the* blazing sky *nor infinite* gravity paradox. Charlier's important contribution was to investigate in detail which kinds of hierarchies are such appealing ones.

Charlier found a criterion which the hierarchy must fulfill so that it will solve the two paradoxes. The decisive factor is how fast the density decreases from one level to the next, and this depends on the ratio of the sizes of the successive elements and on the number of the lower elements forming the upper element. Charlier's criterion is simply: *the size of the upper level element is larger than or equal to the size of the lower level element multiplied by the number of lower elements forming the upper elements.*

In 1922 Charlier revised and completed his older article and corrected an error in the derived criterion for saving the world from the blazing sky: in the above formulation one should say "multiplied by the *square root* of the number of lower elements forming the upper elements." As was independently noted by Franz Selety, this suffices to solve Olbers's paradox and the problem of infinite gravity force. However, Charlier's original criterion is needed to keep the gravitational potential differences and the velocities of celestial bodies finite. [§]

Note that Charlier's hierarchy itself is infinite. If after some level the hierarchy ends and uniformity begins, all the paradoxes again appear!

13.7 Towards hierarchic worlds without a middle point

Neither Fournier d'Albe nor Charlier addressed the question of whether their models have a central point. However, Fournier's attitude was that there is nothing exceptional in our surroundings (nor in the moment at which we are living). This suggests that Fournier may have thought that his model is without a center, as is the world which it attempts to represent.

Franz Selety 1893-1933

In fact, it seems that the Austrian scientist Franz Selety was the first to claim that a hierarchic world may exist without a center. Selety (whose surname up to 1918 was Jeiteles) had obtained his doctorate in philoso-

---

[§] Denoting for each hierarchy level i the sizes (radii) with R_i and R_{i+1} and the numbers of elements with N_i and N_{i+1}, Charlier's first criterion is $R_{i+1}/R_i \geq N_{i+1}$. The second criterion is $R_{i+1}/R_i \geq \sqrt{N_{i+1}}$.

phy at the University of Vienna, in late 1915, with a dissertation on the phenomenological basis of psychology. [¶]

In an article in *Annalen der Physik* in 1922, in which both Fournier d'Albe and Charlier are referred to, Selety pointed out the usual belief that the following assumptions cannot be simultaneously valid:

- *infinite space*
- *infinite total mass*
- *mass filling space so that locally there is everywhere a finite density*
- *zero average density of the mass in the whole world*
- *non-existence of a unique middle point or middle region of the world*

Selety argues that it is in fact possible to create hierarchical worlds which fulfill simultaneously all these conditions, most remarkably the two last ones. Zero density means the limit towards which the average density inside any sphere decreases, as the size of the sphere is increased.

How did he attempt to construct a hierarchy without center? Looking at Fournier's "Diagram of a multiuniverse" (Fig.13.2), one can first describe the way which naturally leads to an infinite hierarchy possessing a privileged point. Start from the central element which contains five subelements and has an obvious center. Now construct the higher hierarchy level so that the whole displayed figure is the central element of the next, bigger figure and its center is kept fixed on the plane. Continue in this way without end. Clearly, the result is an infinite hierarchy with a well defined middle point around which it was built. But Selety points out that there are also infinite many other ways to make the hierarchy starting from this same five-element structure.

At the start one may put the central element not into the center of the higher element, but let it be one of the four corner elements. For example, one could cycle clockwise from one corner to the other, building higher and higher hierarchy levels. In this way the middle point would spiral away from the original center and there would be no unique center in the resulting infinite hierarchy. Selety notes that in general such routes of construction leading to no unique center of mass are much more probable

---

[¶] We wonder if Franz Kafka in his *Tagebücher* refers to the same man, when he wrote in Prague on November 19, 1915: "In der Alt-Neu-Synagoge beim Mischnavortrag. Mit Dr. Jeiteles nachhause. Grosses Interesse an einzelnen Streitfragen". "At the Mischna lecture in the Old-New-Synagogue. With Dr. Jeiteles at home. Great interest in certain controversial issues."

than the ways which restrict the center inside some finite region around the original middle point. This he seems to regard as a hint that in Nature also such centerless hierarchies could be a rule and not an exception. [‖]

Thus Selety appears to be sure that a world without center and with zero average density can exist, and he was aware of its unusual properties. Among other things, Selety notes the intuitive misconception (which is still common to-day) that if the average density is zero then somewhere "at infinity" the local density should be zero, so that there would be empty space at infinity. For instance, Seeliger had written that it is not possible to imagine an infinite space filled by an infinitely sparse matter.

It is interesting that Einstein had correspondence with Selety, where he defended his uniform cosmological model, not accepting the hierarchical system, because he thought it could not be reconciled with Mach's principle. When he replied to Einstein's criticism (which he did not accept), Selety once more emphasized that he had in mind hierarchical worlds which do not contain a middle point.

13.8 Knut Lundmark's great plan

Charlier's articles of 1896, 1908, and 1922 span a highly interesting period in astronomy when the true nature of nebulae came into focus. In 1908 Charlier thought that the Milky Way is one element in the hierarchy, but that the other similar milky ways are very distant – from his first criterion, assuming that the number of elements is one million, he estimated that the closest one would be very small and quite unobservable on the sky (with its stellar magnitude equal to 37, a thousand times fainter than what the Hubble Space Telescope can observe!). But in 1922 he already put the Andromeda nebula on the same hierarchy level alongside the Milky Way. He made a calculation about the upper system considering the Milky Way as a member. The closest nebula should then have an angular diameter of a few degrees and the most distant one a few arc minutes. As the closest one he identified Andromeda, and he predicted that in some near future it would be possible to map completely the whole upper system (which he thought has much less members than the Milky Way has stars).

[‖] There is still something artificial in the structures built around one element, and Benoit Mandelbrot has recently pointed out that such a construction still carries with itself the "memory" of the middle point and does not lead to a genuinely centerless universe.

Fig. 13.4 Knut Lundmark: "We can certainly expect to find a very complicated structure in the doubtlessly gigantic universal system, which is formed by the spiral nebulae." (from his doctoral dissertation written in 1919).

This picture fascinated Knut Lundmark, the successor to Carl Charlier in the professor's chair at Lund University. In the 1920's he applied Charlier's model to data on galaxies and concluded that "the three systems of different orders we now know, the stars, the galactic and the metagalactic systems are so arranged that they fulfill the above found condition". He refers to Charlier's later "square root" criterion. Expressed in modern terms, Lundmark made the first rough estimation of the fractal dimension of the galaxy distribution, deriving a value of about 2. However, the data were still very scarce.

It was Lundmark's dream to build a large database, containing all available information on individual galaxies. These data would be used to study the properties of galaxies and especially their arrangement in space. Do they form the next stage in the hierarchy, a Super Milky Way of milky ways? Sadly, the ambitious "Lund General Catalogue", thousands of galaxies described on separate cards, was never completed, and after the death of its originator, all work on this database ended. The boxes carrying the "local universe" were forgotten, collected dust, and their fate is unknown, according to the writer of Lundmark's biography, Anita Sundman.

Lundmark was the forerunner of eminent modern catalogue builders, such as de Vaucouleurs and Paturel. His idea of a large galaxy catalogue was sound, but at the time it did not seem to raise much enthusiasm among other astronomers and he did not have resources to organize the great plan, work on which continued for three decades. Or as Sundman suspects, Lundmark the perfectionist just could not stop collecting at some point the ever in-flowing data, in order to start publishing at least parts of the catalogue. You can never make a complete inventory of the galaxies! But perhaps Lundmark wanted to reach the border of the Super Milky Way? This hypothetical entity, as it was envisioned by Charlier, contained some one million galaxies, still a formidable number!

Camille Flammarion 1842-1925

Lundmark's work gives the impression of a restless soul, who generously divided his time between different activities of science, teaching and public education, lecturing on astronomy and writing voluminously popular articles and books. His childhood hero and a source of inspiration also later in his life was Camille Flammarion, the French writer on the universe, who thought of astronomy as a science of infinity and eternity. One of the first to step into the realm of galaxies, Lundmark did not feel at home within the cosmology that then emerged, with its primordial explosion a finite time in the past. How could the big bang give rise to a hierarchical system? And why should there be a beginning to time? For him infinity and eternity were the remote dim glow alluring the cosmologist and lighting up his path.

At the present time, some five decades since the last galaxy cards were put into the box at Lund Observatory, Georges Paturel – the "guru of galaxy catalogs" – maintains in Lyon Observatory the huge LEDA database, naturally now computer based, which contains data on about one million galaxies. LEDA is the first modern galaxy database, and much more than just a long list of galaxies. It is a versatile instrument for astronomers investigating the local galaxy universe. And indeed, from the treasurehouse of LEDA emerges an image of a system of nebulae, termed the Local Supergalaxy by its discoverer Gérard de Vaucouleurs in the 1950's, and later suggested by Paturel and others to be just a part of a still larger Hypergalaxy. The dream which Knut Lundmark pursued through his life finally became reality.

* * *

After Charlier and Lundmark little attention was paid to the possibility of a hierarchical distribution of galaxies until Benoit Mandelbrot's and Gérard de Vaucouleurs's work in the 1960's and 70's, and then in more recent years when deep space surveys have revealed large non-uniform structures of galaxies, first analyzed for fractality by Luciano Pietronero. But in the meanwhile novel insights down on the Earth prepared the way to a view of cosmic matter which transcends the old, simple hierarchies.

Chapter 14

The charm of self-similarity

The Fractal Geometry of Nature is one of the major achievements of 20th century mathematics and natural science. Benoit Mandelbrot unexpectedly opened our eyes to hidden structures surrounding us everywhere.

Benoit Mandelbrot was born in Warsaw to a family from Lithuania, in the decade when the galaxy universe was discovered. His family emigrated to France, where Szolem Mandelbrojt, the younger brother of Benoit's father, was a professor of mathematics. The uncle was a founding member of the famous club of mathematicians writing under the name "Nicholas Bourbaki" who wished to present all mathematics in a unified form. The nephew also had talents for abstract reasoning, but unlike Bourbaki he felt attraction to geometry, at the time regarded by many as old-fashioned, even "dead". And he helped geometry to rise from its long sleep.

Mandelbrot recollects how in his youth he kept looking for ways to use his mathematical gifts in dealing with real concrete problems in Nature. "My hopes were thoroughly romantic: to be the first to find order where everyone else had only seen chaos." Mandelbrot's career as a mathematician did not follow a regular line either – he describes it as a fractal orbit.

* * *

14.1 The "fractal orbit" of Mandelbrot

Mandelbrot tells that his versatile research interests were a "wild game" which started paying off in the beginning of the 1960's. He became convinced that he had identified a new aspect of Nature. He had discovered mathematical objects, fractals, which allowed him to study complex natural phenomena that were previously regarded as intractable. * Glimpses of strange entities, related to fractals, have appeared in the works of earlier eminent mathematicians. So Henri Poincaré realized that these may have a place in theoretical physics (and therefore, perhaps, somewhere in Nature). However, it was only Mandelbrot who grasped, as a result of his deep excursions into diverse fields, the underlying "regularity in the irregular".

Benoit Mandelbrot asks why the geometry of Euclid often gives an impression of being "dull" or "dry". And he gives a reason. There are many real-life things which Euclid cannot describe. "Clouds are not spheres. Mountains are not cones. Coastlines are not circles. Trunks of trees are not cylinders and bark is not smooth. Lightning does not travel in a straight line."

And vice versa, one may wonder why we do not get tired in looking at the frothy clouds sailing in the sky, the white-headed waves rolling on the sea, or the living flames playing in the fireplace...

In Antiquity the beautiful geometric forms were put either into Plato's world of Ideas or were connected with the perfect heavenly structures or their models. The irregular and complex "sublunar" world could hardly be captured with such concepts. Later, when the making of mathematical models became a normal practice in physics and cosmology, and the difference between Earth and Heaven eroded, even complicated phenomena became the subject of modeling with simple, regular building bricks.

So, much of the behavior of gas became well understood on the basis of the model of the ideal gas, made of hard spherical molecules continuously colliding in a vacuum and reaching an equilibrium state corresponding to the gas temperature. The density of an ideal gas in its container varies little around its mean value and the variations may be described as fluctuations which are independent of each other. In cosmology, it became normal to postulate a substratum which is homogeneous on the large scale, except for possible small ripples. Even the field of economy borrowed from simple physical models: the volatility of stock prices was understood as a Brownian

---

*After Paris was liberated in 1944, Mandelbrot entered a leading science school École Polytechnique. He defended in 1952 at the University of Paris his PhD thesis on communication and information theory. He has since studied such diverse fields as thermodynamics, economics, and geophysics. During his long career at IBM Mandelbrot became a pioneer of computer graphics. Currently he is Sterling Professor of Mathematical Sciences at Yale University in the USA.

Fig. 14.1 Benoit and Aliette Mandelbrot at the Mittag-Leffler Institute of the Royal Swedish Academy of Sciences, one sunny autumn day in 2001.

motion in analogy with how a particle erratically moves in a fluid where it is randomly bombarded by the surrounding molecules.

But a closer look at many physical systems, including turbulence, coastlines, clouds, even stock prices ..., shows that they cannot be understood on the basis of well-behaving, ideal gas -like models. In fact, Mandelbrot's "fractal orbit" started from economics, when he saw that the erratic variations in prices resemble a process obtained via self-similarity.

14.2 The concept of the fractal

The concept of the fractal grasps an essential aspect of Nature which was previously overlooked – even its rough features have hidden regularities. It also means that apparently chaotic phenomena may have deep structure. The word 'fractal' was coined by Mandelbrot in 1975. He explains that it comes from the Latin adjective 'fractus' which derives from the verb 'to break' or to create irregular fragments.

Karl Weierstrass 1815-1897

Fractals as mathematical entities first appeared in 1875 when Karl Weierstrass discovered a continuous non-differentiable function. Before that mathematicians had thought that all continuous curves have a tangent at

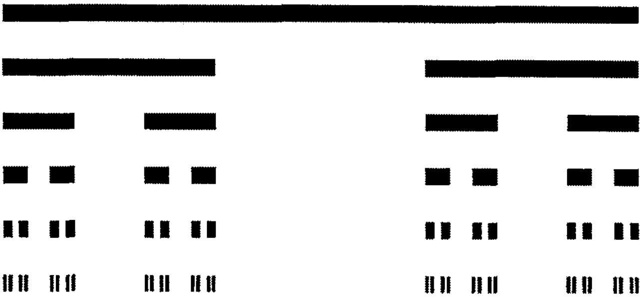


Fig. 14.2 On the infinite path towards the Cantor dust.

every point. This was the beginning of the crisis of analyticity: it was realized that there are mathematical things which are outside analytical description, or curves lacking a tangent at any point. As a result, new branches of pure mathematics were born, such as set theory, topology, and dimension theory. Decisive roles in further study of the new mathematics of non-analytical objects were played by a number of mathematicians, among them Poincaré, Cantor, and Menger.

Georg Cantor 1845-1918

Georg Cantor was born in St. Petersburg, but at the age of eleven moved with his parents to Germany. He was a remarkable student of infinity and created the theory of sets, which is one of the cornerstones of modern mathematics. In 1877 he proved the amazing theorem that the infinite number of points in a one-dimensional line segment is equal (has one-to-one correspondence) to the number of points inside any three dimensional cube (actually he worked with n-dimensional space). This was amazing even for Cantor himself, who wrote that "I see it, but I don't believe it!"

Dreadful things were brought to light, such as the Weierstrass curve, the Cantor dust, the Sierpinski gasket, and the Menger sponge. These inventions of unusual non-analytic structures were received with awe by mathematical society and it was regarded as best to put them in the "Gallery of Monsters". The general attitude is reflected in the words of Charles Hermite from 1893 on "turning away in fear and horror from this lamentable plague of functions with no derivatives". The domain of mathematics had some borders which the prevailing opinion found very difficult to cross.

Later, however, Freeman Dyson noted that "Nature has played a joke on the mathematicians ... the same pathological structures that the mathematicians invented to break loose from 19th-century naturalism turn out to be inherent in familiar objects all around us". One may add another paradox: these repugnant and deformed oddities, once hidden from curious viewers, have as fractals offered for the human eye and mind incredibly beautiful and fascinating experiences.

Fractals as they appear in modern mathematics have many aspects. Here we do not plunge deep into the mathematics of fractal sets, but rather express the essence of fractals in words: *self-similarity, non-analyticity, hierarchy, iteration, randomness, chaos, dimensions*, and *power-laws*.

Often fractals are defined simply as a system, the parts of which are similar to the whole. A study of such a system using a magnifying glass will reveal a new structure that looks similar to what one can see by plain eye. Another way of saying this is that simply from a picture of a part of a fractal structure, one cannot tell its real size.

Similarity is a kind of symmetry, and one of the most fundamental and fruitful concepts of modern science. The beauty and power of the fractals is perhaps explained by the fact that they shed light on the deep symmetries which show how Nature works.

14.3 Koch's curve or snow flake

It is good to have in your mind one strong example of a fractal, simple but offering ample food for thought. Such is the Koch curve. If you find it incomprehensible, you are not alone: it is one of those continuous curves which do not have a tangent at any point (it is nowhere differentiable) and which the mathematicians of the 19th century found repugnant. Helge von Koch, a Swedish mathematician working at Stockholm University, created his celebrated curve in 1904, as Mandelbrot says, "in order to help other mathematicians to get away from the track in which they had been stuck".

Helge von Koch 1870-1924

One starts the construction of Koch's curve with an equilateral triangle. Divide each side into three equal segments and place on the middle segment an equilateral triangle. The result outlines a "Star of David". Then go on with each 12 equal segments and repeat the above operation. Then repeat it again, and again...

At every step the length of the curve increases by a factor of 4/3. This shows that Koch's curve is *infinitely long*, because it is formed when the number of steps increases without limit. At the same time it encloses a finite

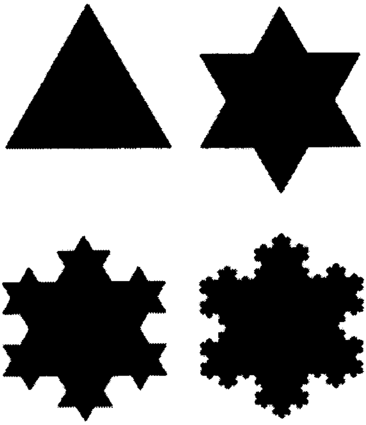


Fig. 14.3 The first steps in the construction of Koch's snowflake

area! Such a curve cannot be quite ordinary – think about an infinitely long thin thread on a plane, bending on itself on a complicated manner in order to produce a closed, finite area inside.

14.4 The simple measure of complex structures

Felix Hausdorff 1868-1942

Even though fractal structures seem very complex and may be very diverse, they have one simple characteristic called the *fractal dimension.* Let us consider the Cantor dust. It is constructed from an interval of a straight line by deleting the middle third and repeating this fragmentation without end (see Fig.14.2). The construction is deceptively simple, but the resulting set has infinitely rich structure. It can be regarded as being made of as many (infinite) points as the initial interval itself, but it is not a collection of continuous one-dimensional small segments. Neither is it "dust" made of clearly separated individual zero-dimensional points. Thus its dimension lies between one and zero! One year after Georg Cantor's death, the German mathematician Felix Hausdorff introduced to mathematics the concept of non-integer dimension, which generalizes the familiar notions of length, area and volume. When one speaks about the fractal dimension, one usually has in mind the Hausdorff dimension. His definition gives the Cantor dust a dimension of about 0.63.

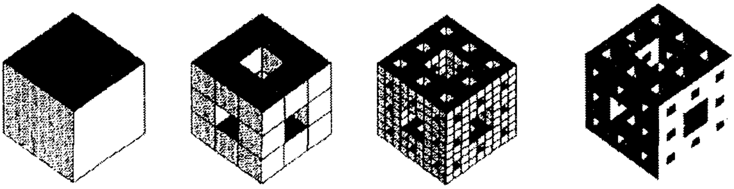


Fig. 14.4 Making the Menger sponge.

Let us inspect another regular fractal structure in three dimensional space, which illuminates the concept of fractal dimension. This is the Menger sponge which is useful when we later discuss the spatial distribution of galaxies. Its creator, Karl Menger, was an Austrian mathematician who emigrated to the USA in 1938. He is best known for his work on the concept of dimension.

In the first step of making the Menger sponge, take a solid with a unit side, unit mass, and unit density (see Fig.14.4). Divide each face into nine equal squares. In the next step make three holes through the central squares of the cube. Then repeat this procedure for all the remaining 20 smaller cubes. The new structure encompasses the same volume, but the mass has decreased by the factor 20/27. Hence, this operation led to a structure with the same cubic outline, but smaller density. One may continue building the sponge repeating the same iterative procedure, using for the next stage the new smaller cubes. The structure resembles that of the familiar bath sponge where all the holes are connected. It is this property which makes the sponge so useful: it absorbs water voluminously into its holes and all the water is easily squeezed out.

Karl Menger 1902-1985

The genuine Menger fractal is the result of an infinite number of such steps and extremely difficult to imagine. However, if we apply this model to some physical objects, Nature puts some size limit to the smallest elements and also the initial cube is finite. The main fractal features can be clearly seen within these length limits.

For example, let us stop the fragmentation after n steps, when there are 20^n small cubes. Take one such element and draw around it a sphere. The main property of a self-similar fractal structure is that the number of elements inside this sphere grows as a power of the radius of the sphere. If observers sitting on the points count the number of the objects around them inside spheres of different radii, they will find a simple relation between number and radius:

$$\textit{number} = \textit{constant} \times \textit{radius}^{\mathcal{D}}$$

The exponent $\mathcal{D}$ is the fractal dimension of the cluster. For the Menger sponge the fractal dimension is about 2.7. The general formula for the fractal dimension for a regular structure is

$$\mathcal{D} = \log N / \log R.$$

The numbers N and R say how the fractal is generated: for each step the previous element is fragmented into N equal subelements so that the ratio of sizes is $R > 1$. [†] $\mathcal{D}$ may well be a fraction or even a real number.

The Cantor dust, the Koch snowflake, and the Menger sponge illustrate the general procedure for generating a regular (non-random) fractal: 1) an initial element, 2) fragmentation or aggregation, following 3) an iteration rule, 4) self-similarity, and 5) a power law number–radius dependence.

For uniformly distributed points, the fractal dimension $\mathcal{D} = 3$. Roughly speaking, such a point set obediently delineates the underlying familiar space. For general fractal structures, embedded in 3-D space, the fractal dimension is less than three. Such a structure no more "fills" the space in such a dull manner as a uniform distribution does. When the fractal dimension is less than 3, the above formula tells that the number density of points increases when the considered volume is decreased. Hence on smaller and smaller scales, where the eye would need a zooming magnifying glass, the rising density allows the appearance of "new" points, making rich structures there where the uniform distribution would offer no surprises.

14.5 The fractal dimension of Fournier-Charlier worlds

Fournier's world model was in fact an early attempt by a physicist to make a protofractal. For regular, hierarchical clusters, discussed by Fournier and Charlier, the fractal dimension is simply obtained from the above formula containing the number N of elements per cluster and the size ratio R. The fractal dimension is the ratio of their logarithms.

[†] For the Menger sponge, $N = 20$ and $R = 3$, hence $\mathcal{D} = 2.7268\ldots$ For the Koch curve $N = 4$ and $R = 3$. Hence $\mathcal{D} = \log 4/\log 3 \approx 1.26$. When Cantor's dust is created, each segment is divided into three equal parts out of which two are left for further use, hence the fractal dimension $\mathcal{D} = \log 2/\log 3 \approx 0.63$.

Thus it is immediately seen that Charlier's first criterion ($N \leq R$) means that the fractal dimension of such a universe is equal to or less than one. This criterion guarantees the absence of Olbers' and gravitational paradoxes. The latter one may appear in two forms, involving either infinite forces or potentials. Charlier's second criterion ($\sqrt{N} < R$) implies a fractal dimension less than two. This resolves Olbers' paradox and keeps gravitational forces finite, but the gravitational potential may be infinite and hence the speeds of stars much higher than observed.

In Fournier's model mass (number) increases proportionally to radius. Hence the exponent $\mathcal{D}$, i.e. the fractal dimension, is equal to one. It is interesting to calculate the fractal dimension from Fournier's original picture (Fig. 13.2). It gives $N = 5$ and $R = 7$, thus $\mathcal{D} = 0.83$. And if we add two elements in the third dimension of his figure, then the resulting hierarchy has just the critical dimension one!

When the fractal dimension is less than two, then the sky is not completely covered by stellar disks. Here one finds an interesting connection with rain clouds. There is a theorem about the projection of a fractal in 3-D space onto a plane: if the fractal dimension is larger or equal to 2, then the dimension of the projection is exactly 2. On the plane, dimension 2 means that the points cover the plane uniformly, without holes. The typical fractal dimension of clouds, consisting of water drops, is about 2.5. Thus the clouds can cast shadows on the ground without holes.

14.6 Creativity of fractals

The concept of the fractal has as it were liberated enormous energies for creating hierarchical objects which are very rich in form. The creativity of fractals, as Mandelbrot calls it, is emphasized already in the title of his first book *Fractals: form, chance, and dimension.* In contrast to simple hierarchies, which bring little surprise, fractals open the gate to a secret garden populated by an infinity of forms.

Complicated fractal shapes may be created from remarkably simple recipes. The most famous fractal, the Mandelbrot set, is generated in a complex plane by a simple mathematical process (the iteration $z \rightarrow z^2 + c$) which defines a procedure, where the output of the calculation is the input for the next calculation. Starting from one point in the plane and taking different values of c, different endless series of numbers are obtained which,

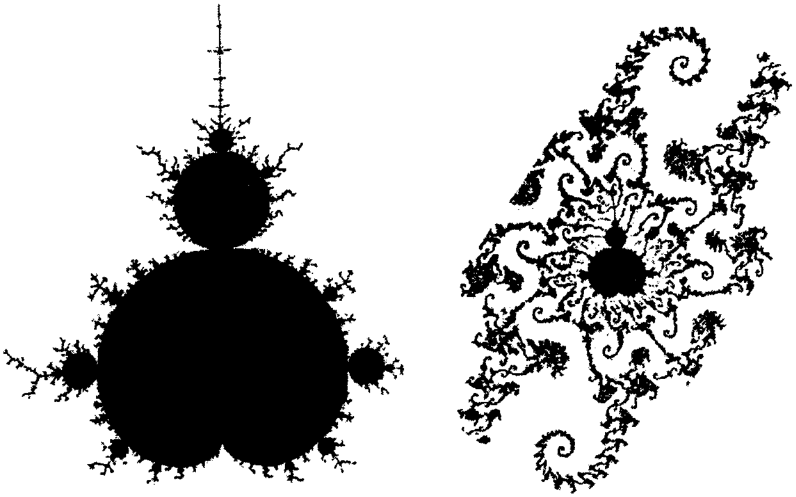


Fig. 14.5 The Mandelbrot fractal is, in the words of its discoverer "an astonishing combination of utter simplicity and mind-boggling complication". The right hand pattern is a highly magnified tiny detail of the infinitely fractured border of the left image. Inside it one may recognize the initial image.

according to a certain rule, will define a set of points. [‡]

The border of the resulting image contains infinitely small details. Looking at any one structure through a microscope, one finds new, smaller fascinating details. They are not exact copies, but are variations on the same theme. The Mandelbrot set has been called the most complex object ever seen. Its infinitely rich structures pose a challenge for mathematicians. There was a conjecture by Mandelbrot himself that the border of his set has the fractal dimension two. This was proven around 1990 by Mitsuhiro Shishikura, a mathematician in Kyoto. The Mandelbrot set would be invisible without computers, because making even a rough image of it, one needs millions of calculations.

[‡] The numbers are complex ($z = a + ib$), but the process is quite simple. Take the initial point $z = 0$ in the complex plane and take a number c. Then calculate $z_1 = (0)^2 + c = c$, and go on with $z_2 = z_1^2 + c$, and $z_3 = z_2^2 + c$ *etc.* If the sequence 0, c, $c^2 + c$,... keeps within a bounded distance from the origin forever, then the chosen point c is a member of the Mandelbrot set. But if the sequence diverges from the origin, then the point does not belong to the set. Repeat this for as many c as you like.

14.7 Random fractals and Brownian motion

Above we considered regular fractals which are based on regular iteration rules and on exact self-similarity, i.e. sub-elements are exact copies of the elements at the previous level of the hierarchy. They are important for mathematical studies of fractal properties, but of course they are too rigid as models for natural processes.

Random fractal structures are free from strict regularity. Their construction is more flexible, utilizing iteration rules which include probability laws for choosing the value of fragmentation or aggregation parameters. For example, the Cantor dust may be randomized so that one casts the die at each step of the construction, thus choosing randomly which part of the line segment is thrown away. An important special case of random structures is a uniform Poisson distribution of points. It has $\mathcal{D} = 3$ and it represents what is called the uniform distribution of any of the discrete objects in Nature (molecules, galaxies...).

A good example of random fractals is offered by the famous Brownian motion, named after the Scottish botanist Robert Brown who discovered it in 1827 when he looked through his microscope at pollen grains flowing in water. These small particles underwent rapid jumps in different, random directions. But it took almost a century before this phenomenon was explained. A particle is from time to time kicked by molecules surrounding it. As a result, the particle moves erratically.

Robert Brown 1773-1858

The most amazing feature of the Brownian motion is its self-similarity. If one increases by ten times the magnification of the microscope, one again sees a quite similar erratic motion, but with the jumps ten times smaller. In fact, the Brownian motion "draws" a random fractal curve!

Random fractals have the property of statistical self-similarity. This means that only on the average the different steps are the same. But how then to calculate the fractal dimension? Usually physicists measure the so-called box-counting dimension. They cover the fractal structure by a net of boxes and count the number of boxes occupied by a part of the structure. This is repeated for smaller and smaller boxes. When one plots, in logarithmic scale, the number of occupied boxes versus the size of the box, the self-similarity is seen as a straight line and its slope gives the fractal dimension. For the Brownian motion the fractal dimension is two.

14.8 Percolation – a process leading to fractals

When matter condenses, fractals may form. It is useful to look at a simple model describing an often encountered process, so-called percolation (from the Latin word 'to flow through'). Consider a large square net or lattice. Each knot of this net may be bright (i.e. occupied) or dark (i.e. empty). The probability that a knot is bright is P (less than 1). One can say that P is the brightness density of the net. If there is a continuous lighted path (a chain of bright neighbors) from one edge of the net to the opposite edge, then it is said that the lattice percolates. Such a process occurs in a coffee machine when water percolates through ground coffee.

Let us light up randomly more and more knots. The smallest brightness density (probability) for which the infinite lattice percolates is called the critical density or percolation threshold (for a plane lattice this is about 0.59). At the moment of percolation threshold, lighted fractal clusters form in the lattice. For plain lattices, the percolating clusters have the fractal dimension $\mathcal{D} \approx 1.9$ and in three-dimensional space, the critical fractal dimension is very close to two.

14.9 Fractal structures versus smooth distributions

Let us inspect two examples of quite different structures consisting of points in space. They help one to understand the difference between rough irregular fractals and smooth regular structures. This picture is useful, when we later on describe the galaxy universe.

The top left panel of Fig.14.6 displays a distribution of points made from a background of constant density plus an overdensity in the center. Apart from the small scale granulosity the structure appears rather regular and may be characterized by a few parameters: there is a well defined value for the background density, and one can identify a position for the structure (it is located at the center), it has a certain width and a certain height (amplitude) which can be easily estimated. One can define a density profile as shown in the top right panel. This profile can be well approximated by a smooth curve, in this case a constant plus a Gaussian "bell" curve. If one studies the dynamical evolution of this structure, taking into account the interactions between its points, one can rely on usual analytical mathematics. From this perspective the structure is essentially described by three

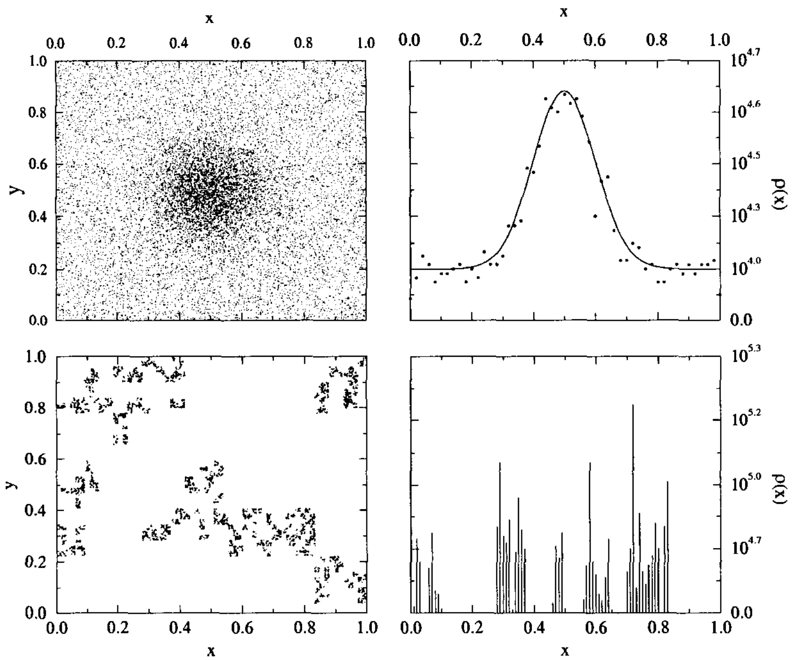


Fig. 14.6 A comparison of regular and fractal structures, as visualized by Francesco Sylos Labini. Above: a simple regular cluster on an uniform background and its density profile; below: a complex fractal distribution of points in a plane. In this case the density profile is different for different chosen scales within which the density is calculated.

elements: position, size, and amplitude. The typical result of such a study is to understand the motion and deformation of the structure. This is the traditional approach to the study of structures based on the assumption of regularity which had been implicitly adopted in statistical physics before the advent of self-organized critical phenomena a few decades ago.

In the bottom left panel of Fig.14.6 we instead show a strongly irregular structure for which all the concepts used to characterize the previous structure loose their meaning. There is no background density, there are structures in many zones and on various scales, but it is not possible to assign them a specific size or amplitude. This situation is also illustrated by the density profile (bottom right) which is highly irregular on any scale. In order to give a meaningful characterization of the properties of this structure, one has to look at it from a new perspective. This structure, a simple random fractal, also has its regularity, but it is hidden in the self-similarity.

A simple fractal is characterized by one fractal dimension. But for many things in Nature this is a too crude description. And as if the simple fractal were not subtle enough for the modest organ of imagination that most of us posses, we are now confronted by the *multifractal*. Multifractals emerged as a new concept in 1969 in Mandelbrot's study of turbulent phenomena. More recently they have been applied to understand the volatility of stock prices and the distribution of positions and luminosities of galaxies. [§]

Multifractality is also a step into a more realistic natural fractal. For example, galaxies are not mathematical points and furthermore, they are far from identical. There is a wide distribution of luminosity and mass in the galaxy population, and the multifractal is the mathematical instrument for studying such a complex system. In the simple fractal case one refers to the properties of a set of points and one needs only one fractal dimension. In the complex case, when the fractal properties can be different for galaxies having different luminosities, one has to introduce a continuous set of fractal dimensions to characterize the system, i.e. a multifractal.

14.10 Fractal view of Nature

It is a miracle which brings to mind Plato's concept of heaven, how abstract mathematics always seems to find correspondence with "real" natural phenomena. Integers, irrational numbers, imaginary numbers, n-dimensional spaces, curved geometries – all lie at the foundation of important physical theories. And now the entities having fractional dimensions join their company. Also fractals come in two categories:

- *ideal*
- *natural*

Ideal fractals exist in two forms – regular and random. These are mathematical objects with an infinite number of elements and iterations. Natural fractals appear in the structures and processes of Nature and usually occupy a finite volume with a finite number of elements. They are generally better modelled by random fractals than by regular ones.

[§] The term 'multifractal' was first used by Uriel Frisch and Giorgio Parisi in an article on "Fully developed turbulence and intermittency", in 1985.

Many nice fractals occur in diffusion processes. The growth of crystals and snowflakes are familiar examples. Another everyday fractal appears when cream is poured into coffee, forming patterns called "viscous fingers". In his book "Snowflakes and world views" (unfortunately available only in Finnish), Raimo Lehti traces the history of the study of the snowflake. The origin of its symmetric, lace-edged, fractal-like structure, which also shows so much variations on the same theme, has puzzled thinkers from Kepler to Descartes to Newton and many others. Nowadays it is a popular example of natural fractals. But of course to say so does not *explain* why the snowflake is what it is.

In Chapter 2 we mentioned that one of Kepler's feet was rooted firmly in the past, and snowflakes offer an illustrative example. Although Kepler also searched for physical reasons, in our sense, for various phenomena, he was strongly influenced by the idea that inner mathematical archetypes, Plato's regular geometric forms emerge in Nature. Sometimes these appear as a result of the concrete situation, e.g. bees build their honeycomb in the form of a rhombic dodecaedron, because this is a reasonable choice (it is a very economic way to use the valuable bee wax). But Kepler believed a geometrical form may burst out just as a reflection of a soul or an animating principle inherent in a thing. He could not figure out any physical necessity for the six-cornered starlets of snow, so he turned to the properties of the soul of the Earth which was reflected in meteorological phenomena. For us it is odd, but for Kepler it was natural to think that the icy flowers decorating his window glass on a winter day had formed with the *aim* of having such charming patterns! Along this line of thinking, complex natural fractal structures would be something that a phenomenon aims at reaching as a part of natural design. But a modern physicist sees the snowflake as a complex result of elementary forces.

Lehti points out that understanding the snowflake was impossible, when we lacked the knowledge of the fundamental physical phenomena on which the growth of a snowflake is based. Only after the theory of water molecules and icy crystals and other physics – issues of the 20th century – did it become possible, utilizing iterative methods and fractals, to attempt describing the birth and patterns of real snowflakes. Nevertheless, Euclidean geometry is still much involved, because the seed of a snowflake's outer appearance is the regular symmetry of invisible water molecules, which suffice to explain the simple growth of crystals in such ideal conditions when the surface tension is the dominating force. But it is the presence of heat diffusion which creates the porous, needle-like fractal surfaces during a snowflake's one hour descent from the heavens.

Fractals are powerful tools in physics, chemistry, computer science, geography, economics, and even in music and art. In recent years the top journals for scientific breaktroughs, *Nature* and *Science*, have been an arena for a new debate around fractality – on the issue of whether even the living world is governed by laws based on fractal geometry! Although fractals may be applied to so many phenomena, one should say, in the words of Mandelbrot himself, that fractals are not a panacea. Complex phenomena still continue to be complex. But at least we now have a new tool for probing the rugged surface of reality and measuring the hidden order in it.

14.11 The fractal dimension of abstract art

Most people learn about fractals from the attractive, often hopelessly enigmatic pictures generated by mathematical methods and computers. Indeed, fractal geometry has a message for the eye, as had the old regular shapes of Euclidean geometry. The simple mathematical formulae defining a fractal, only burst into life when given a visual representation. Fractal graphics created with computers present mathematical fractal shapes, this art faithfully following the orders of the formula.

As artists are the professionals of both what is seen in Nature and what is pictured, one may look at their creations from a fractal viewpoint. Nature being rich in fractality, it is no wonder that a "realistic" painter can catch it, like Leonardo da Vinci in his detailed pictures of turbulence or Aivazovskij in his breath-taking sea views (Chapter 4). But at a closer look also "abstract" art, intentionally distanced from our concrete experience, contains elements of fractality, and again this is not quite unexpected, as abstract artists work with geometrical forms and colors and experiment with processes that have their kinship in Nature.

It is in fact striking to read from an essay of a student of abstract art, written by Jean-Clarence Lambert in 1975 when the modern concept of the fractal was just emerging elsewhere, how "It seems that ... they [abstract artists] have succeeded in capturing on canvas something from the cosmos itself. They show, though not explicitly pictured, galaxies from interstellar space and playful shadows of a beautiful foliage; subtle games of clouds and the mute rise of the sun on the desolate sky; the swinging movements of a

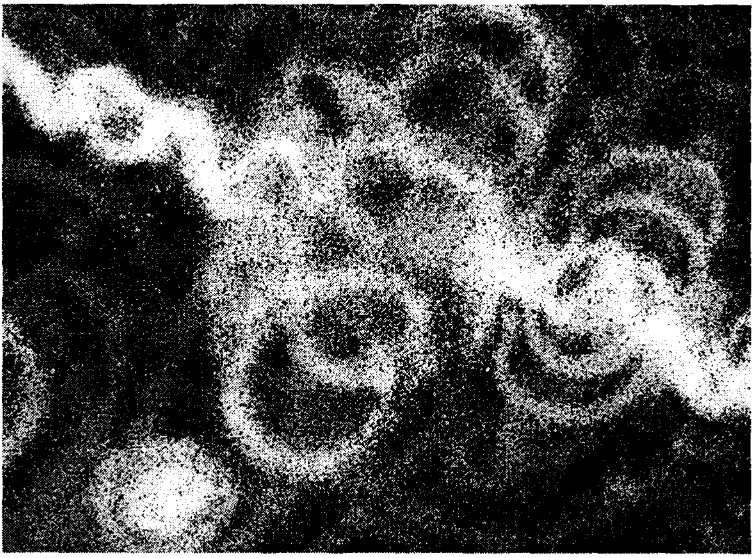


Fig. 14.7 "The Dance of the Plasma" by Stephen Goodfellow shows an artist's view of the origin of complex structures in the universe. Plasma, a hot magnetized gas, naturally creates fractals. The original color painting is housed at Tuorla Observatory.

piece of rush in the river, the eddies of the water;..." Many of these things are examples of natural fractals.

Jackson Pollock 1912-1956

Jackson Pollock was born two years after Vasilij Kandinskij, the Russian founder of abstract art, started his famous "Compositions". Pollock is regarded as the first spontanist in abstract art, who utilized "randomness" in his work. He abandoned the easel and put the canvas horizontally on the ground. He let a stream of paint drip from a stick or some other thing onto the canvas, while he at will moved around it. Thus Pollock's paintings are made from criss-crossing trajectories, resembling something chaotic.

What is the secret of Jackson Pollock's charm? In 1999 an interesting study by three physicists was reported in *Nature*. They had made a fractal analysis of Pollock's drip paintings, calculating the fractal dimension of the patterns on the canvas. This was done using the "box-counting" method. The studied painting (or rather its photograph) was covered by a mesh of identical squares and the number of squares that contained part of the pattern was counted. This was repeated for different square sizes. If Pollock's pattern were made of simple, non-fractal curves, then the number of squares intersected by a curve would increase simply as $1/size$, as the square size is decreased. But the number increases more quickly, which means that the pattern has a fractal dimension larger than one.

Actually the study revealed two fractal dimensions, one for small lengths on the canvas (less than a few centimeters), and another for longer lengths. In different paintings and at short lengths the dimension was always smaller than at long lengths, where it could be almost 2. So the painting *Blue Poles*, which was bought by the Australian National Gallery in 1973 for the sensational price of 2 million dollars, has the fractal dimensions 1.72 and 1.96. The analysts concluded that the two fractal dimensions reflect two kinds of movements involved in Pollock's technique: the dripping fluid motions (giving rise to small scale structure) and his dashes around the canvas (larger scale structure). Both processes can introduce chaos and fractality of its own kind to the pattern.

* * *

Fractals appear truly everywhere on Earth. Benoit Mandelbrot likes to tell how in medieval cathedrals his attention is drawn to the architecture displaying self-similarity. Sometimes he may see fractals under his own feet, on the decorated floors of churches, sometimes he raises the eyes towards the paintings on the walls. The human mind and hand can create fractals, and with particular vigor and subtleness this is done by all-embracing Nature. And not only here in the sublunar world, but also in the realm of celestial bodies.

Chapter 15

Fractal and chaos: planets, stardust, dark haloes

The success story of fractal geometry shows that this new concept is not just temporarily fashionable, but has given birth to a new level of comprehending Nature. It has become possible to study quantitatively phenomena which previously were considered too complex and irregular for the methods of exact science, and hence not even offering anything interesting to science. Fractals are quickly colonizing the space of the Solar System, the substance between the stars, and the dark massive halo of the Milky Way.

Chaotic behavior of many dynamical systems ruled by regular forces was discovered about the same time as fractal geometry. In fact, fractals and chaos are tightly connected. Sometimes fractals may be regarded as instantaneous pictures of chaotic motion, as in turbulent water, an eternal and deep image of chaos. Generally, fractals hide in the mathematics of infinity describing a chaotic system, invisible to the eye. The new theory of chaos is developing rapidly and finds more and more applications in science.

* * *

15.1 Order and chaos revealed by the Solar System

For two thousand years the harmonic motions of the heavenly spheres served as a great manifestation of strict order in Nature. After Copernicus's reconstruction of the planetary spheres, Kepler's discovery of the elliptic orbits, and Newton's great idea to base all physics on the universal laws of gravitation and motion, the orderliness of the Solar System became a new scientific triumph. It was possible to predict the motions of celestial bodies from natural laws. Nowadays the planetary surveyors are launched with confidence to the planets, asteroids and comets – their very exact trajectories leading them right to their targets, computed using the Newtonian laws. This apparently confirms the famous proclamation by Laplace about the clockwork universe in which the entire future may be calculated precisely.

However, it happened that the same Solar System also revealed phenomena which are the Nemesis of order – chaos, irregular, nonpredictable motion. This surprising generation of chaos from order dispersed the Laplacian dream. The mathematical research into chaos began when Henri Poincaré posed the question about the stability of the Solar System. He discovered the chaos in the orbital motions of three bodies, caused by their mutual gravity force. One cannot exactly predict their future trajectories.

In his *New Methods in Celestial Mechanics* from the year 1892, Poincaré describes the complex mathematics behind three-body trajectories: "...these intersections form a type of trellis, tissue or grid with infinitely fine mesh. Neither of the two curves must ever cut across itself again, but it must bend back upon itself in a very complex manner in order to cut across all of the meshes in the grid an infinite number of times. ...One must be struck by the complexity of this shape, which I do not even attempt to illustrate. Nothing can give us a better idea of the complexity of the three-body problem." In fact, one reads here an attempt to picture the fractal structure of a strange attractor – when you put three particles in motion, you cannot know in practice along which of the infinitely densely intermingled orbits your system will start to evolve.

Gösta Mittag-Leffler 1846-1927

Poincaré's discovery was a result of years of hard work, partly put in motion by a substantial prize, established by King Oscar II of Sweden and Norway, to be awarded to the first person who obtained the general solution of the n-body problem. It happened that the problem itself remained unsolved, but another mathematical treasure was unearthed... A key person in instigating this competition was Gösta Mittag-Leffler, a prominent Swedish mathematician and the editor in chief of *Acta Mathematica.* He worked four years as professor at Helsinki University, before accepting the chair of mathematics at Stockholm University. Nowadays Mittag-Leffler's former residence in Djursholm, peacefully placed in the outskirts of Stockholm, houses the research institute of mathematics which bears his name.

There is a convoluted story surrounding Poincaré's winning the contest in 1889. Poincaré found an error in the article which won him the prize but only after he had submitted it. Several months of further intensive work led

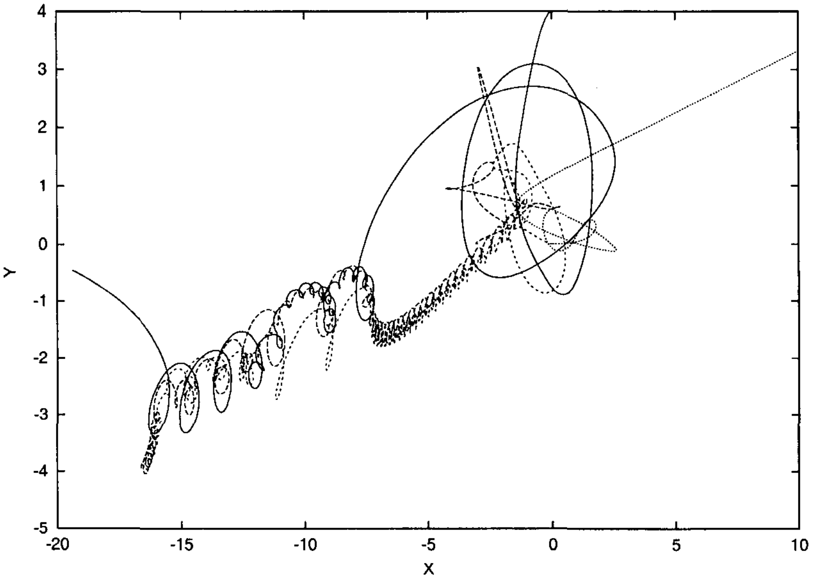


Fig. 15.1 Chaotic trajectories of four gravitating bodies which started their motions around the (0,0)-point. After some revolutions a strong interaction occurs which ejects one body to the right, and a pair and a single body to the left. The single one joins the pair, steels from it a companion and all three fly, spinning around each other, further to the left. Then the original single body is ejected from the three body system, while the pair is again united. This history came from a four-body computer experiment made by Seppo Mikkola at Tuorla Observatory. Generally the end results of such experiments are totally unpredictable, being extremely sensitive to the initial configuration and velocities.

him to the correct solution. It was this extra penetration into the subject which led him to the very discovery of chaos which he had overlooked in his actual prize essay on three bodies.

Newton lamented (or was glad) that the only problem that ever made his head ache, was the calculation of the Moon's orbit. Indeed, our faithful companion is attracted not only by the Earth, but also by the Sun, making it a member of a three-body system. Luckily it is not one of the most unpleasant ones – after all, we can fairly confidently consult the calendar for the date of the next full moon.

Our Solar System contains much more than three objects (the Sun, nine major planets, tens of their moons, thousands of asteroids, and other smaller stuff). This makes one suspect that there should be some chaos, unpredictability, around our home star. Where then is the chaos lurking?

A modern branch of celestial mechanics is fully devoted to investigations of the chaos in the Solar System. It is usually stated that a gravitating system is chaotic if the orbits of originally quite closeby masses increasingly diverge from each other so that in practice it is not possible to predict their positions after some time. In the Solar System the chaos appears as an erratic change of the orbit, for example its ellipticity may change dramatically. Small bodies, such as the huge number of asteroids populating the asteroid belt between the orbits of Mars and Jupiter, are known to display chaotic behavior. A sign of this is the absence of asteroids at some orbits of the belt. These "Kirkwood gaps" have been cleaned up by chaotic change of the orbits, leading to elongated trajectories intersecting the orbits of inner planets. In such irregular journeys asteroids may collide with each other, generating many small pieces. Thus the theory of chaos explains the meteorites falling to Earth!

The orbits of planets are in principle chaotic too, but it takes very long times before anything drastic happens. For example, computations show that Mercury may in a thousand billion years face the fate of a close encounter with Venus or a plunge into the Sun. Of course, this future is so distant that Mercury has more to worry about when the Sun begins expanding into a Red Giant in some five billion years...

We related in Chapter 9 that nuclei of galaxies generally contain compact supermassive objects, possible black holes, and in some of them there may even be two or more of these objects orbiting around each other. In some situations one or two of these may be expelled into the intergalactic space, as predicted by the Slingshot Theory originated by William Saslaw, Mauri Valtonen, and Sverre Aarseth. In recent years it has been developed especially by Valtonen's team at Tuorla observatory in Finland, as an alternative way to understand double radio sources ejected from active galactic nuclei. This is a process which is essentially chaotic in nature. An example is given in Figure 15.1 – just think that the "particles" in this four-body experiment have masses of a million suns.

15.2 Chaos, strange attractors, and fractals

20th century mathematics and physics discovered an amazing thing: even very simple physical systems, obeying strict laws, can display irregular, unpredictable motion. As we mentioned above, the first example was found by Henri Poincaré in the physics of three gravitating bodies. He is now hailed as the initiator of the dynamical system theory.

Generally speaking, any system obeying deterministic rules is called a dynamical system. * The discovery of chaotic behavior means that even such a system may be unpredictable in the long run, though it has a predictable behavior in the short term. In other words, there is a "predictability horizon" for a dynamical system. A necessary condition for chaotic behavior is the non-linearity of the system, which means that a change in the state of the system depends on its present state. † Many, if not all, real physical phenomena are described by non-linear dynamical systems, which are chaotic.

In the 1960's Edward Lorenz, a meteorologist from the Massachusetts Institute of Technology, discovered a surprising fact about the deterministic mathematical model used for making long range weather predictions with the computer. It lacks predictability! This epochal discovery opened the door to chaos for regular dynamical systems. Essentially, he found the "sensitivity effect", i.e. the fact that a small error in the input signal propagates with an ever increasing amplitude into the end result of the calculations and totally masks the original true signal.

The extreme sensitivity of the system to the starting values for the calculations characterizes chaotic systems. A popular illustration of chaos is the "butterfly effect": if the weather-man ignores the fluttering of a butterfly's wings in his calculations, his long term prediction will inevitably fail sooner or later. In other words, it is impossible to predict the weather... More exactly, it is said that the predictability horizon in weather forecasting is three weeks. Beyond that the proverbial "old wives" succesfully compete with the meteorologists at their supercomputers.

We prefer the above form of the butterfly effect instead of the oft heard

---

[*] Deterministic means that the evolution of the system is strictly determined by the physical causes and laws influencing its components. Knowing all these laws and the state of the system perfectly in all its details at one moment permits one to know – in principle at least – its state at any later time.

[†] For example, $y = kx + b$ is a linear deterministic law for which $dy/dx = k$ does not depend on x. But the quadratic law $y = kx^2 + b$ is non-linear; its derivative $dy/dx = 2kx$ depends on the value of x. This quadratic law is a basic example of a simple dynamical system capable of generating complex chaotic behavior. Here we can only mention that by investigating the simple "quadratic iterator" $x_{n+1} = ax_n(1 - x_n)$, Mitchell Feigenbaum discovered the universal route from order to chaos, governed by "period-doubling" and the Feigenbaum constant 4.6692....

one: The flap of a butterfly's wing could stir up a tornado in Texas. This gives a highly exaggerated significance to one small event among practically infinite number of other small events – it "could" stir up a tornado, but the probability is extremely tiny, and we can never know if it really did.

In order to describe the possible states of a dynamical system, mathematicians have introduced the concept of an *attractor.* It is the set of points (each point actually depicting the state of the whole system) which the system occupies during its whole dynamical evolution. For example, if the system is strictly periodic, like a frictionless pendulum or a planet circling a star, it is "attracted" into a regular succession of states and the attractor has a relatively simple structure. But if one adds suitable forces to the system, its time evolution may become extremely complex – in particular it will never return to a previous state (otherwise it would start all over again and be periodic). Its orbit from one state to another never intersects itself, but fills the allowed space in a complicated manner. In fact, this tangled pattern has a fractal structure and is now termed a chaotic or strange attractor. [‡]

Strange attractors and fractals are intimately related. As a geometrical pattern a strange attractor is a fractal, and as a dynamical object it is chaotic. One may say that chaos is a process and fractal is the geometry of chaos. In comparison, simple geometry is a hallmark of a regularly behaving dynamical system. For example, Kepler's elliptical orbits correspond to the regular motion of a planet around the Sun.

It is intriguing that the notions of chaos, strange attractors, and fractals are still without a final mathematical definition and the full mathematical understanding of these objects is a great challenge for future science.

15.3 How a pendulum connects chaos and fractals

Chaos is not just a question of practical calculation accuracy, but it has much deeper meaning related to the fundamental unpredictability of physical systems. Because of the endlessly dense distribution of numbers along the real number axis, there is no "accurate" initial value which could be ascribed to the position or velocity of a particle. Furthermore, along the

---

[‡] The term 'strange attractor' first appeared in the 1971 article "The nature of turbulence" by David Ruelle and Floris Taken. 'Chaos' – in the modern sense – appeared twelve years after Lorenz's discovery, in a paper by Tien-Yien Li and James Yorke.

endless path towards the unachievable correct initial value, the corresponding end results of calculation may fluctuate strongly, so that there is no convergence towards some definite state. Nevertheless, the chaos is not just "a state of confusion": after all, the evolution of a physical system is ruled by definite laws – a particle in a chaotic system does not "feel" anything mysterious, just the influence of ordinary forces. [§] Hence, the calculated end results for the system, though apparently "chaotic", actually reveal complicated order, fractality. In fact, the words by the physicist Joseph Ford on the mystery of chaos have a definite feeling of fractal within them: "It is a paradox, hidden inside a puzzle shrouded by an enigma."

The link between chaos and fractals is difficult to show without going into mathematical details, but it may be illustrated by a simple toy called a spherical pendulum. The pendulum is hung so that it may swing freely in any direction and has in its end an iron bob. Directly below it is a square of four magnets, painted green, red, yellow, and violet. Push the pendulum a little aside and then let it swing. Which magnet will trap the bob? If the starting point is close to any one magnet, then that magnet will finally catch the bob. However, if the pendulum is released close to the two perpendicular straight lines with the magnets symmetrically in the quadrants, then it is increasingly complicated to predict the final magnet.

One may decide to make repeated experiments, noting each time the initial bob position and also the colour of the trapping magnet. If one plots each initial position with the color of the corresponding end state, the region around the red magnet will be covered by red – the bob initially close to the red magnet always stops finally over it. The same is true for the regions around the other magnets, each becoming painted with the colour of their host magnet. But close to the perpendicular boundary lines there appears quite a complicated mix of differently colored areas. A small shift brings you to another colour, i.e. if you hold and release the pendulum, its ultimate destination will abruptly change from one magnet to another. And if you use sharper and sharper color pencils to mark the initial points, smaller and smaller color structure appears. In these boundary regions the outcome of the experiment is virtually unpredictable. There the infinite succession of color patterns is fractal, the geometry of chaos!

[§] There should exist a definite end state when the system starts from an *exactly* known initial state. In fact, in 1906-9 the Finnish mathematician Karl Sundman (1873-1949) proved the *existence* of a general mathematical solution to the three body motion.

15.4 "Protochaos" in Swedenborg's vision of evolution

We have here and now referred to Emanuel Swedenborg's early writings for the simple reason that they are a gold mine of interesting ideas coming down from the childhood of science, even if they did not much influence the development of science. It is apt to mention his views on predictability, evolutionary chains, and the diversity of worlds.

Swedenborg was fascinated by the possibility to gain by a scientific method, which included experience, theory and rational reasoning, knowledge of the world small and large. He gave as an example an ideal super-scientist who oddly reminds us of the almighty Newtonian intelligence later described by Laplace and whom we recall from Chapter 2. Swedenborg wrote: "Such a man would be capable of taking his station as it were in the center; and surveying from thence the whole circumference of his system at a single glance, he would be able to make himself acquainted with things present, past, and future, from a knowledge of their causes, and of their contingent given or supposed."

But Swedenborg acknowledges that such a super-scientist, continuously receiving and immediately analyzing all information, is just a dream: in reality Man is much less ideal and more helpless in his attempt to understand the world, to see the past and to predict the future.

Swedenborg's cosmos was very dynamic and evolving. It was full of planetary systems, "worlds", each having evolved into their present state along long chains of steps, believing "That no world can exist, abounding in any variety of objects and phenomena, without first passing through a succession of states and intervals of time." Though the laws or principles driving such an evolution are the same, the end results are varied. He gives an example in which 1000 steps were required to reach a present state. But in that chain even a small difference in a single step gave immediately rise to a new chain, and a quite different final state. There is not only much variety in the present "worlds", but also infinitely many chains of evolution.

"Now we may likewise have an equally great variety of series; for were there the least diversity in any intermediate change, it would immediately give rise to another, collateral and perfectly different series of things successively and simultaneously existing. ... there may be as many series as there are worlds; or as many worlds as there are series."

Swedenborg also touches the subject which we would call exobiology: "If ... in these earths we could suppose the existence of an animal kingdom of the same kind as our own, then we must also suppose it subject to the same contingencies, changes, modes and series, through which it must pass to arrive at the same perfection; but since we cannot presume that, in these respects, all other worlds are absolutely similar to our sun, so we cannot presume them to be tenanted by a precisely similar race of living creatures."

Swedenborg's cosmos was not only a static self-similar structure. It was dynamic and – in the modern sense – chaotic. For the Swedish thinker the evolutionary tendency into a great variety was the true hallmark of the universe. He wrote that the world itself is a miracle. Its beauty and perfection lies in the great number of modifications. And paradoxically (and now based on the deep reasoning of chaos) "in this therefore consists our highest wisdom, that we know how small is the extent of our knowledge".

15.5 To the microcosmos – and back to the planets again: Nottale's fractal space-time

Laurent Nottale, a French astronomer working at the Meudon observatory, has bravely speculated about fractality in the ultimate substance of modern physics, space-time itself. Classically, the coordinates describing space-time have been regarded as continuous and differentiable. In a remarkable series of articles Nottale proposes that this is not so and that there are scales of space and time where one must take into account the non-differentiability of space-time. Furthermore, he has demonstrated that a *continuous but non-differentiable space-time is necessarily fractal.* Fractality means structures at all scales or on a wide range of scales, and Nottale generalizes the principle of relativity so that *the laws of nature must be valid in every coordinate system, whatever their state of motion and of scale.* The state of scale is related to the space and time resolution with which the physical system in question is described. He calls the resulting theory *Scale Relativity.*

Though Nottale's theory is still developing and not yet a generally accepted part of physics, there are already many exciting views and predictions surfacing from the new formalism. It is concerned in particular with the frontier domains of modern physics, i.e. small length- and time-scales (microworld, elementary particles), large length-scales (cosmology), and long time-scales. The limit of long time-scales relates to chaotic phenomena, where after a certain time the system is shrouded in unpredictability and the information on its initial conditions is completely lost.

Fig. 15.2 Laurent Nottale pictured during the informal cosmological meeting organized by Georges Paturel in April 2000 at Lyon Observatory. In the warm atmosphere of a French spring, the discussions concerned different aspects of cosmology and fractals.

In the microworld scale relativity throws light on quantum phenomena which are now said to be various manifestations of the space-time fractality. In fact, understanding of quantum mechanics (small length- and time-scales) was the original motive for scale relativity. Then it was extended to macroscopic chaotic phenomena which can be described in terms of a non-deterministic, quantum-like theory, valid on very long time scales.

As to the astronomical observations, perhaps the most interesting – and controversial – prediction of Nottale's scale relativity is the universal structure of solar systems. With his equation for the probability density of planetary orbits around a star, Nottale has seemingly come close to the old analogy which saw a similarity between our solar system and an atom in which electrons orbit the nucleus. But now the analogy is deeper and mathematically and physically supported: it comes from the suggestion that chaotic planetary orbits on very long time scales have preferred sizes, the roots of which go to fractal space-time and generalized Newtonian equation of motion which assumes the form of the quantum Schrödinger equation. ¶

¶ "Actually my claim is that there are structures – 'quantizations' – in phase space..., that the position structures are well known even if they are still misunderstood, they are what we call galaxies, groups, etc., but that the corresponding velocity structures also

As in the usual theories, Nottale assumes that the planets are formed by accretion of planetesimals. But the orbits of these small building blocks do not fill randomly the plane of planets, but their probability distribution is given by the generalized Schrödinger equation. Nottale has presented evidence from our own and other planetary systems that planets orbiting stars seem to follow the predictions in a statistical sense. The predicted distribution of the orbits is quite simple and depends only on the mass of the central star and a universal constant. [‖] Nottale's formula gives the positions of the peaks of the probability distribution – individual planets are not predicted to follow exactly this law. Tens of exoplanets have been detected. With increasing statistics from future discoveries of still more extra-solar planetary systems it will be exciting to see if this profound new physical idea is consistent with observations.

Indeed, we have entered the epoch when other planetary systems have become a subject of observational study. Beginning from Swedenborg's hypothesis "On the universal solar and planetary chaos and its separation into planets and satellites", there have been general theoretical ideas pointing at the universality of planetary systems around stars. In the 1990's these were backed up by observations of dusty "protoplanetary" disks around young stars. And in recent years astronomers have been able, if not yet directly to see, but to "feel" planets circling their host stars (which they cause to slightly wobble as expected from Newton's laws). From the tiny movements of a star one may calculate the orbit period, the orbit size and the mass of the invisible planet.

Iosif Shklovskii 1911-1985

In their classic book "Intelligent Life in the Universe", published in 1966, Iosif Shklovskii and Carl Sagan discussed what they called "the assumption of mediocrity", the idea that our surroundings are more or less typical of any other region of the universe. For example, assuming that our Sun is an average star allows one to get a realistic picture of interstellar distances, even without advanced astrophysics. Sometimes it is risky to utilize the

Carl Sagan 1934-1996

---

exist and are nothing but the various Tifft-Guthrie-Napier...effects: but the problem is that present cosmology works with only half of the information, and has rejected the most important information since the basic quantization is in terms of velocity." (From a letter by L. Nottale to us.)

[‖] The preferential orbit radius/AU $\propto (mass/M_{sun}) \times n^2$, where n is an integer. Nottale suggests that Tifft's redshift quantization (Ch.7), can also be explained by his theory. The constant above contains the velocity 144 km/s, the multiples and submultiples of which have been claimed to appear in extragalactic systems.

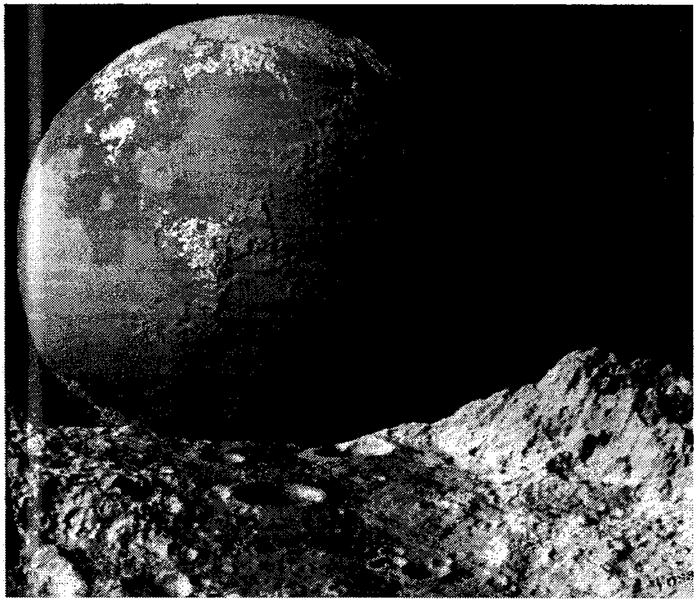


Fig. 15.3 "Souvenier from a space mission that never was", as Benoit Mandelbrot calls this impressive image. On the horizon of the fractal moon-like landscape there is an unknown exoplanet rising, possessing its own continents with coastlines of the universal fractal. Are there other eyes than ours looking at the cosmic fractals – other minds wondering about their meaning? This question is so natural to pose, after a long tradition of contemplation of the plurality of inhabited worlds. But only in our days has it become a subject of scientific study.

principle of mediocrity – we cannot with any confidence say that there are civilizations or even simple life around other suns. A single cell of a mundane organism is such a complex miracle that we simply do not know how far away we could find another, born independently of the origin of our life (that's why all that discussion on possible alien life on the nearby Mars is so important). But it seems that one striking prediction of mediocrity, aided by our general knowledge of celestial bodies, has proven true: planetary systems are common in space!

15.6 Rugged planetary landscapes

The moon was the first celestial body to which Galileo turned his brand-new telescope. In his *Siderial Messenger* of 1610 he described the observations showing that the Moon is not a perfectly smooth sphere, but has a rugged surface, with mountains, holes and deep valleys, but also wide smoother regions. The roughness on the Earth had found an unexpected rival in the celestial realm, a realm which had usually been thought of as the abode of regularity.

In fact, Anaxagoras, the early advocate of the assumption of mediocrity, believed that he saw plains and valleys on the surface of the Moon. His "terrestrial" view of the heavenly bodies was much influenced by the fall of a meteorite at Aegospotamoi. He was said to have held that the stone came from some celestial body on which there was an "earthquake". Now we know that the pock-marked appearance of the Moon was caused by impact of thousands of large meteorites. This bombardment occured with particular fervor when the young Solar System contained much debris left after the accretion of the planets from the protoplanetary disk around the Sun. For Anaxagoras the meteorite was an important message from another celestial body. We recall this vividly now when the two-kilogram ALH 84001, found in Antartica in 1984, is regarded as having been ejected from Mars some 16 million years ago as a result of an impact. About 13000 years ago its orbit led it fall onto the Earth. Of course, this is the famous stone which possibly carries traces of past Martian life. And only 200 years ago the mere possibility of stones falling from the heaven was ardently disputed by many scientists!

An early appearance of natural fractal in geography, before fractals were invented, was the measurement of the coastline length by the British physicist Lewis Richardson. He noted that the total length depends on the size of the measuring rod, and he discovered that this dependence is a nice power law. He had no explanation for this law. In fact, these measurements were published only after his death.

Lewis Richardson 1881-1953

Mandelbrot posed the proverbial question "How long is the coast of Britain?" and answered using Richardson's result that the coastline has the fractal dimension $\mathcal{D} = 1.25$. Our intuition would say that the coastline is really a line, i.e. has a dimension equal to one. Actually, its dimension is more than one. Figuratively speaking the coastline is no longer a line, though it is not yet a ribbon on the 2-D plane. After Mandelbrot opened the gate to fractals, a lot of other applications in earth sciences have appeared, among them clouds, mountain landscapes, and turbulent atmospheric flows.

Lewis Richardson did much more than just measured the coastline. This pioneer of numerical weather forecasting also paved the way for theoretical studies of turbulence, again topics linking him intimately to chaos and fractals. Turbulence in gases and fluids is characterised by unpredictable fluctuations in density and velocity in a cascade of whirls. Richardson's famous lines catch some of the complex physics in a way Leonardo's drawings of "turbulenzia" opened our eyes to the bubbling hierarchies of eddies: "Big whorls have little whorls, which feed on their velocity; And little whorls have lesser whorls, And so on to viscosity."

When one breaks a solid body, the resulting fracture surface is a fractal structure, similar to a rough mountain landscape. Both fracture surfaces and mountain landscapes have fractal dimensions around 2.3. They are no longer classical surfaces with $\mathcal{D} = 2$ as our common sense might tell us. Now when it is known that roughness on the Earth follows the fractal law, it is no longer surprising to hear that elsewhere in the Solar System one has found fractal structures too, e.g. in the craters on the Moon and planets and in the rings of Saturn.

The ejecta outlines of the craters on Venus were studied by a team of astronomers headed by Jouko Raitala from Oulu University in Finland. Following the classical method, they measured the length of the outlines with rulers of different size. Fig.15.4 shows the resulting "Richardson plot" for the huge crater Adivar, having a diameter of 31 km. The diagram reveals, as do similar plots for other craters, that there are two regions of scales, with differing fractal dimensions. On small scales $\mathcal{D} \approx 1.09$, while on large scales the dimension is ≈ 1.25. Fractal dimensions may be telling us about different geological or impact processes. Like the fractal dimension is Jackson Pollock's fingerprint in his action paintings, various processes in Nature leave their signature in the structures they engender.

15.7 Dense dust clouds – cocoons of stars

It has been estimated that in our Milky Way about ten new stars are born every year. Where do they come from?

There is much gas and dust between stars, with composition rather similar to that of the stars. It is this non-uniform interstellar medium that gives rise to new stars. The mechanism was already described by Newton in his letter to Bentley: a cloud tends to contract under its own gravity.

However, an interstellar cloud is far from simple. It is a mixture of gas and dust, has temperature, may rotate, and is permeated by magnetic

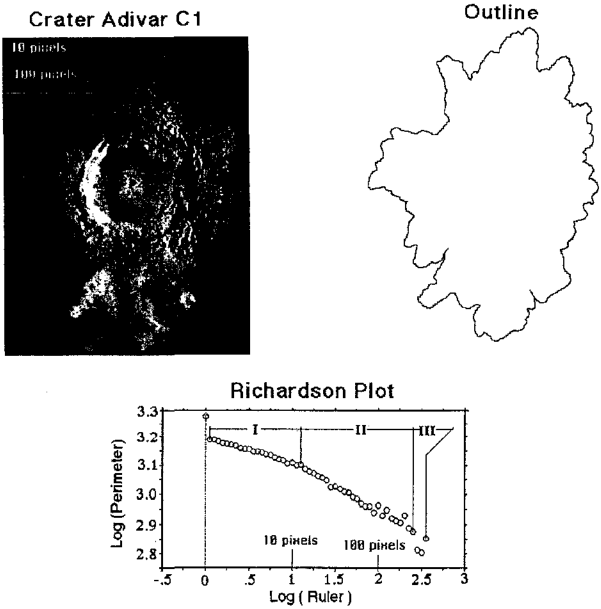


Fig. 15.4 An image of the crater Adivar on the surface of Venus and the analysis of the fractal properties of its ejecta outline (J. Raitala – Oulu University).

fields, filled by cosmic rays, and kicked by shock waves coming from supernova explosions. It even has internal fractal structure. Contraction into dense clouds is influenced by complex factors, making it hard to construct a theoretical sequence of what happens before a star is born. Neither is it easy to *see* the various phases involved. Contraction from a gas cloud into a young star takes from 100 000 to 1 million years. This is too slow a process for even generations of astronomers to follow from start to finish.

Nevertheless, astronomers think that in general outlines the birth of stars is understood. In the beginning there is a giant cloud of molecular gas and dust, one of the many found all around the disk of the Milky Way. The cloud contains denser regions which tend to contract under their own gravity. The contracting cloud is heated by the released gravitational energy and starts to shine even before it is a true star. Such a *protostar* is a strong source of infrared radiation coming from the heated dust surrounding it.

How does a true star differ from a protostar? When the temperature

Fig. 15.5 The Great Nebula in the constellation of Orion. This aggregation of gas and dust, which can just be seen by the naked eye, is a place where new stars are being born. One may see complex structures typical for many such clouds in the Milky Way

inside the contracting protostar reaches about ten million degrees, a new energy source is ignited and prevents further contraction: the major constituent of stars, hydrogen, starts to be converted into helium via a nuclear fusion reaction. Now one can speak of a true star, shining thanks to the released fusion energy and having for a long time the stability that we are accustomed to in our own Sun.

15.8 A case study of natural fractals: interstellar clouds

In Cosmological Letters in which he described how the band of the Milky Way made him think about a giant solar system, Johann Lambert also noted that its border is quite irregular, its width on the sky varying between 3 and 25 degrees, as if made up of pieces some of which seem to have broken away from the main body. Lambert wonders whether this irregularity actually contains harmony and order. And he concludes that it reflects

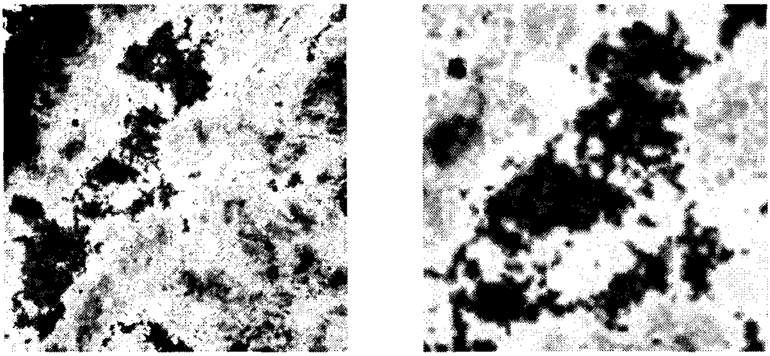


Fig. 15.6 Two images of the interstellar dusty medium in our Milky Way, in the direction of the constellation Taurus. In the left panel a 30 × 30 deg area. In the right panel new details in the heavenly "coastlines" emerge after zooming in by a factor of three. This image was obtained with the IRAS infrared satellite at 100 μm wavelength.

the existence of subsystems in the Milky Way. Indeed, perhaps the ancients were not so far from the truth when they named the milky ribbon in the night sky the Cosmic River with its jagged coastline and maelstroms.

On photographs taken by optical telescopes and maps made by radio telescopes or infrared observatories in space, the gas and dust of the Milky Way exhibit clear cloudy structure. Years ago it was usual to picture interstellar matter in terms of "standard" clouds, a few parsecs in size and rather similar to each other. Now we know that such a simplified picture lacks essential properties. The salient feature of cosmic clouds is that their edges are quite crinkly, and when they are studied with different magnifications, rather similar structures are again found. With a better look, the previously envisaged standard clouds evaporate into smaller clumps.

In other words, the appearance of interstellar clouds is reminiscent of fractals, which has been confirmed by detailed studies. These have mostly been concerned with the fractal dimension of the perimeter of the clouds. In order to measure this dimension, one needs to answer the question: How does the length of the perimeter depend on the observed area of the cloud?

One should first explain what one means by the perimeter of a fractal object, when – recall Koch's snowflake – the actual length of the curve may approach infinity. Take a meterstick of length ϵ and approximate the outline of the object as a polygon with side ϵ. Then its "epsilon perimeter" is the total length of the polygon sides. The length thus defined will depend on the meterstick ϵ, as Richardson noted with the coastline.

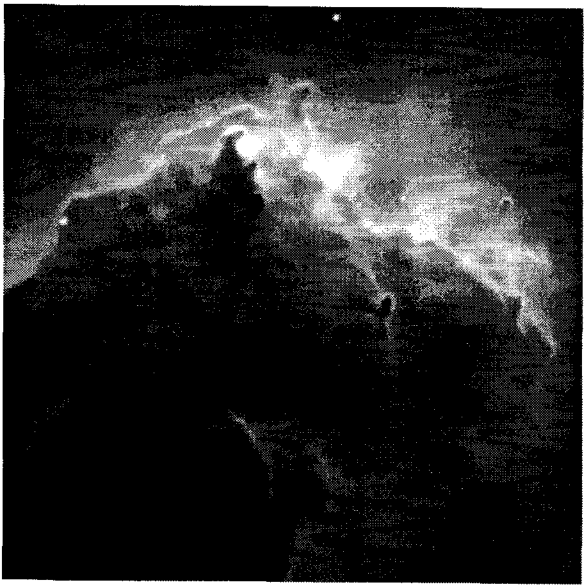


Fig. 15.7 The Eagle nebula in the constellation of Serpens. In this Hubble Space Telescope image one sees several globules, carrying embryos of stars.

There is a simple relation between the epsilon perimeter and the area (measured in units of ϵ^2) of fractal objects of different sizes:

$$perimeter = constant \times area^{\frac{\mathcal{D}}{2}}$$

The fractal dimension $\mathcal{D}$ of the perimeter is an indicator of how complex the outline of the object is. If the object is smooth and regular like, say, a circle or a square, then $\mathcal{D} = 1$, and the area grows as the square of the perimeter. If the perimeter is distorted from a normal line, its fractal dimension is somewhere between 1 and 2, and the epsilon meterstick "finds" increasingly finer details when the size of the fractal object is increased.

15.9 Dark clouds, molecular complexes, cirrus filaments...

Perhaps the first astronomer who undertook such a study of perimeters of interstellar clouds was the Canadian Martin Beech. In 1986 he analyzed dark clouds, i.e. dust clouds which are seen in the sky because they block from view stars which lie behind them. He found a clear relation between their perimeters and areas. It deviated from what is expected for regular edges (i.e. $\mathcal{D} = 1$), and suggested that the fractal dimension is about 1.4.

After this result on dark clouds, other kinds of interstellar clouds have been similarly studied by astronomers. These include infrared "cirrus" clouds (first detected by the space observatory IRAS), molecular clouds, and clouds of hydrogen. It has come as a surprise that all kinds of clouds seem to exhibit similar fractality, with the dimension of the perimeter around 1.2 - 1.4. The self-similarity is observed in a very wide size range, from small sizes of 0.0001 pc to large clouds with diameters of 100 pc.

The interstellar clouds are in many ways extremely different from the patches sailing in our atmosphere. Nevertheless, to refer to them by the terrestrial name 'cloud' seems to be quite apt, since our clouds show fractal structure, too. Shaun Lovejoy, at the time a post-doctoral student of Mandelbrot, measured from radar and satellite pictures in 1982 that the outlines of rain clouds have fractal dimension $D \approx 1.35$, quite similarly to their 30 000 billion times larger relatives in interstellar space.

The above dimension refers to the perimeter of the clouds, which is easy to measure. A more fundamental quantity would be the fractal dimension of the clouds themselves, telling how they are composed of sub-units and how they fill space. Unfortunately, astronomers cannot make accurate 3-D maps of the spatial arrangement of interstellar clouds. However, there is a theorem in fractal geometry, which may be of help.

If one takes a slice through a fractal object with $\mathcal{D}$ larger than or equal to 2, the dimension of the resulting curve is $\mathcal{D}-1$. Though the observations of clouds do not represent such slices, but projections of the whole cloud on the sky plane, one may suppose that the dimension $\mathcal{D}$ of the cloud itself is about the perimeter dimension plus one. In fact, there is some experimental evidence for this. For instance, Beech has investigated crumpled sheets of paper, objects so familiar to any writer. They are fractals with $\mathcal{D}$ around 2.4. The shadows cast by such paper balls on a screen have outlines which are characterized by perimeter dimension ≈ 1.4. This delightful experiment made Beech conclude that the fractal dimension of interstellar clouds may be around 2.4. The crumpled paper ball is a rough picture of the clouds which are not at all uniform, space-filling structures. **

**Recently, Bruce Elmegreen and Edith Falgarone have shown by actually studying the distributions of sizes and masses of the interstellar clouds that the fractal dimension is about 2.3, in agreement with what might be guessed from the outlines of the clouds.

It is still a matter of conjecture why different kinds of interstellar and even terrestrial clouds have such similar shapes. It is usually thought that this has something to do with turbulence, again showing the connection between chaos and fractals. Another kind of model for making fractal interstellar clouds from small clumps was developed by Peter Tarakanov at St. Petersburg University. He considered a process in which small clouds are ejected from stars, together with the "stellar wind" and aggregated into large structures. These turned out to be fractals with a dimension of 2.35!

15.10 Galaxy haloes – dark mass hiding in fractals?

Recall that galaxies are surrounded by massive dark matter haloes, as betrayed by the swift rotation of the outskirts of spiral galaxies and by the deflection of light coming from more distant objecs. But the structure and composition of the dark haloes and even their size remain a mystery.

A very interesting fractal model for the massive haloes of galaxies has been created by Daniel Pfenniger of Geneva Observatory and Françoise Combes of Paris Observatory. They suggest that the dark matter surrounding spiral galaxies is essentially baryonic and in the form of cold molecular hydrogen. Of course, this does not preclude the existence of other forms of non-baryonic matter. According to their model, the cold gas forms a fractal structure with dimension $\mathcal{D} \approx 1.6$. The smallest fragments in this structure, called "clumpuscules", would have a mass comparable to that of Jupiter, but be about the size of the Solar System.

The fractal haloes would be flattened like the observed gaseous disks of spiral galaxies. Remember that according to big bang nucleosynthesis 90 percent of baryonic matter is dark and we do not know its composition. The cold gas remains one of the best candidates for such dark matter now that compact objects like brown dwarfs or small black holes have been ruled out by microlensing experiments searching for MACHO's (Ch.10).

Because of the coldness of the gas and because of its fractality, it would be difficult to detect observationally other than via its gravitational influence. Fractality ensures that a large part of the gas is hiding in small-sized, compact clouds which cover a small fraction of the sky.

If true, this hypothesis leads to a new view on how galaxies are born and how they evolve. For example, it is a part of this picture that the dark mass provides continuously fresh gas for star formation in galaxies and that

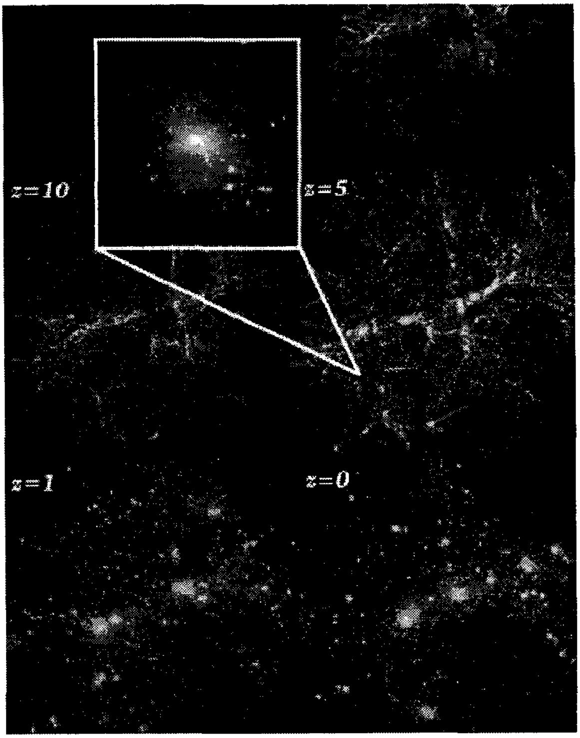


Fig. 15.8 The hierarchical formation of an artificial "Local Group" consisting of two main galaxies ("milky way" and "andromeda") surrounded by fractal haloes. From a high-resolution, N-body computer simulation by Ben Moore et al. showing how primordial cold dark matter seeds evolve to structures like the Local Group.

there is a gradual evolution from one galaxy type to another (Sc → Sb → Sa). The common view is that different spiral types were determined long ago, already in their protogalaxy stage.

Recent cosmological computer simulations of how galaxies were formed give new insight for the structures of the massive galaxy haloes, made of some sort of cold dark matter. The high resolution N-body calculations using 3 million gravitating "dark matter particles" reveal that the resulting haloes around the galaxies of the present epoch are built from hierarchically organized sub-elements (see Fig.15.8). Such calculations are very demanding even for supercomputers and one could not reach hierarchy levels below the minimum mass of one million Solar masses. The simulation of smaller details of the dark mass halo requires more powerful computers.

15.11 Fractal gas clouds between galaxies

A natural next step would be to ask whether intergalactic gas clouds are also fractals. But before that it is advisable to know if there are any gas clouds in the space between galaxies at all. Namely, our knowledge of the intergalactic medium and its different forms is still so scanty, there is no data available even close to the rich observational material on interstellar clouds in our Milky Way. Nevertheless, one kind of intergalactic gas cloud is revealed by their fingerprints in the spectra of very distant quasars. Such spectra, numbering presently over 200, show the so-called Lyman-alpha absorption lines of neutral hydrogen, which originate at different redshifts between us and the quasars. These lines may be so many and dense in one spectrum, that one speaks of a "Lyman-alpha forest". It is possible that these gas clouds (which we see as they were billions of years ago) have been left over when galaxies were born, or some of them have since collapsed and formed galaxies.

Vitalij Gorbatskij, astronomy professor at St. Petersburg University, has analyzed how the widths of the Lyman-alpha forest spectral lines are distributed. He managed to show that the clouds are composed from some elementary "cloudlets" in a fractal manner, so that the fractal dimension is roughly 2.5. The uncertainty in this determination is still large, and it may well be that the true value is somewhat smaller, close to that of the interstellar and atmospheric clouds.

* * *

Our home galaxy is filled with phenomena which display both strict order, bubbling chaos, and complex fractals. The Solar and other planetary systems are prototypes of order, but they are also generators of chaos, especially among their swarms of minor inhabitants. The space between the stars is an ocean of gas and dust, from which stars with their planets are born and when they die, they enrich it with the elements produced during their lives. Islands of dense molecular clouds, the cocoons of future stellar nuclear reactors, are themselves "chemical factories", where complicated molecules are formed. This happens in conditions where turbulence, self-gravity, and hierarchical fragmentation act in concert. Let us now go further, leaving behind the stormy Milky Way with its billions of stars, into the tranquil depths of the extragalactic world with its myriads of galaxies.

Chapter 16

Redshift – the quiet cosmographer

The redshift was the first, unanticipated discovery in the virgin forest of the galaxy universe. The expansion of space is its generally accepted explanation. But independently of any such interpretation, the accurate Hubble law (redshift proportional to distance) now serves as a powerful tool for exploring the 3–dimensional cosmic structures. For this task, astronomers are working hard at their telescopes measuring redshifts for thousands of galaxies. Step by step, they are penetrating deeper into space, making cosmic maps for regions "to boldly go where no one has gone before".

And they are excited by the views beyond the homely Local Group of galaxies, the views which are opening into the closest Supercluster in Virgo. Long filamentary structures, great walls made of galaxies, and giant voids form the rugged landscape of our cosmic neighborhood. In the middle of this silent and outlandish territory, one suddenly realizes that the map maker – the Hubble law – should not operate here at all! There are nowhere plains, no uniformity, which as we think gives rise to the linear law of redshifts, our distance indicator. The quiet cosmographer has its secrets...

* * *

16.1 Hubble's law of redshifts is a distance indicator

Edwin Hubble found his law of redshifts using 24 galaxies. The most distant ones were members of the Virgo cluster, at a distance of about 20 Mpc (or 25 times the distance of our near neighbor, Andromeda). What do we know about the Hubble law today, some seventy years later?

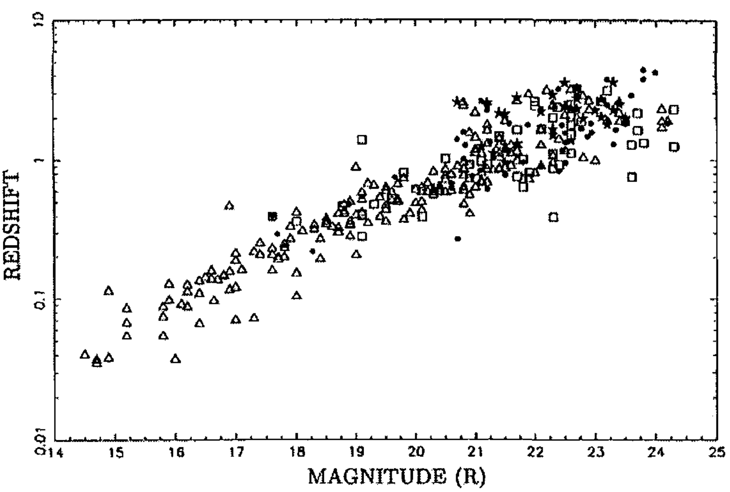


Fig. 16.1 The Hubble diagram for giant elliptical galaxies which are also double radio sources like Cygnus A in Fig.7.5. The Hubble law is visible up to $Z \approx 3$, whereas in Hubble's discovery article the largest redshift was ≈ 0.03! (The diagram by T. Pursimo.)

Milton Humason 1891-1972

After Hubble's "brightest stars", many new standard candles (or ways of inferring cosmic distances) have been discovered. Hubble had an observer colleague Milton Humason, the legendary astronomer who had started his career as a mule driver at Mount Wilson where the world's largest telescope was located. They found that the brightest members of galaxy clusters are apparently quite similar galaxies, almost "twins". Because they are very powerful emitters of light (contain many stars) they can be seen from large distances. Modern studies with larger telescopes and more sensitive detectors have confirmed that the tight relation between redshift and distance continues at least up to redshift ≈ 1, which means a distance two hundred times greater than in Hubble's discovery study. Supernovae, radio galaxies, and quasars give evidence for the Hubble law at still larger distances.

The Hubble law offers astronomers a powerful method to measure distances to galaxies, because the redshift is much easier to measure than anything else for most galaxies. Then the distance is the redshift multiplied by the speed of light and divided by the Hubble constant ($R = Zc/H$). And even if one is uncertain about the value of the Hubble constant, in this fashion one may always derive *relative* distances at least (e.g. that the distance of one galaxy is twice the distance of another).

16.2 The Hubble constant measured before Hubble...

Redshift alone gives good relative distance measures. But the astronomer often wants to know the absolute distance in a standard length (such as the megaparsec). For this one has to know the value of the Hubble constant H accurately, which, so to speak, fixes the cosmological distance scale. It is also needed for the calculation of the age of the big bang universe. No wonder that its accurate value is a big thing for astronomers.

In principle, it is easy to measure the value of the Hubble constant. Just take a galaxy, measure the redshift of its spectrum, and also measure its distance. Then multiply the redshift by the velocity of light in order to get the recession velocity and finally divide by the distance ($H = cZ/R$).

Georges Lemaître 1894-1966

The first estimation of the Hubble constant was made in 1927 by the Belgian astronomer Georges Lemaître, who was also a Roman Catholic priest. He determined it even before the Hubble law was discovered. How was this possible? In that year Lemaître published an article, in which he showed independently of Friedmann that Einstein's equations for a universe filled uniformly by matter allowed a dynamical solution. This article also introduced Lemaître's law which connected redshift to the scale factor at the moment of the light emission (Ch.8). He also predicted that in an expanding universe the redshift should at small distances grow directly in proportion to distance.

Lemaître had at his disposal redshifts and apparent magnitudes for a few tens of nearby galaxies. From their magnitudes, he estimated that their average distance is 0.95 million parsecs. Combining this number with their average radial velocity (from the redshifts) 600 km/sec, he obtained 625 km sec^{-1}/Mpc for the ratio between velocity and distance. The corresponding Hubble time for Lemaître's universe is about 2 billion years.

When Hubble wrote his landmark paper on the redshift law in 1929, he was apparently unaware of Lemaître's work and the predicted linear redshift–distance law. Written in French, it had been published in a not so widely read journal. But it is also important to note that Howard Robertson had in 1928 published a remarkable article in which he also predicted the linear law and inferred a value for the "Hubble constant". In a letter to us Allan Sandage recollects: "Robertson was my professor of mathematical physics at Caltech in 1951 and I got to know him moderately well. He told me that he had discussed this paper with Hubble in 1927 or 1928, including his discussion of an expansion with a rate $\approx$ 530 km/s/Mpc."

16.3 The Hubble constant: 100 or 72 or 50?

Measuring the redshift is not difficult, though this requires a large telescope equipped with a modern spectrograph. However, the measured redshift should be "cleaned" from its non-cosmological parts, due to the motion of our Milky Way and the wandering of the measured galaxy

The roamings of individual galaxies and large scale galaxy streams are difficult to determine and always cause uncertainty in measuring the Hubble constant. However, if one can reach very distant galaxies, then such "peculiar velocities" (typically 100 km/sec) become petty in comparison with the cosmological part of the redshift. Unfortunately, this advantage in redshift tends to be more than compensated by an increasing error in the distance itself.

Indeed, it is the measurement of distance that has caused the famous uncertainty and debate on the value of the Hubble constant. Everybody agrees now that Lemaître's 625 km $\sec^{-1}$/Mpc was much too large (as well as Hubble's own "inverse value" 526 km $\sec^{-1}$/Mpc), and that the overestimate was essentially caused by underestimated distances. Still, it is not yet quite clear what the correct value of the Hubble constant is.

In the 1970's and 80's, the debate on the extragalactic distances developed into a struggle between two sides, represented by the heavy-weights in the realm of galaxies, Allan Sandage and Gérard de Vaucouleurs. In round numbers, they derived that the Hubble constant is 50 and 100, respectively (we often drop the cumbersome unit). To-day, Sandage – and his longtime European associate Gustav Tammann – still remains faithful for his 50 to 60, while now the "other side", a younger generation of astronomers, has been favoring values around 70 to 80. But the trend is towards the lower value and an older universe, as first inferred years ago by the last of the great lone observers, Allan Sandage.

16.4 Distances to galaxies – a mission impossible?

Distances of nearby stars may be measured from their parallax angles, which reflect the Earth's yearly motion around the Sun. But these are such tiny swings in the sky that for distant stars and for other galaxies one has to invent other techniques. The chain of such methods, starting from the Sun, tying together nearby and distant stars in our Milky Way, jumping to

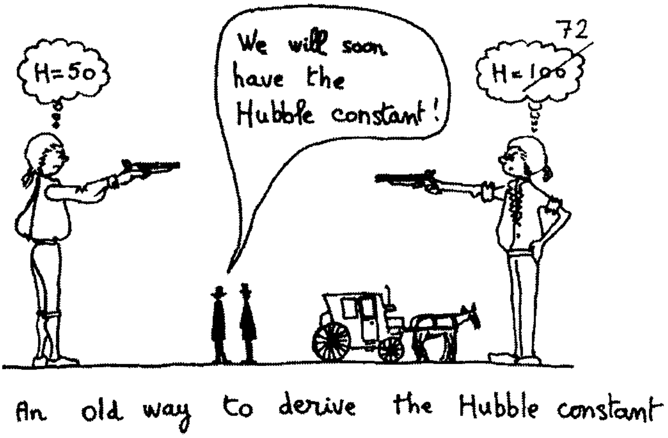


Fig. 16.2 In 1983 Georges Paturel stands with his pencil in the middle of the duel on the value of the Hubble constant. By 2001 the higher value has come down to 72.

nearest galaxies, and stretching out to more and more remote objects, this chain has the picturesque name: the *cosmic distance ladder.*

This ladder has not been drawn beforehand in the cabinet of some clever astronomer – it is the result of endless years of hard work to take advantage of what Nature happens to offer measurers of the universe. Really, it is one thing to capture the faint image of a distant galaxy on a photograph, and quite another thing, almost impossible, to measure its distance. The differences in the derived values of the Hubble constant, sometimes incredibly large, reveal how difficult the art of distance measurement is.

The distances of remote galaxies are mostly estimated from apparent brightness (flux of the received light) or angular size. Suppose that we transfer by some magic the Andromeda galaxy to a distance ten times its present distance. Clearly, after this shift it will be seen fainter and smaller in the sky (but no longer by naked eye!). Its new brightness (flux) is only 1/100 of the old Andromeda galaxy, because the flux of light drops as distance squared in transparent space.

If all galaxies were identical twins to Andromeda, one could measure their distances simply by comparing their appearance with the image of Andromeda. The natural unit of distance would be the distance of Andromeda, likely called one "andrometer" by the inhabitants of such a monotonous galaxy universe.

A world of identical galaxies is too good to be true. Their true luminosities vary greatly from one galaxy to another. Using the Andromeda galaxy as a meter stick for all other galaxies would often result in incredibly large errors in the derived distances. Just look at the photograph of Andromeda together with its companion M32, the roundish blob just to the right of Andromeda's center in Fig.3.1. (M32 was discovered by Guillaume Le Gentil in 1749.) Both these galaxies are at about the same distance. On the basis of the much smaller size of the dwarf elliptical galaxy M32, an astronomer believing in the world of twin andromedas would put it some 20 times farther than Andromeda!

Guillaume
Le Gentil
1725-1792

If the galaxies had labels with big letters saying "my luminosity is 3 Andromeda galaxies" or "I contain 100 billion suns", one could measure distances as easily as in the world of andromedas. Of course, such helpful labels are just dreams, and the astronomer has to do it the hard way.

The distance ladder has, after a large effort, allowed astronomers to measure with some accuracy distances to galaxies that are beyond the Andromeda galaxy though closer than, say, one hundred "andrometers". Briefly, it is the problem of identifying galaxies which have almost similar luminosities. For instance, among spiral galaxies those with more strongly developed spiral arms are also more massive and luminous. This morphological method, invented by Sidney van den Bergh, might be compared with inspection of passport photos in order to decide whether the owners are small or big guys. The currently much used Tully-Fisher rotational classification (named after Brent Tully and Richard Fisher) is based on the fact that more luminous spiral galaxies spin faster than the less luminous.

Assume now that the astronomer has some such means of recognizing galaxies which have similar true luminosities. He can then calculate relative distances for such galaxies. But he usually wants to know more, he wants the distance in parsecs. For this is needed the distance of at least one galaxy in parsecs. If the astronomer is lucky, there may be one galaxy, the closest one, for which he can make such a distance measurement. This galaxy is called the *calibrator*. For example, in the fancy world of Andromeda twins, Andromeda would be the calibrator, and in fact, its distance is known in parsecs from measurements of its Cepheid-stars (Ch.8):

distance to Andromeda = 670 000 parsecs

So in the world of andromedas all the distances easily measured in "andrometers" could be as well expressed in parsecs or kilometers. In our real

Fig. 16.3 Geoffrey Burbidge and Allan Sandage (right) at the end of the 1950's. Burbidge is one of the founders of the theory which describes how chemical elements are created inside the stars. For years he has pointed out observations which do not seem to fit the standard views in cosmology. Sandage is the father of modern practical cosmology, who is known worldwide for his extensive and careful work on the cosmic distance ladder and the Hubble constant. They are also known as editors of the prestigious Annual Review of Astronomy and Astrophysics. (Olin Eggen Photo Archives, NOAO)

world, Georges Paturel has devised a special distance indicator, the method of *sosie* galaxies where distant "Doppelgängers" of closeby calibrators are intentionally searched for among the rich variety of galaxy forms. *

In practice, it is difficult to collect galaxies into classes of similar luminosity. As a result, some inferred distances are too large, while others are too small, and the astronomer is happy to achieve a modest accuracy of 25 percent in his measurement of the distance to one galaxy. But more dangerous than such random "plus" and "minus" errors are the systematic errors, which always produce distance estimates which are too short, due to the so-called Malmquist bias.

---

*From *Larousse Classique*: Sosie – Personne avant une ressemblance parfaite avec une autre. Or "Sosie – A person with a perfect resemblance to another."

16.5 The notorious Malmquist bias

Observational data, our invaluable and subtle link to the world large and small, have also their Scyllas and Charybdises, ever ready to lure an explorer astray.

Malmquist bias is a splendid representative of the selection effects which constantly hide the truth about the universe and hamper the work of astronomer. One may even say – with slight exaggeration – that cosmology is just a study of selection effects! In a similar vein, Sir Arthur Eddington once compared cosmologists to detectives figuring out what took place at the scene of mischief. No genuine eye witnesses will ever tell what really happened in the depths of space and time. Indeed, what Sherlock the astronomer has to work with is little more than traces on a photographic plate, made by photons that have arrived from a known direction. From these meager clues he must construct the universe, its components and the physical phenomena that occur within it.

The weak light from remote celestial bodies has been attenuated by the distance itself and also by intervening cosmic dust. From very large distances – from the galaxies seen in the Deep Fields of the Hubble Space Telescope – the number of arriving photons is one billionth of the photons from the faintest stars visible by naked eye. A candle on the Moon would be brighter. Thus the astronomer can actually only see the intrinsically very luminous objects when trying to see what happens very far away.

Gunnar Malmquist 1893-1982

Gunnar Malmquist, was a pupil of Carl Charlier in Lund and later Professor of astronomy at Uppsala. He attempted to determine the average luminosity *in space* of stars belonging to a particular spectral class. This is not the same as the average luminosity of similar stars *on the sky.* At large distances, from where only the most luminous stars are visible, the less luminous stars will be missed. Hence the average luminosity of stars on the sky is larger than the true average luminosity in space. In 1920 Malmquist derived a famous formula which gives this difference in luminosities. †

Malmquist's theory may be applied to galaxies as well. When the astronomer measures distances to galaxies, he can select his galaxies only from the sky, and not from space, because his ability to travel between the

---

† The average absolute magnitudes "in space" and "on the sky" differ by an amount $1.382 \times \sigma^2$, where σ is the dispersion of the absolute magnitude distribution of the stellar class, assumed to be normal (Gaussian). This formula also presupposes that the stars are uniformly scattered in space. For a fractal distribution it becomes $0.461\mathcal{D}\sigma^2$.

galaxies is badly limited. Also, he cannot see very faint galaxies. Thus the standard candles which the astronomer can reach at large distances are bigger and more luminous, on the average, than the nearby calibrator galaxies. This mismatch is the Malmquist bias.

Malmquist bias tends to creep silently into any astronomical data. If the astronomer forgets this distortion, he will derive too small distances to remote galaxies and he will determine a too high value of the Hubble constant and, consequently, a too young age for the expanding universe. And such things have really happened in astronomy.

16.6 What, after all, is the value of the Hubble constant?

The issue of measuring the Hubble constant touches one of us (P.T.) who has attempted to come to grips with the mischievous Malmquist bias. The riddles of the cosmic distance ladder led him to collaborate with research teams in the observatories of Paris and Lyon. These French groups, grown around Lucette Bottinelli, Lucienne Gouguenheim, and Georges Paturel (and from the younger generation Gilles Theureau), have collected a uniquely large galaxy sample called KLUN (Kinematics of the Local Universe). It consists of 6600 spiral galaxies for which distances are known from the Tully-Fisher method of rotating galaxies. The future big brother "KLUN+" is expected to contain 20 000 galaxies, all eager to be used for still deeper exploration of the galaxy universe.

To gather so many galaxies requires much work and dedication from the team, years of planning and making databases, slowly advancing observations with large radio telescopes, countless nights for measuring redshifts under both northern and southern skies. But then something unique is born, a piece of the Universe is on the tray before you.

One motive for this effort was the measurement of the Hubble constant in a special way that utilizes so-called normalized distances. This method was invented in order to overcome the Malmquist bias. The analysis of the KLUN data led to $H = 55$. With its error marginals of ± 5, this value is the same as that obtained by Allan Sandage, who has used other samples and techniques, but has always been aware of the dangerous bias.

A team of astronomers working with the Hubble Space Telescope published their measurement of H in 2001. It turned out to be 72 ± 5, again a reminder that the debate is not yet over and there is still uncertainty

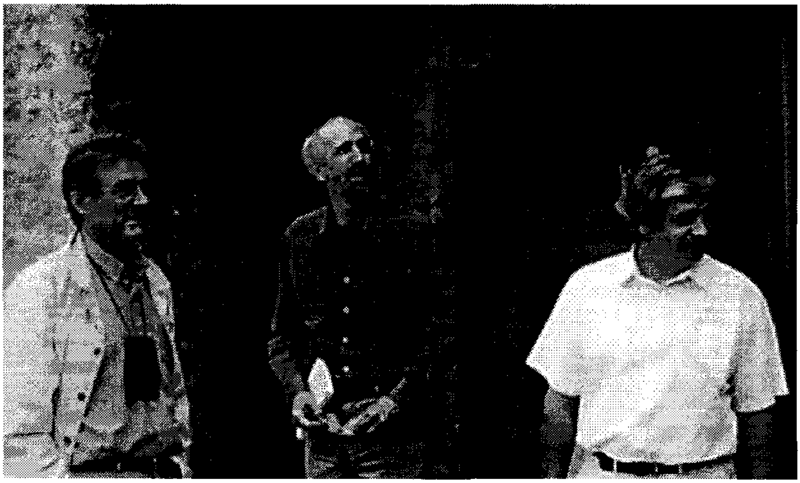


Fig. 16.4 Georges Paturel, Barry Madore, and Pekka Teerikorpi in Lyon in 1999, members of the jury for a doctoral dissertation on the value of the Hubble constant. Madore works in the Hubble Space Telescope Key Project to measure the Hubble constant.

in the cosmic distance ladder. Though it may well be that the high value again reflects another caprice of Dr. Malmquist, it is wise to allow for the possibility that the Hubble constant may lie anywhere between 50 and 70.

Recently a new kind of age determination for the universe was reported by Roger Cayrel and others. This international team have applied the method of radioactive dating for an old star in the Milky Way, detecting and measuring uranium outside of the Solar System for the first time. The very latest news from this project, which uses the powerful ESO Very Large Telescope (VLT), is that the star's age is 14 billion years. The universe must naturally be older still. If this result really holds, it casts a shadow of doubt on big bang models with $H = 72$, which are younger than the measured star. But a speeding-up $H = 60$ universe would offer enough life time, 16 billion years, for a Milky Way this old.

16.7 Galaxy clusters painted on the celestial sphere

For millenia the stars remained fixed on the celestial sphere, until their differing distances became both natural to think of and measurable. After the invention of the telescope, Man became aware of nebulae. Evidently the nebulae were at different distances, too, but one could only peer at those fuzzy blots with no means of measuring what the distances really were.

The ancients recognized various constellations, which are not real groupings of stars, but just interesting projections on the sky. But there are also genuine stellar swarms, such as the Hyades and the "Seven Sisters" or the Pleiades. Something similar took place to the astronomers when they mapped out the positions of nebulae on the sky. The comet hunter Charles Messier found a concentration of nebulae in the constellation Virgo. William Herschel found a collection of many hundreds of nebulae in Coma Berenices and several other groupings. Then his son John obtained hints for the existence of what is now called the Local Supercluster, the Virgo concentration "being regarded as the main body of this system" and the Milky Way "placed somewhat beyond the borders of its densest portion".

Photographic surveys of the sky from the early years of the 20th century revealed thousands of new nebulae (c.f. Fig.7.1), and such work continued with increasing fervor when the nebulae were revealed as galaxies. It became clear that a band of nebulae goes through Virgo and is perpendicular to the nebula poor Milky Way ribbon. Mainly through the efforts by de Vaucouleurs, the existence of the flattened Local Supercluster became established in the 1950's. There is indeed a structure formed on the sky by the galaxies something like the Milky Way seen on the starry sky!

The galaxies on the sky showed a clear tendency to cluster also outside the band of the Local Supercluster. These clusters became individuals in their own right and the subject of study. In 1927 Lundmark plotted 55 clusters on the celestial sphere in an article in which he also recognized that galaxies are often found in pairs, forming binary galaxies. The chart hinted that the clusters themselves may be clustered. Were there genuine clusters of clusters, i.e. superclusters? It took decades of ardent arguings before this issue was resolved.

16.8 The origin of the debate on superclusters

The first phase of the debate occurred in the 1930's. Hubble argued that clusters of galaxies were the largest units in the distribution of matter, whereas Harlow Shapley presented evidence on still larger structures. Shapley was the director of the Harvard College Observatory and a leading American astronomer who towards the end of the 1910's used globular star

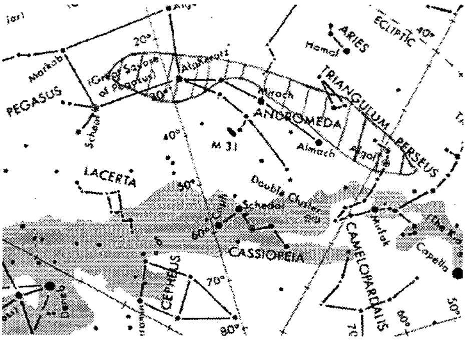


Fig. 16.5 Clyde Tombaugh discovered an elongated supercluster of 1800 galaxies stretching between the constellations of Perseus and Pisces in 1936. It is outlined on the upper part of this map. Note the grey band of the Milky Way, Cassiopeia within it (and inside the Local Spiral Arm), and our neighboring galaxy Andromeda (M31).

Harlow Shapley 1885-1972

clusters (and Cepheid stars) to chart the Milky Way, finding its center to be far away in the direction of the constellation Sagittarius. This momentous discovery moved us away from a central position yet again – according to the model popular before Shapley's work (the so-called Kapteyn Universe), we were thought to be rather close to the Milky Way's center.

Clyde Tombaugh 1906-1997

Shapley initiated wide photographic surveys of galaxies. The Shapley-Ames catalogue of 1249 bright galaxies from the year 1932 formed the basis for de Vaucouleurs's Reference Catalogue in the 1960's. Inspecting the distribution of galaxy clusters, Shapley came to the conclusion that there are "metagalactic clouds" (today's superclusters), for example in the constellations of Coma, Centaurus and Hercules. The Centaurus cloud is nowadays called Shapley's supercluster. It is interesting to mention that Clyde Tombaugh, the discoverer of the planet Pluto, noted as a by-product of his extensive planet searches the Perseus–Pisces supercluster. He counted 1800 galaxies in this elongated cloud which is now a much studied agglomeration of clusters of galaxies at a distance of about 100 Mpc. Working at Lowell Observatory, Tombaugh had as boss Vesto Slipher, the discoverer of cosmological redshifts!

Contrary to Shapley, Edwin Hubble thought that "the tendency to cluster appears to operate on a limited scale. No organizations on a scale larger than the great clusters . . . are definitely known." At the time the number of clusters was still small, and Hubble's greater authority carried the day. Incidentally, in the 1940's and 50's Shapley was closely surveilled, like many of his compatriots, by the Federal Bureau of Investigation, in a futile attempt to prove a connection between him and communism (as reported by the January 2002 issue of *Astronomy*). He was described as being "a stubborn advocate of his own ideas" and having "an inherent dislike for authority" and willing "invariably do the opposite to what he is told or supposed to do". Perhaps the sincere informant regarded these traits as something negative in society, but in science it is sometimes good to go one's own way...

16.9 Abell's rich clusters of galaxies

George Abell 1927-1983

An enormous increase in the number of known galaxy swarms came with George Abell's catalogue of 2712 *rich clusters of galaxies*, published in 1958. A cluster is said to be rich, if it contains many bright galaxies. Such clusters are rare, but can be seen from large distances in space. By the way, our neighboring Virgo cluster, so big that it is slowing down our recession from it, is far from being among the richest in the world! Abell's collection was one outcome of the photographic survey of the entire northern sky, made by the large 48 inch Schmidt telescope at Palomar Observatory – an incredibly important observational programme which gave astronomers huge amounts of data about stars and galaxies. The nine hundred 60 × 60 cm copies of the Palomar Sky Atlas photographs were a basic tool of observational astronomy at observatories all around the world for decades.

Now the question was: Do Abell's rich clusters form superclusters? Abell himself considered that his clusters were themselves clustered. With the large number of clusters at hand, it was relatively easy to answer "yes".

Jerzey Neyman 1894-1983

Elizabeth Scott 1917-1988

An interesting study during this "2-D period", in 1967, was made by Tao Kiang from the Dunsink Observatory in Ireland. He applied a statistical model to the celestial distribution of Abell's clusters. In this model, which had been introduced by the statisticians Jerzey Neyman and Elizabeth Scott in the 1950's, objects occur only in clusters which are randomly scattered throughout space. If we mean by "objects" the cluster of galaxies, is their observed distribution on the sky something like predicted by the statistical model?

In order to study this question Tao Kiang used a computer to generate artificial distributions of objects projected on the sky, adopting the space distribution specified by the Neyman–Scott model. Comparison with the true surface distribution allowed him to conclude that Abell's clusters do themselves cluster, but not in the simple hierarchic manner of galaxies forming roughly identical clusters and these forming identical superclusters. Observations could not be reconciled with superclusters of one fixed size.

Kiang concluded that clustering of galaxies occurs on all scales – there are no clear-cut hierarchic levels. His hypothesis of *indefinite clustering of galaxies* reinstated as it were the galaxy as the fundamental building brick of the universe. He was compelled to visualize the arrangement of galaxies so that the various clusters interpenetrate each other, for were it otherwise "the mean density must decrease as the volume in which the mean is taken, is increased". Thus, he came close to the very modern view of fractal clustering, but preferred to be guided by Einstein's cosmological principle, keeping the average cosmic density the same everywhere.

16.10 Looking through the dusty window

However, not all astronomers were satisfied with the apparent clumping of clusters on the sky. Some warned about the influence of cosmic dust. In fact, already in the 1940's Victor Ambartsumian made fundamental theoretical studies on how the cloudy distribution of interstellar dust would influence the apparent distribution of galaxies. And when Neyman and Scott introduced their statistical clustering model, in 1952, they were careful to point out two alternative explanations of the observed clustering: 1) Not only the apparent but also the actual spatial distribution of galaxies is clustered, or 2) the clustering is only apparent, caused by the effects of extinction of light by dust clouds. And they added that undoubtly both factors play a role. But which factor dominates in superclustering? This question was important in the next phase of the supercluster debate.

Of course, the Zone of Avoidance tells us that in the plane of the Milky Way there is a lot of dust. What has been less clear is how much dust obscures the view when we look perpendicular to the plane, towards the Galactic Poles. In the past it was often thought that through the polar caps our view into space is practically unhindered by dust, making those regions very valuable for extragalactic astronomy. However, gradually various observations have drawn a different picture: also the polar caps are strewn with a little dust, although its relatively small amount makes the dirty window look almost perfectly clean. Nevertheless, there is enough

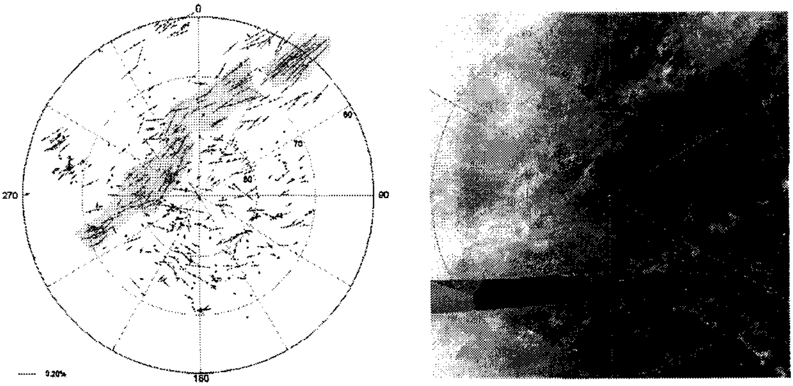


Fig. 16.6 The Northern Galactic Pole (in the constellation Coma Berenices) is a rather dusty window into space, as shown by these two maps of the polar cap (Galactic latitude $> 60°$). The right panel displays infrared radiation emitted by cool dust, as measured by the IRAS satellite (the black narrow region is an artifact, due to a lack of data). There are long filaments and a complex fractal distribution of the dust. The left map shows polarization directions of the light from distant stars, as measured by A. Berdyugin and P. Teerikorpi at the Nordic Optical Telescope (La Palma). Elongated dust grains, oriented by the interstellar magnetic field, polarize stellar light. The shaded region is an elongated dust cloud, first found by T. Markkanen in the 1970's.

dust (and gas) to influence extragalactic observations.

For example, building the cosmic distance ladder requires that we have accurate information on the amount of light extinction by dust (dust makes galaxies look fainter, fooling the astronomer into a too long distance estimate). And different amounts of dust in opposite directions on the sky may give rise to erroneous "streaming motions" of galaxies. The dust also emits faint thermal radiation, and may also leave its mark on the cosmic background radiation. As the astronomer cannot brush away the dust, he has to learn to live with it.

Inspecting the lumpy distribution of nebulae in Fig.7.1, Charlier allowed for the possibility that it could have been deformed by "dark matter in space". Similarly, in the 1960's and 70's a few astronomers tried to understand the clumping of galaxy clusters, not as real superclusters in space, but as an artifact caused by the nonuniform light extinction in the Milky Way or even intergalactic space. Such a possibility makes sense, because in dusty directions the fainter, more numerous galaxies in a cluster will be dimmed out of sight and hence the cluster itself may go unnoticed by the astronomer. Clusters are best seen when they lie behind almost transparent regions with little dust.

Erik Holmberg 1908-2000

Erik Holmberg, an eminent Swedish investigator of galaxies and one of the pupils of Knut Lundmark, was quite worried about the role dust may play in molding the apparent pattern of galaxy clusters on the sky. In his article of 1974 "Distribution of clusters of galaxies as related to galactic absorption" he presented evidence that the regions of the sky which contain many clusters are more transparent (contain less dust) than the sky in general. He concluded that "the random distribution of the galaxy clusters has thus been proved in an indirect way". Indeed, at the time the issue of superclusters could not yet be approached in a direct way, by working in three-dimensional space.

It is interesting to note that Fritz Zwicky, who found dark matter in galaxy clusters and whose extensive galaxy cluster catalog was utilized by Holmberg, was himself of the opinion that superclustering is due to *intergalactic* extinction. At one end of the spectrum of opinions was Boris Fesenko from Pskov who radically held that even galaxy clusters themselves are transparent holes in the dusty Milky Way.

Now that we know superclusters really do exist, the worries about fluctuating extinction may sound obsolete, or even just historical curiosities. In the specific issue of superclusters this explanation, although reasonable and very worth probing, was destined to fail. But there is no reason to ignore the tiny grains of dust altogether. Even though the spatial distribution of galaxy clusters is far from random, Holmberg's result – more clusters in transparent regions – remains valid, and we should not forget that our lines of sight reach the galaxies and quasars through the patchy, dusty window.

16.11 3-D astronomy from the vertex of a space cone

Astronomy students learn an old and ever important discipline, spherical astronomy. In practice we are still, so long after Ptolemy, stuck as it were at the center of a huge sphere onto which stars and galaxies are projected. Only if we know their distances, can we shift them into their true positions around us. Then if we look at a limited region on the sky, we get a cone, from the vertex of which we view its contents. This would be a three-dimensional map of a part of the universe.

But even if he knew all the distances in a space cone the astronomer

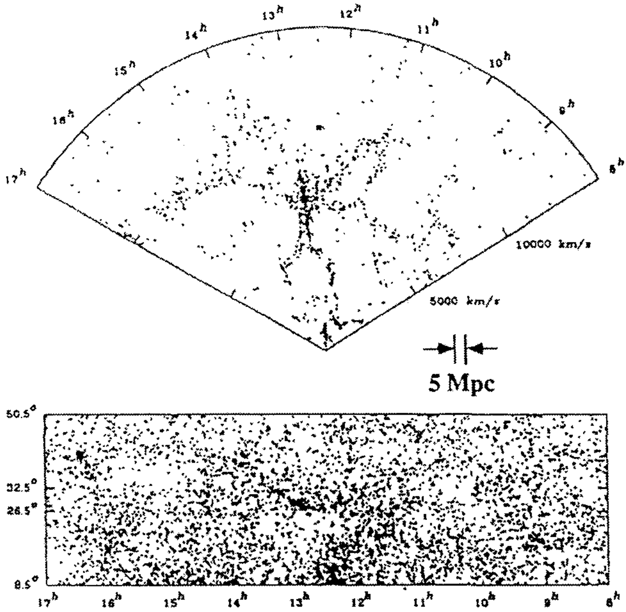


Fig. 16.7 The upper figure is the wedge diagram of the first CfA redshift survey. This 3-D map revealed large non-uniform structures and voids which are not at all seen if one just looks at the sky, without information on the redshifts (at the bottom). Some rows of galaxies pointing at the observer (at the vertex of the cone) are "Fingers of God" caused by the rapid movements of galaxies inside clusters.

would still feel our inconvenient position on the "bottom of the well". We encountered one of these problems in connection with the Malmquist bias. The cone will be filled with galaxies in a very distorted way. Intrinsically luminous galaxies are found throughout the cone, whereas dwarf galaxies concentrate into a small volume close to the vertex, because we simply cannot see those at large distance. [‡]

Sometimes our vertex position leads to especially bizarre problems. For instance, there is a popular theory (called "unification") according to which the different kinds of active galaxies are actually one and the same thing.

[‡] Note that the volume of a spherical shell with thickness Δd and radius r is proportional to $\Delta d r^2$, the volume increasing as the *square* of the radius (distance). In fact, it is this volume effect, together with the inverse square law for the flux of light ($f \propto r^{-2}$), which leads to the formula for the Malmquist bias (footnote above), and may also be used to demonstrate the blazing sky paradox in infinite classical space (Ch.3).

An active galaxy is seen as a radio galaxy, or as a quasar, or as a so-called blazar, depending on the direction from which the galaxy is viewed. If one has a "face-on" view of the accretion disc around the supermassive object, one sees the active nucleus in all its luster – this is a quasar or, in the extreme case, a blazar. If one looks at the same galaxy more from the side, the active nucleus is shadowed behind thick dust and – hey presto! – there is a radio galaxy before one's eyes. It would be easy to confirm this theory: just go and look at a quasar from different directions... Such an enterprise being out of question, the astronomer has to device a plethora of indirect tests of the unification theory.

Nevertheless, in spite of the fact that 3-D maps do not permit the astronomer to see everything he would like to, it was realized that they would be a great thing to posses. Just think about the above debate about superclusters of galaxies, which relied on 2-D projections. But in order to make such maps for the galaxy universe, you need distances. The distances of galaxies simply inferred from their apparent sizes or brightnesses are generally so inaccurate as to shuffle completely any attempted 3-D map. And only recently has the number of galaxies with more accurate modern photometric distances started to increase. But redshift is still by far the best indicator of distance for galaxies.

Gérard de Vaucouleurs's Reference Catalogue of Bright Galaxies was published in the 1960's. It contained redshift measurements for about 1500 galaxies. This was enough for the first attempts to make 3-dimensional maps, or at least maps of slices across space. For example, one may restrict the redshift between two values and plot all such galaxies on the map. They should then lie within a spherical shell around us and the structures would not be smoothed away because of different distances.

The astronomers became aware of these new developments in September 1977, in the first international Symposium devoted to the large scale structure of the universe. It was held in Tallinn, Estonia. Many of the persons appearing in our book were among the participants (De Vaucouleurs, Peebles, Davis, Tifft, Burbidge, Chernin, Parijskij, Zeldovich, Rudnicki, Abell, Holmberg, Jaakkola, Fesenko, Karachentsev, Kiang, Huchra...). On the third day of the meeting Mihkel Joeveer and Jaan Einasto, astronomers from Tartu Observatory, made a startling announcement about the space distribution of galaxies. Their paper was entitled "Has the Universe a Cell Structure?" They had used redshifts to build 3-D maps. Far from a smooth distribution of galaxies, the maps revealed astonishing structures around us in space, in the form of long filaments and giant walls marking out a kind of honeycomb. There were huge voids which contained no galaxies.

At the time, the available observations suggested that the "cells" are about 30 Mpc wide, something like the size of the Local Supercluster. More recently, Einasto and his team have presented evidence for a roughly regular network of clusters, where the size between the walls is about 120 Mpc. The current (wide-angle) galaxy maps do not allow one to see how the clusters are distributed at distances larger than a few "cell-sizes".

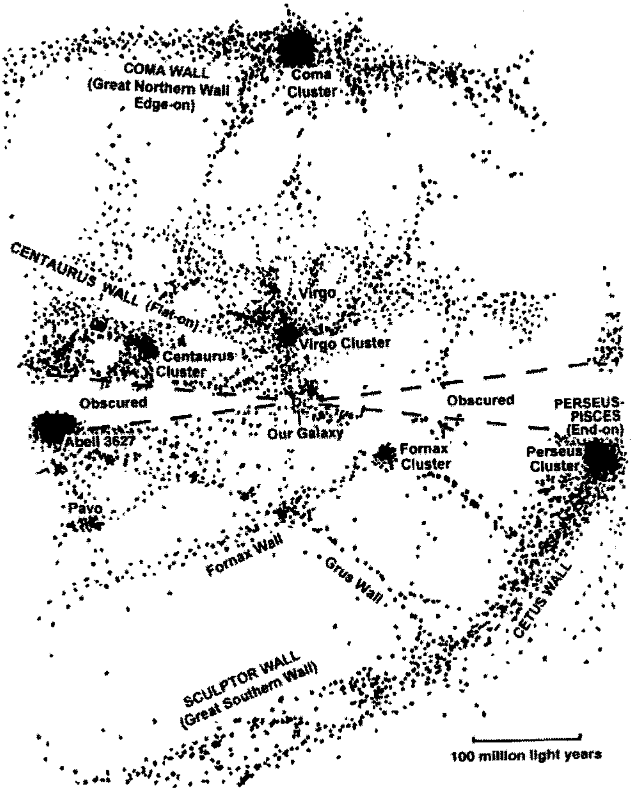
16.12 Excursions into the local galaxy universe and beyond

The cellular structure revealed by Jaan Einasto's team using such sparse redshift catalogues motivated the speeding up of efforts to measure many new redshifts all over the sky. Many astronomers were worried about the heterogeneous, incomplete data then available. The discovered architecture of the galaxy universe was quite unexpected and it was not readily accepted.

The first survey, accurately planned to produce a complete sample of redshifts on a large region of the sky, was carried out by astronomers at the Harvard-Smithsonian Center for Astrophysics in the USA. It is good to mention some names behind this pioneering effort: Marc Davis, John Huchra, Margaret Geller, Valérie de Lapparent... They measured redshifts for all galaxies in Zwicky's catalogue which were brighter than $m = 14.5$ mag. This survey came to be known as the CfA. Later John Huchra recalled his reaction when he saw the first 3-D plots of the survey: "...my first, very conservative reaction was whoops! what did I do wrong?" He had not expected anything extraordinary in the distribution of galaxies. By the way, it was Huchra who in 1984 found the famous Einstein cross, Fig.10.2.

The new map, published in 1986 in a paper titled "A Slice of the Universe", confirmed the Estonians' discovery of shell-like galaxy clustering and found still more diversity in the galaxy distribution. To-day one speaks fluently of The Great Wall across Coma Berenice Hair, the Cetus Wall across the Perseus-Pisces supercluster, the Sculptor Void between the Fornax and Great Southern Walls.

We recall how the dream of Lundmark appeared to come true, when in the 1950's de Vaucouleurs established the existence of a Milky Way of galaxies, the Local Supercluster. But perhaps that really happened only when in the 1980's Georges Paturel and his collaborators Lucette Bottinelli,



A. P. Fairall - University of Cape Town

Fig. 16.8 Clusters, superclusters, filaments, and voids in the local galaxy universe within about 100 Mpc. The Milky Way is in the middle. (Drawing by A. Fairall).

Lucienne Gougenheim, and Pascal Fouqué pointed out that a very large structure is visible in the extragalactic sky, a kind of Milky Way of clusters of galaxies. Its "hypergalactic" plane almost coincides with the plane of the Local Supercluster and contains several other neighboring superclusters. Similar observations were reported by Brent Tully.

Other astronomers attempted to made deep stings into space. In this way an intriguing observation was made by Richards Ellis of the University of Durham and his colleagues in 1990. They performed "pencil beam" redshift surveys, concentrating their attention to a very small patch of the sky where it was possible to measure all redshifts of galaxies with a tolerable effort and time. Inspection of the redshifts (i.e. distances) revealed apparent periodicity in the space density of galaxies. Densely populated narrow walls of galaxies were separated by almost empty voids. The intervals between peaks had lengths of about 120 Mpc. They were seen in two opposite directions in the sky. Could this phenomenon be related to Einasto's more local network? This is not known, but it shows again the strong nonuniformity of the galaxy distribution.

Here we should mention one more problem with 3-D astronomy as it is practiced at the vertex of the cone. Observed redshifts of galaxies always include a part due to real motions of galaxies caused by gravitational effects of nearby masses. Hence the redshift has always a small non-cosmological part and the distances derived with the Hubble law always contain some error. The maps based on redshifts are not quite faithful copies of the true three-dimensional galaxy universe.

One type of error literally strikes the eye. Dense clusters of galaxies appear strongly elongated on the maps, as if they were "pointing" at the observer. These "Fingers of God" are no true wall-like structures in space. They are caused by the high velocities of the member galaxies inside massive galaxy clusters. Those galaxies which happen to be moving almost towards us or away from us are also shifted in the redshift map in the same sense, making the cluster elongated. Recall from Chapter 10 how Zwicky found that galaxies in clusters move quickly, most likely because of the gravity of dark matter. The Fingers of God allow you to "see" the invisible masses.

16.13 The mysterious quietness of the Hubble flow

But how accurate is the Hubble law, i.e. redshift, as a distance indicator? The redshift itself may be measured very accurately, so the error in the inferred distance depends essentially on how tight a dependence there is between the redshift and distance. The expansion component of the redshift would be directly proportional to the distance, if the galaxies, others as well as our Milky Way, were at rest relative to expanding space. But galaxies can not be considered to be nailed to space, not even in the space of the expanding big bang.

As we mentioned above, galaxies have their own motions, apparently induced by the gravitational attraction of their nearby neighbors and more distant massive galaxy concentrations, such as Abell's rich clusters. But

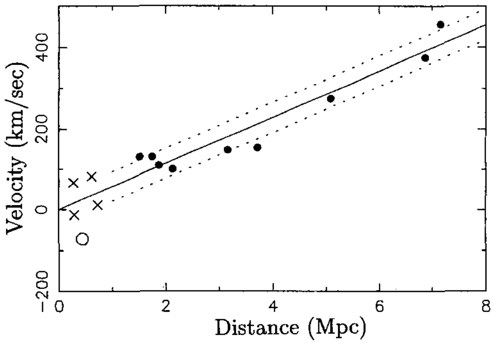


Fig. 16.9 The Hubble law in the environment of the Local Group as revealed by modern cepheid-based distances. The law starts immediately at the border of the Local Group, at a distance of 1.5 Mpc. Its scatter is remarkably small, less than 40 km/sec. (From a study by T. Ekholm, Yu. Baryshev, P. Teerikorpi, M. Hanski, and G. Paturel.)

the rich are rare in the kingdom of galaxies and between the clusters the Hubble law is amazingly accurate. Astronomers say that the Hubble flow is very "quiet", with little scatter due to the galaxies' own motions. The remarkable linearity and smoothness of the distance–redshift relation allows the astronomer to make an even better map for the extragalactic space than for the stellar realm of our own Milky Way!

If we could measure perfectly accurate distances for a sample of galaxies, their Hubble diagram would directly tell how precisely redshifts indicate distance. Any scatter away from the thin line representing the Hubble law is then caused by the galaxies' own motions. [§] In real life one cannot easily separate unknown deviations from the expansion redshift and distance errors. The goodness of the Hubble law is best studied close to the Local Group, where one may use the accurate Cepheid distance indicator.

Already in 1957 de Vaucouleurs concluded from the meager galaxy data then available that in the local galaxy universe the deviations from the regular Hubble flow (apart from the streaming motion towards the Virgo cluster) are small, with a scatter of less than 100 km/sec. In their famous

---

[§] The galaxy streams add genuine Doppler shifts to the expansion redshift. Fortunately, at distances beyond the Local Supercluster, the resulting errors in the redshift distance are much smaller than the size of the large structures being studied. The observed redshift contains the cosmological and Doppler parts: $1+Z_{obs} = (1+Z_{exp})(1+Z_{Dop})$. Thus a motion of 300 km/sec gives an error of 5 Mpc in the distance determination.

series of "The steps towards the Hubble constant", in the 1970's, Allan Sandage and Gustav Tammann pointed out that the scatter is even smaller, less than 60 km/sec. In another connection, Sandage predicted that the better one can measure the distances, the smaller the scatter becomes. This profecy was tested using galaxies with cepheid-based distances (see Fig.16.9). And indeed, now the scatter in the local Hubble flow came down to the record value < 40 km/sec! Recently Igor Karachentsev and Dmitrij Makarov have presented evidence for an even smoother flow.

Recall that the regular Hubble expansion is a consequence of uniformity. This argument has created a mystery as Allan Sandage rightly calls it: as the local space of galaxies is very non-uniform, how then can the local Hubble law be so smooth? Why is our closest environment around the Local Group expanding at a similar rate to much larger volumes? ¶

The problem of "too good" a Hubble law has its history. In 1972 Sandage, Tammann, and Eduardo Hardy were very puzzled by the linearity of the Hubble law amongst the very non-uniform galaxy distribution. They suggested that either the mean density of matter is very low or there is a dominant smoothly distributed dark substratum. An early expression of this paradox is also hidden in the natural reaction by Steven Weinberg in *The First Three Minutes* when he described how Hubble discovered his law in the local, lumpy universe: "In fact, we would not *expect* any neat relation of proportionality between velocity and distance for these 18 galaxies – they are all much too close, none being farther than the Virgo cluster".

In the tenth Petrie Lecture, delivered at Montreal in 1989, James Peebles considered the cosmological significance of the nearby space. Inspecting the Hubble diagram for the galaxies within 10 Mpc, he wondered why they all are drifting away from us at the same rate as the pure Hubble flow for distant galaxies. He was excited that "this nearly homogeneous expansion is in striking contrast to the extremely clumpy space distribution".

¶Fabio Governato and others have shown using N-body simulations within CDM models that the velocity dispersion in the vicinity of the Local Group is expected to be in the range 150 to 700 km/sec. Another problematic issue is our velocity of 400 km/sec relative to the cosmic radiation. When we subtract from it the rotation of the Sun around the Milky Way's center, it is found that the Local Group has a speed of 600 km/sec within the expanding universe and it is strange to realize that its nearby environment expands uniformly, too. Johann Lambert already pondered, what our velocity relative to (Newtonian) absolute space is, when we participate in different revolutions around the distant centers in his hierarchic world. Now we do know our velocity, but we do not as yet know all those systems whose gravitating masses have given us that speed.

It is odd that the Hubble law, so universal and familiar in the light of our cosmology, should offer such a problem – one would rather expect that its absence somewhere could be a source of worry, and not its excellent validity. But the phenomena of even the local universe are far from fully understood: just recall dark matter and dark energy! We return to the Hubble law in Chapter 18.

16.14 The redshift of quasars as a distance indicator

A promising way to probe the structures on the very largest scales is to employ redshift surveys of quasars, which can be seen from very far away. Of course, the maps made using quasars tell us about the large scale universe only if the redshifts of quasars are good distance indicators. Are they?

Astronomers have always been struck by the large scatter in the Hubble diagram of quasars. This has sometimes been viewed as indicating that the redshifts of quasars include large non-cosmological components (anomalous redshifts, which we discussed in Chapter 7). But more usually it has been explained as reflecting a great variety of luminosities: there are very bright quasars and quite faint ones, and everything in between. So quasars as a whole are far from being a "standard candle". This is problematic, because for some classes of galaxies, the appearance of a tight relation in the Hubble diagram was seen as evidence for both the Hubble law and the standard candle. And for quasars there are no "Cepheid-stars" or other methods which could be used to measure their distances independent of the redshift.

Ajit Kembhavi and Jayant Narlikar, in 1999, wrote in their book on quasars: "Of course, it may be argued that the scatter arises because of the variation in luminosity from quasar to quasar. In that case, one should try to identify a 'standard candle' class of quasars...The identification of such a class for galaxies (viz. the brightest member of a cluster) led Allan Sandage and his coworkers to a tight Hubble relation...Such an exercise has not so far been succesful for quasars."

As it happens one of us (P.T.) has come across evidence for a separate class of the optically most luminous (radio loud) quasars in the Hubble diagram, and thus also evidence for the cosmological redshift. The roots of this investigation of very distant objects go, oddly perhaps, to the nearby dirty window of our Milky Way – it started from the question of how to use quasars as indicators of the presence of dust, and led to the correction

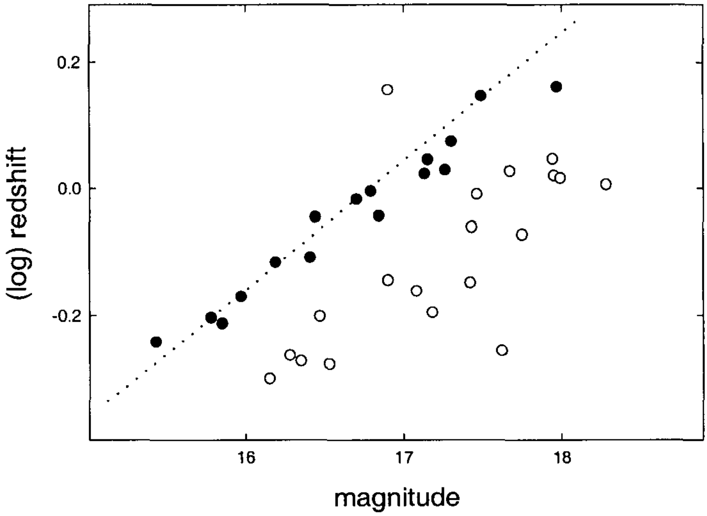


Fig. 16.10 One way to probe whether the redshift of quasars is a distance indicator, is to study their Hubble diagram. Here we show the magnitude–redshift plot for radio loud quasars, excluding those whose light is very variable. The magnitudes have been corrected for the extinction in the Milky Way. Note the strip of quasars, with the expected slope 0.2, enveloping the bright side of the quasar population. Across the gap some properties of quasars change. That one may see such a structure in the Hubble diagram suggests that redshift is a distance indicator for quasars. It is hoped that in the future they may be used to test the world model as is done using the supernovae.

of their magnitudes from the dimming caused by this same dust. A further motivation, after the separate class emerged from the extinction corrected data (for the first time in 1981), has been the idea that a luminosity class may reflect a specific host galaxy class. This has gained support from the recent Hubble Space Telescope discoveries that evidently all galaxies contain supermassive objects in their cores and the masses of these energy machines for quasars are proportional to the masses of the host galaxies. The proposed quasar class might be hosted by some subclass of the brightest cluster galaxies, known to have a very narrow luminosity (and mass) distribution. [‖] By the way, the reality of the separate class is also supported by the observation that its powerful member quasars stand out from

---

[‖] These host galaxies may represent a stage in Boris Komberg's evolutionary classification of quasars. Komberg, from the Astro Space Center in Moscow, envisions different host generations, arising from the mergers of galaxies containing supermassive objects.

the crowd: they are meeker (optically less variable), and as double radio sources they are the largest and most symmetric.

If this hypothesis is correct, it would have an immediate physical implication: for such a class to emerge from the observations, the redshift must be a distance indicator for quasars, too, at least for the radio loud ones. True, most quasars are radio quiet, with no similar evidence for subclasses. But at small redshifts there are cases, in which a quasar is found in a cluster of galaxies, both with the same redshift. In our opinion, it is natural to assume that quasars as a whole follow the Hubble law. Of course, should it happen that the space maps drawn by quasars show only very blurred or no structures, one may have to check if a non-cosmological part of redshift is the culprit – in a sense, the maps will also test the reality of anomalous redshifts!

* * *

In an anecdote a cosmologist is said to be like a man searching for a lost key in the patch of light under a street lamp. When asked why he was searching that particular place he replied that, of course, because here there is light... Yes, light is important, even with all those pesky selection effects restricting our view. A weak flow of photons, our only bridge to a distant galaxy, may carry dramatic information in a subtle form, as the cosmological redshift also testifies.

The enigmatic quietness of the Hubble law, its small scatter outside clusters of galaxies, is also a blessing for the astronomer who is attempting to draw a map of the galaxy universe. The amazing results from the first redshift surveys proved that it pays well to spend time and effort on gathering night after night faint light from the multitude of galaxies. When astronomers entered the 3-D epoch, it finally became possible to discover the true nature of the distribution of galaxies. But how far do the structures spread which we have found in our extragalactic neighborhood?

Chapter 17

Fractal structure of the galaxy universe

In June of 1987 an article on astronomy came out in the journal *Physica A* which usually covers topics in statistical physics. Its author was a professor from Rome University, Luciano Pietronero, a specialist in this field of physics. The title, however, did not sound like anything coming from an Earthly laboratory: 'The fractal structure of the Universe: correlations of galaxies and clusters and the average mass density.' For the cosmic community this was an unusual message about the break-down of the Cosmological Principle, arriving from a remote field of expertise. At that time the concept of the fractal, the creation of Benoit Mandelbrot in the 1960's, had hardly entered astronomy. Fractal geometry is a splendid example of an interdisciplinary approach for solving problems in different sciences. But still crossing the borders is not always welcomed.

It took a decade of deepening observations and ardent debates, before the January 1999 issue of *Nature* acknowledged that the universe is "essentially fractal" on scales below 100 Mpc. While views on galaxy fractals still diverge for larger scales, the observations go deeper, and with the change of the century the news came that galaxies, clusters of galaxies, and quasars reveal evidence for structures on scales up to 300, 600, 900 Mpc...

* * *

17.1 Einstein's Cosmological Principle

In 1917, when Europe was in the flames of war and revolution, the modern principle of cosmic harmony was born. On the 6th of February Einstein presented his epochal work on cosmology to the Prussian Academy, in Berlin at the legendary street Unter den Linden. Einstein applied his brand-new general relativity to the cosmological problem, i.e. constructing a model of the universe as a whole. Not knowing about galaxies, he imagined a world filled with stars and argued that the stars have a natural spatial distribution, which is uniform: matter concentrations around any preferred center should with time evaporate and disperse uniformly all over the universe. * Later, in discussions with Selety, Einstein rejected the hierarchical distribution of stars. One reason was its privileged center, on which point Selety and Einstein disagreed (see Ch.13).

Einstein accepted the principle of no center, postulated a uniform matter distribution, and put relativistic gravity into cosmology. The result was a world with uniform geometry. Even earlier, perhaps the closest step towards the uniformity of both the matter distribution and geometry of space had been made in 1900 by Karl Schwarzschild who assumed uniformly distributed stars in his discussion of stellar parallaxes and the curvature of space. He concluded that the radius of curvature of a spherical space (if ours is such) must be at least 160 million times the Earth-Sun distance. In this case the distribution of observed stellar parallaxes could be consistent with uniformly scattered stars.

Besides the absence of a center, another plus of uniformity was a simplification of Einstein's equations, which permitted him to derive the static spherical world model. He compared himself to a geodesist who describes the average form of the Earth by an ellipsoidal figure, though the details of the surface are complicated (now we know so rough that fractal mountains and coastlines may appear...). Finally, in 1922 Friedmann liberated the universe from this stiff state, allowing the uniformly distributed matter and space to expand.

The name Einstein's Cosmological Principle for the hypothesis of the uniformity of the universe was coined by Edward Milne who analyzed the foundations of cosmology in the 1930's. In these early years of modern cosmology there was no direct observational evidence for the uniformity of the universe and it was theoretical reasoning which guided the cosmologist.

---

*This view curiously reminds one of the atomist Epicurus who could not accept the Stoic world in which there was one big island of matter within an infinite empty void: "If space were infinite and the bodies were limited in number, these could not stay in some one place, but would be moved into infinity, they would be dispersed without any assistance or propulsion other than collisions."

17.2 Many faces of the Cosmological Principle

In the beginning of our story, we told of the time when the first ancient principles were stated by Greek philosophers, and envisioned to govern the whole of the universe. But we should also mention cosmological views in ancient India as explored by Konrad Rudnicki. Indian philosophy was immersed in poetry and there was no science in the modern sense. The oral teachings, coming down the epochs from several thousand years B.C. were intended to be experienced and not logically analyzed. Rudnicki discerned an underlying idea which he called the *ancient Indian* cosmological principle: *The universe is infinite in space and time and is infinitely heterogeneous.* Whatever the truth about the entire world is, cosmology has rather started from the simpler view of regularity and uniformity which is easier to imagine, has aesthetic value, and is mathematically tractable.

Plato's heavenly spheres, the prototype of regularity, can be expressed in modern terms as the *principle of local isotropy*, in which isotropy means the absence of a privileged direction on the sky. In fact, the sphericity of the Earth itself nicely followed this rule. Strictly speaking, Plato had to complement the spheres with the principle of circular motion, which did define a preferred direction – the axis of rotation. And of course his universe had a special place, occupied by the Earth.

The view of the absence of a privileged place also comes from Antiquity. The atomists had maintained as a guideline of cosmology that "it is not important in which region you stay". In 1440 Nicholas of Cusa explicitly expressed the principle that "the center is everywhere". In the following century Copernicus, changing the traditional mathematical world model, shifted the center of the universe from the Earth to the Sun. It did not take long before Bruno emphatically wrote that there is "no center in the universe". His doctrine also included the universality of Earthly laws.

While the principle of isotropy was relatively easy to accept under the revolving starry sky, it required a much longer time and dramatic events in science in order to adopt the vision of no center. In the hindsight we may see that the Cosmological Principle is a kind of unification of the old principles of no center and local isotropy. In Newtonian cosmology it gained the status of a mathematical principle in the form of a uniform stellar distribution.

A new face of the Cosmological Principle was unveiled by Hubble after the discovery of galaxies. In his Halley Lecture at Oxford in 1934 he suggested that "the Observable Region is a fair sample, and that the nature of the universe may be inferred from the observed characteristics of the sample." This Hubble Hypothesis of Fair Sample states that it is possible for an observer to find a local volume which is big enough to reveal global properties of the universe. It generalizes "no center" to all local properties, and includes the local isotropy as a particular feature.

In the 1930's Milne formulated the modern version of the Cosmological Principle as being that the *whole world-picture as seen by one observer (attached to a fundamental particle or galaxy) is similar to the world-picture seen by any other observer.* This includes the physical laws: they are the same everywhere and produce similar things. This assumption is at the heart of practical cosmology, too. For example, Lundmark defined a distance indicator as a group of cosmical objects having the same physical properties in different galaxies. The principle of uniformity of natural law is assumed to be valid when one jumps from one galaxy to another. One may now rephrase Rudnicki's formulation of the Copernican Principle (Ch.2): The universe as observed from any galaxy looks much the same.

But why do we think intuitively that the universe looks alike at all places? One reason comes from historical experience: one after another we have rejected proposed centers of the universe in the Earth, the Sun, the Milky Way. It is as if Nature is hinting that She has no need for "a place to stand on". Furthermore, the absence of absolute space, time and motion in modern physics also facilitates the acceptance of this idea.

More philosophically, and in the spirit of the anthropic principle, perhaps for the universe to be conceivable by the human being, it really had to be designed according to the master plan laid out by the Cosmological Principle. Einstein may have seen a glimpse of this when he said "The most incomprehensible thing about the universe is that it is comprehensible". A universe randomly drawn from nothingness would be an incomprehensible mess – and moreover devoid of Einsteins and the rest of us. Only to a big, orderly and symmetric world could the gift of life be given. But from our side, we view our inseparable companion, the beautiful and comprehensible universe, as an incredible gift which we do not seem to deserve...

17.3 The derivation of uniformity from local isotropy

There was a guess in the air that there must be a way to prove that from the isotropy observed at one point, together with the principle of no cen-

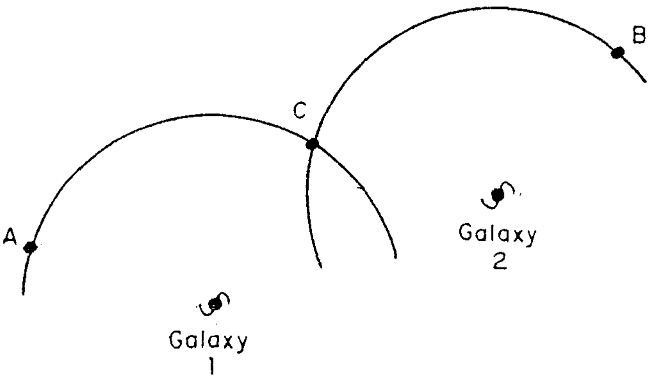


Fig. 17.1 How isotropy and homogeneity are connected. From isotropy around each point it follows that the density on the circle around galaxy 1 is the same on each point of the circle, and on the circle around galaxy 2 the density is also constant. Because the circles have a common point C, the density in fact is the same on both circles. By a similar construction of intersecting circles, one may show that at any point the density is the same, i.e. the matter is distributed uniformly. The assumption of a smooth density distribution around every point is essential.

ter, follows that there is uniformity everywhere. This seems intuitive: if every place is as good as any other, then every observer in the universe (not only us) sees isotropy, and this cannot happen if there are large scale inhomogeneities. Indeed, in 1944, Goeffrey Walker, the British mathematician whom we encountered in Chapter 8 and who worked closely with Milne, proved this conjecture starting from his Hypothesis of Local Spherical Symmetry which supposes that isotropy exists locally about each point.

A simple reasoning leading to homogeneity when there is isotropy around each point, may be found in Steven Weinberg's *First Three Minutes.* Fig.17.1 shows how one can go from any one point to another arbitrary place along circle arcs on which the density remains the same. Hence the density is the same on every point. However, strictly speaking this conclusion is based on a hidden mathematical assumption of regularity, i.e. the existence of a smooth density on each point, and only then from the left-hand side does follow the conclusion:

$$\textit{Isotropy} \ + \ \textit{No Center} \ + \ \textit{Regularity} \ \Rightarrow \ \textit{Uniformity}$$

Here "isotropy plus no center" means that all points are equivalent and around each point the density law does not depend on the direction (though it might depend on the distance from this point – the proof shows that it doesn't!). The "regular" matter distribution is described by continuous, smooth mathematical functions. In the chapter "Simplifying assumptions of cosmology" in his *Introduction to Cosmology* Jayant Narlikar explicitly introduces the assumption of the *smooth fluid approximation*, which essentially means going over from a discrete distribution of particles to a continuum density distribution. This is what we call "regularity". It means that one may use the concept of the mass density at each point of space, like in a fluid. In this case a distribution of points may be looked at as containing a smooth "signal" superposed on discrete "noise" (see the upper panel of Fig.14.6). It is thus the union of local isotropy, no center, and smoothness which gives homogeneity, the fundament of modern world models.

17.4 The galaxy universe may seem rather smooth...

In the early decades of modern cosmology there were three types of indirect reasonings which seemed to confirm the uniformity of the galaxy universe and gave motivation for the theorists to develop homogeneous world models.

In the 1930's Hubble counted galaxies from the photographs of several patches of the sky. For relatively bright galaxies (brighter than 17 mag) he found that the counts follow a "$0.6m$-law", but for still fainter galaxies the numbers increase slower, following a $0.5m$-law. A formula of statistical astronomy gives the expected number of galaxies which are brighter than a given magnitude m. If galaxies have a uniform distribution, then one expects the $0.6m$-law. †

Indeed, the observed $0.6m$-law was widely regarded as a proof of uniformity of the galaxy universe, also by Hubble himself, who concluded that the observable local universe is a fair sample of the universe as a whole. And Einstein was pleased that his cosmological assumption, which previously was a theoretical guess, thus had received a kind of observational confirmation, at least for brighter galaxies.

---

† The observer counts in the sky the numbers of objects brighter than apparent magnitude m. In the general case of a fractal distribution of light sources in Euclidean space the resulting counts are given by the relation $\log N(m) = 0.2\mathcal{D}m + const$, and for the homogeneous distribution ($\mathcal{D} = 3$) there is the famous Seeliger's result of stellar statistics: $\log N(m) = 0.6m + const$.

But why do the galaxy counts for fainter galaxies deviate from the $0.6m$-law? Hubble examined this question in his book *The Observational Approach to Cosmology.* Faintness means remoteness and he noted that the deviation could be due to the galaxy density thinning away from the Milky Way. But this would imply a curiously unique position for us, against the Copernican spirit. One wonders how Hubble might have reasoned, had he known about fractals which show a thinning of density around every observer.

Hubble tried to explain the deviation as an effect of the expanding, non-Euclidean space. This is nowadays the usual start for understanding the behavior of the galaxy counts, but now we also know that galaxy counts are heavily influenced by unpleasant other factors. Besides the geometry of space and the true distribution of galaxies, the counts depend on the spectra of galaxies and the cosmic evolution of the brightness of galaxies. And counting very faint galaxies is difficult. Thus the counts, though at first sight so promising, are not a reliable way of measuring the large scale spatial distribution of galaxies (nor the geometry of space, Ch.12).

A stronger argument came from the Hubble law. After Lemaître, the models of expanding universes were quickly adopted for understanding the redshift–distance law. Its linearity follows from the uniform distribution of matter (Ch.8). And vice versa, the observed Hubble law was seen as observational evidence for the uniformity of the galaxy universe.

For a long time, data on the true distances to galaxies were scarce and only the projection on the sky was available. Nevertheless, the celestial distribution gave still another indirect argument which comes from the fact that the sky is rather smoothly covered by faint, distant galaxies. This isotropy, together with the principle of no center, is usually taken as evidence for a uniform galaxy universe.

17.5 . . . but the uniformity is elusive

The above indirect arguments were clearly concerned with relatively deep space, because the bright local galaxies over the face of the sky did not suggest anything like uniformity, on the contrary, they exhibited strong clustering. But it was thought by many astronomers that when one looks at large enough volumes of space, the galaxy universe would eventually become smooth.

It is striking how the paradigm of uniform matter has a strong grip on our minds. One has often read over the years how the galaxy universe is locally non-uniform, but on larger scales certainly becomes uniform. A lonely voice in the wilderness, Gérard de Vaucouleurs, saw the oddness of this strong belief, when he wrote in 1970:

In the 1930's astronomers stated, and cosmologists believed, that, except perhaps for a few clusters, galaxies were randomly distributed throughout space; in the 1950's the same property was assigned to cluster centers; now the hope is that, if superclusters are here to stay (and apparently they are), at least they represent the last scale of clustering we need to worry about...

It turned out that superclusters were "here to stay". And the trend continued: during the past decades the evidence for non-uniformity has extended to increasingly large distances.

17.6 Carpenter – de Vaucouleurs's law of galaxy clustering

Already the first inspections of brighter nebulae in the sky showed that they form clouds and clusters. A most interesting finding in the 1930's was that these non-uniformities are not random, but possess a special regularity. The man behind this finding was Edwin Carpenter, who was the director of the Steward Observatory in Arizona from 1938 until the end of his life.

Edwin Francis Carpenter 1898-1963

Carpenter inspected clusters of galaxies and found that their galaxy number density (number of galaxies per unit volume) depends on the cluster size such that the density is smaller in larger clusters. If the density remains the same, the total number of member galaxies would grow as the cube of the size. He calculated that instead of the cube, the population grows slower, as the size raised to a power of one and a half.

Carpenter regarded such a relation as a cosmic restriction so that a cluster of a given extent may have no more than a limited number of members. This relation extends from pairs of galaxies to large systems of hundreds of members, which showed for him that small groups and large clusters do not essentially differ.. He wrote the prophetic words: ... *the objects commonly recognized as physical clusterings are merely the extremes of a nonuniform though not random distribution which is limited by density ...* Nonuniform though not random – this is reminiscent of the objects called fractals!

Carpenter recognized a deep connection where others might see only different clumps of galaxies. However, it was left to another astronomer to

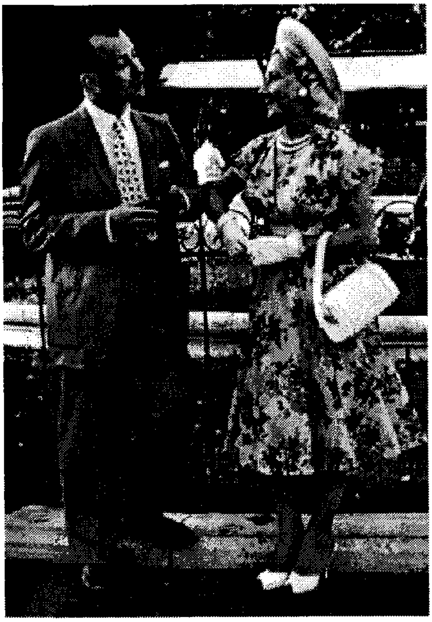


Fig. 17.2 Gérard and Antoinette de Vaucouleurs in Paris, 1962

realize the cosmological significance of the density–size law.

Gérard de Vaucouleurs 1918-1995

Gérard de Vaucouleurs, who moved from France to the United States, was one of the great observers of the 20th century. Starting his career with the planets of our Solar System, his attention moved to distant galaxies, which eventually became his main targets. With his wife Antoinette he initiated the *Reference Catalogue of Bright Galaxies*, an invaluable tool for every student of galaxies. His work on the distances of galaxies and on the value of the Hubble constant led to a lengthy debate between the two schools of "high" and "low" values, which still continues.

Antoinette de Vaucouleurs 1921-1987

De Vaucouleurs was an ardent student of the local universe. He found evidence for the flattened Local Supercluster around the Virgo cluster of galaxies and claimed that its mass causes deviations in the cosmological expansion. Though controversial at the time, now we know that the Hubble law *is* distorted close to the Virgo cluster. Even our Local Group on the outskirts of the Supercluster has been "slowed down" by about 200 km/sec by the massive cluster, having attracted us for billions of years.

Inspired by the work of Carpenter, de Vaucouleurs calculated from new data the density of matter inside galaxy clusters of different sizes. He published results of his calculations and thinking in 1970 in the article "The Case for a Hierarchical Cosmology". He suggested the existence of a universal density – size law in the galaxy universe (with $\alpha \approx 1.7$): [‡]

$$\textit{density} = \textit{constant} \times \textit{size}^{-\alpha}$$

Gérard de Vaucouleurs connected this law with the idea that hierarchic clustering is fundamental in cosmology. But he emphasized the need to replace the original oversimplified regular hierarchies by more general models of statistical density fluctuations, resembling turbulence with its hierarchy of eddies. He intuitively envisioned important, but odd properties which he thought such worlds would possess. *First*, there is "no privileged position in space", which makes such models reasonable cosmology candidates. *Secondly*, in a sense a hierarchic distribution of matter is uniform, because any two separate large and equal volumes (encompassing an element of the hierarchy) on the average contain the same total mass. *Thirdly*, there is the especially puzzling and non-intuitive feature: "All observers, wherever located (but within the hierarchy), will find that the average density decreases" when the depth of observations increases around the observer.

It seems that before de Vaucouleurs these properties had not been summarized with such clarity in print. But his ideas on hierarchy were received with rather little enthusiasm. There were also observational objections. The counts of brighter galaxies seemed to follow the uniform distribution law. It was also pointed out by Allan Sandage that the Hubble law is good all the way from small to large distances, whereas deep inside a hierarchic galaxy system one would expect a strong deviation from the linear relation between velocity and distance. De Vaucouleurs did present evidence for a curved Hubble law in the local universe (the Hubble "constant" increased with the distance). But then it was understood that this curvature was caused by a selection effect (like the Malmquist bias) influencing the distance measurements. In fact, the riddle of the curved Hubble law opened for one of us (P.T.) the gate into the exciting field of cosmic distances.

Perhaps disappointed by the cool reception, de Vaucouleurs did not develop his hierarchic cosmology further and remained silent about it to the end of his life. But he had contacts with the one man who in all the

---

[‡] This result was not far from that of Carpenter which in terms of number density n was: $n \propto N/R^3 \propto R^{-1.5}$.

world understood what he had been up to. This man was Mandelbrot who already had started thinking about the fractal structure of the galaxy universe in the 1960's. "I am not alone any longer!" he exclaimed, when he first saw de Vaucouleurs's article on hierarchic cosmology. He later wrote that the study of galaxy clusters has greatly stimulated the development of fractal geometry. The ultra-massive elements in real space and the points in mathematician's abstract heaven had found each other.

17.7 Mandelbrot's fractal view of galaxy clustering

In 1977, in his book *Fractals: form, chance, and dimension*, Benoit Mandelbrot foresaw that galaxies are fractally distributed and gave the first mathematical description of the fractal properties of such a distribution. He recalls how around 1965, his ambition was to implement the law of decreasing density with a model where there is no "center of the universe".

Mandelbrot views the fractal galaxy distribution as a major conceptual step in the description of the cosmological matter distribution. It is a kind of synthesis of hierarchical structures ("thesis") and homogeneity ("antithesis"), essentially based on randomness. Indeed, there is a fundamental difference between true random fractals and stiff hierarchical protofractals. Into protofractals the hierarchy is injected "ex-nihilo", by defining explicitly its levels. But fractals internally contain a scale invariance (self-similarity) and the impression of a hierarchy follows as an unavoidable consequence. A useful example is the Lévy dust which is created by a random walk process in which the direction of each step is chosen isotropically and the length of a step follows a certain probability distribution (see Fig.17.3).

Fractality carries within itself also a trace of uniformity. Within a fixed radius, i.e. for a fixed scale, every observer counts the same number of elements, *on average*. But upon changing the radius, a "new uniformity" is found with a new mean number density. Furthermore, there is no center for random fractals – this is another "relic" from homogeneity.

Thus Mandelbrot made the first step for genuine fractals in cosmology, generalizing Einstein's cosmological principle corresponding to $D = 3$ by fractality, which allows a non-uniform galaxy distribution with $D < 3$. His "Conditional Cosmographic Principle" states that all observers see similar cosmic landscapes around them, but only under the condition that they make observations from a structure element (galaxy).

It is interesting that the idea of "conditionality" came to Mandelbrot in quite another context, when he studied the temporal clustering of errors in data transmission on telephone lines (see his Foreword to our book). This he did together with Jay Berger at IBM, where Mandelbrot has long worked. Telephone lines seem rather far from cosmology, but in fact the article which was born from this collaboration in 1963 was the first in which the "previously esoteric notion" of fractal dimension was interpreted as a fundamental physical quantity. Mandelbrot writes that "my research, hence my whole life, was changed by accidental circumstances that led to this article". This reminds us of something he said when we were discussing with him at the Mittag-Leffler Institute in Stockhom in November 2001: "I believe in a universe where a set of zero contributes something significant." Sometimes an innocuous event, or a flash of idea in another field, may set in motion an important chain of developments.

"Conditional" in the spatial context emphasizes that each observer occupies a material element of the structure. There is an essential difference between the elements of fractal structure and emptiness. Only *if* the observer sits on a structure point, does he see that the matter distribution follows a fractal law. If the observer for some peculiar reason finds himself in a large void, then the fractal decrease of density can be detected only on scales much larger than the size of the void. The asymmetry between observers on structure points and observers randomly scattered in space is very important. A random point would most likely be in an indefinitely large void (in an infinite fractal universe) and such an observer would be unaware of almost anything.

Mandelbrot's cosmological principle – that the observers attached to the material structure elements are equivalent – is close to what Milne presented. Thus the fractality of the universe perfectly satisfies Milne's Cosmological Principle. It also automatically makes what Igor Karachentsev has called "the ecological correction to the Copernican principle" – the real observer can *live* only on or close to a material celestial body.

17.8 Does isotropy always imply uniformity?

An exact description of fractal structures involves non-smooth mathematical objects. Hence, in the realm of fractals one cannot always trust the conclusions based on the usual techniques of smooth mathematics. A case in point is the proof, discussed above, that the combination of local isotropy and no center implies a uniform matter distribution.

Let us look at Fig.17.1 again, focusing attention on the intersections of the various circles. Usually the density at a point is defined as the limiting value of the mass divided by volume, when one allows the volume to approach zero. But if the density on a circle cannot be defined in such a smooth fluid-like manner, then one cannot ascribe to each point a value of the density. For example, for fractals there is no such limit; instead the density calculated inside a radius around a structural point changes all the way from large to arbitrarily small volumes (though the change follows a nice power law, the hallmark of a fractal). Furthermore, the intersections of two circles may not even contain a structure point!

One realizes that the proof of uniformity, which is based on the density being smooth around all points, does not always work. It is valid for regular distributions (also discrete ones), but not for fractals. It is smoothness which wipes out fractality. Thus strictly speaking from local isotropy and the principle of no center one cannot infer uniformity. But is this important in practice? Yes, because it prompts one to ask if the observed local isotropy, usually regarded as a sign of homogeneity, could be consistent with fractality which is essentially inhomogeneus, though has no center.

Extragalactic radio sources and x-ray emitters smoothly fill the sky. This isotropy is often interpreted as being inconsistent with the fractal space distribution of galaxies, noting that radio and x-ray sources are also galaxies. However, redshift measurements have shown that these galaxies are typically at distances of a thousand or more megaparsecs, so such an argument tells nothing about the more local universe, at distances from 100 to 500 Mpc, where fractality is a hot issue. Furthermore, as we discussed above, isotropy does not automatically imply homogeneity.

How in general would fractals appear when projected on the celestial vault? At first sight, one might anticipate large voids and clusters on the sky if fractals exist up to very large distances. Indeed, first examples of the Lévy dust suggested in the 1970's as models for the galaxy distribution had such large voids that they could not reproduce the observed distribution. This failure promoted general scepticism on the application of fractals in the cosmological context. These early examples were constructed to give a low fractal dimension $\mathcal{D} = 1.2$, and was accompanied by strong clustering in isolated clumps. A higher fractal dimension, as current observations suggest, produces less pronounced clumpiness.

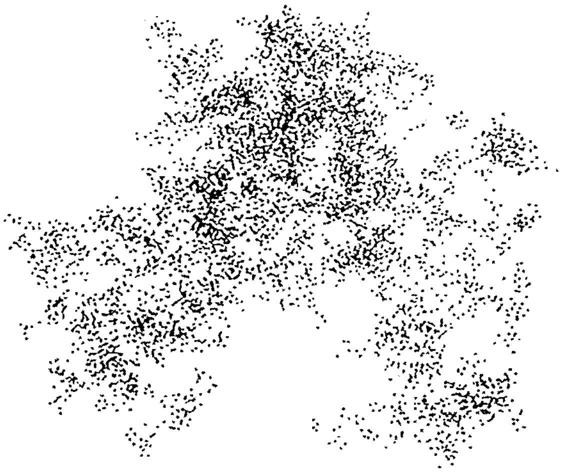


Fig. 17.3 A new generation Lévy dust on the plane, with the fractal dimension 1.26. The lacunarity of this dust, produced by an algorithm recently invented by Mandelbrot, is much smaller than, say, for the simpler fractal in Fig.14.6.

Furthermore, it is now known that the patchiness on the sky depends not only on the fractal dimension, but also on the so-called *lacunarity*, which is a measure of how frequent large voids are. The term lacunarity, introduced by Mandelbrot, comes from the Latin word lacuna, meaning hole. Numerical simulations have shown that fractals with a small lacunarity can have rather smooth projections on the sky.

Another factor which smooths out the patchiness, is the large differences in the luminosities of celestial bodies. As a result two objects with equal apparent brightness actually may have widely different distances. This mixing of nearby and distant objects hides clusters and fills in holes, decreasing the celestial anisotropy for very distant radio sources. Of course, only future 3-D maps for these objects will tell if the isotropy is really due to homogeneity or fractality with low lacunarity on such large scales.

17.9 Do we live on the peak of a mountain?

The emerging fractality brings about surprises for all of us who have used to think in terms of ordinary "ideal gas" -like distributions. For example, the observer sees the density of matter decreasing outwards, as if he stood on the top of a mountain. And this is true for every observer, which radically distinguishes fractality from usual clumpiness. If not aware of fractality, each observer will think that he is in the true center of the universe. They may be content with this state of affairs, until an extragalactic communication network starts working. It will be a shock to learn that everyone has a privileged position.

We have been involved, together with our French colleagues, in a study which addresses the intriguing question: can we see that the density decreases around us as expected from fractality? In fact, we had for years studied extragalactic issues closely related to fractality, even without yet knowing that concept, such questions as the hierarchic distribution of galaxies and its influence on the Hubble law within the lumpy galaxy universe. New fine possibilities were opened by the large KLUN sample of spiral galaxies. As related in Chapter 16, the sample was assembled for measuring the Hubble constant, and distances to all of its 6600 galaxies are known from the Tully-Fisher method. But it can also be used to derive how the average density around us depends on distance.

The method to derive the density is such that it was not possible beforehand to guess the result. Hence, it was quite exciting to see the first diagrams drawn by the computer. At small distances, where there are few galaxies, the derived density fluctuates chaotically. Then around 20 Mpc it starts to follow nicely the density law predicted by a fractal of dimension $D \approx 2$ The average density decreases steeply up to 200 Mpc, the achievable limit of KLUN! So the flattering conclusion seems to be that we live on the peak of a good-sized mountain. However, if you as a cosmologist are not so happy with this eagle's nest, you may accept the more humble view that the Milky Way occupies a mediocre position within the fractal structure.

It is amusing to imagine another possible end result. What if we had inferred that the density increases away from us? Would this mean that the fractal dimension is larger than 3? No, it would not, because in a regular three-dimensional space the fractal dimension cannot exceed three. In that case one should sadly admit that we are living at the bottom of a pit. [§]

[§] That $\mathcal{D}$ cannot be larger than 3 is easy to see for the case of hierarchic point clusters characterized by the number of elements N and the size ratio R. If the fractal dimension exceeded 3, then from $N \propto R^{\mathcal{D}}$ follows that the elements of the lower level making the element of the upper level occupy more space than is available, and must therefore interpenetrate.

Accepting that everyone appears to live on a mountain top is a hard nut to crack for what is called "common sense". Over the centuries we have become used to the idea that common sense is a poor guide in science, and it is often asserted that it actually hinders science's development. There are many examples: the difficulties to accept the idea of a moving, spherical Earth and more recently the relativistic and quantum phenomena. In this context common sense has been seen as the reluctance to adopt new knowledge, or the dogmatic insistence on old views, or the incapability to understand theoretical concepts.

However, there are also positive sides to common sense, better reflected in the corresponding words in the Russian and Finnish languages, which mean literally "healthy sense". Indeed, human beings possess an intuitive feeling for truth, a need for clear argumentation, the ability to discern logical connections, and of course, they are guided by their life experience as a valuable "data base".

Furthermore, common sense is not something that is given forever. It changes with time as a result of the increasing and widening storage of human experience, science, and education. Aristotle and Ptolemy regarded the Earth as inmovable, not because they were reluctant to adopt new knowledge or wanted to brake the advancement of cosmology, but on the contrary, because their intuition, scientific arguments and experimental data were in agreement with the Earth at rest. When understood as a generally accepted world picture or paradigm, common sense always contains obvious things which are replaced sooner or later by other concepts. How to sort those which are of more permanent value from those which are just temporarily obvious, is a great riddle of science.

One might summarize the above by saying that the complexity of fractals (and the wonders of the microcosm, the paradoxes of relativity...) are truly hard nuts to crack, *but highly interesting and rewarding*, for modern common sense. It is only appropriate to admit that the universe – and thus the theories describing it – contains things which may easily outpower our capabilities of imagination. After all, we are living midway between the quantum world and the cosmological realm, neither of which we can freely visit. In the same way as our eyes are optimized to see the maximum in the spectrum of sunlight, our brains may have been constructed to tackle most effectively the problems encountered in our local environment. No wonder that our attempts to penetrate the very small and the very large scale regions of the world are so exciting and challenging excursions!

Fig. 17.4 Luciano Pietronero, Benoit Mandelbrot and Georges Paturel during a discussion at the Sesto Pusteria Workshop (Italy) "Observational Cosmology" in July 1996.

17.10 Modern redshift surveys of galaxies

When de Vaucouleurs suggested in 1970 that there is a universal density-radius law for the distribution of galaxies, his study was limited mainly to clusters of galaxies, for which there were data on sizes and masses. For some time it was not possible to check this law directly, nor Mandelbrot's fractal view, from the distribution of galaxies in space. Then a new and efficient technique arrived. In large parts of the observable universe redshifts of galaxies have been measured in massive surveys, continuing the pioneering work on galaxy clustering by De Vaucouleurs, Zwicky, Einasto and others. From redshifts are derived distances, using the Hubble law. The resulting 3-D maps reveal the real space distribution of galaxies, apart from small errors due to the intrinsic motions of the galaxies.

Presently astronomers possess redshifts for about a hundred thousand galaxies from the Center for Astrophysics survey (CfA), Southern Sky Redshift Survey (SSRS), the Las Campanas Redshift Survey (LCRS), the European Southern Observatory Slice Project (ESP) and other large observing programs are going on. The LCRS team used multi-object spectrographs for the first time to measure redshifts of up to 120 galaxies simultaneously.

Table 17.1 Large surveys of galaxy redshifts from which the fractal dimension has been or will be measured. Ω is the fraction of the whole sky covered by the survey, R_s is the depth of the survey in Mpc for $H = 60$ km s^{-1}/Mpc, N is the total number of galaxies, D is the fractal dimension, as estimated by F. Sylos Labini.

Survey	Ω	N	R_s	D
CfA	0.15	1845	260	1.9 ± 0.2
SSRS	0.14	1773	200	2.0 ± 0.2
LCRS	0.01	26000	800	1.8 ± 0.2
ESP	0.0005	4000	1000	1.9 ± 0.3
LEDA	(≈ 1)	70 000	(500)	2.1 ± 0.2
2dF	0.05	250 000	1000	–
Sloan	0.25	1 000 000	1000	–

The first systematic survey, CfA, extended in space to the depth of about 200 Mpc and together with SSRS covers 1/3 of the sky. It can be used to study the space distribution of galaxies on scales which are less than 50 Mpc (this is the radius of the largest sphere which the survey volume can totally embrace). The LCRS and ESP surveys are narrow slices in space. They are very deep, but cover much smaller areas in the sky, and the maximal spheres have sizes of 10 Mpc only.

Though still a rather young branch of astronomy, such cartography has already brought unexpected news, as we saw in Chapter 16. The galaxies are distributed highly non-uniformly with rich forms of structure. But what is the nature of this non-uniformity? Lev Tolstoi wrote in Anna Karenina that happy families are all alike, while every unhappy family is unhappy in its own way. One may admit that it is much the same with the happy state of uniformity, which is, in a manner of speaking, always similar, but distributions may be non-uniform in many different ways.

17.11 Pietronero and the five megaparsec mystery

The sky peppered with faint galaxies appears relatively smooth. This seemed to support the view held for years that the galaxy universe should be uniform above a scale of a few megaparsecs and clustering could be present only on small scales. This naturally led to the idea that the galaxy distribution can be described by the theory of small fluctuations from a

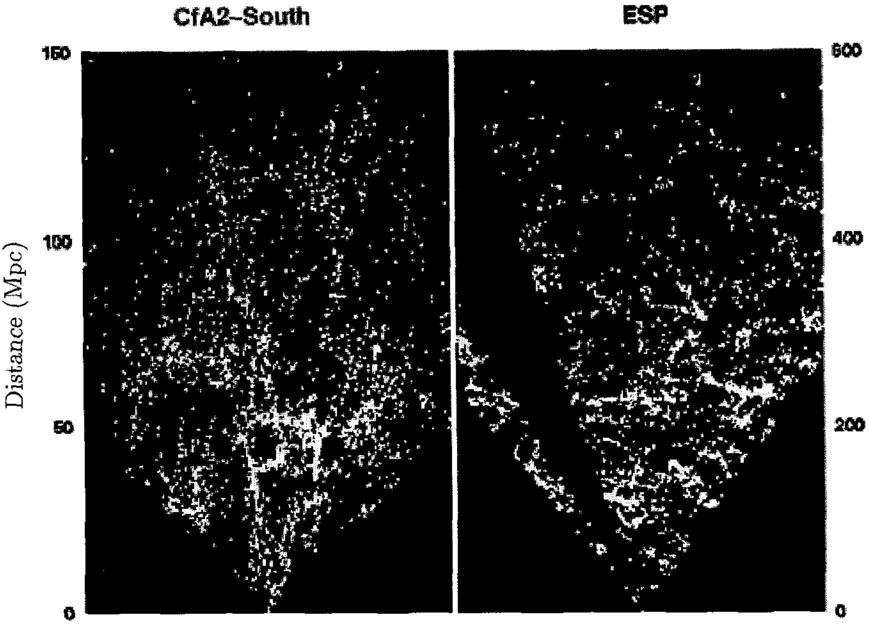


Fig. 17.5 Comparing the CfA and ESP maps. The latter shows that fractal structure continues to a hundred Mpc. Note that at the right panel 5 Mpc corresponds to 0.5 mm, but one clearly sees lumps 100 Mpc across.

well defined average density.

The harmonious picture was badly shaken when the redshift turned the distributions in the sky into 3-D space. The smooth sky was replaced by large clusters and voids extending to the scales of the entire maps. The situation was highly unexpected. The first 3-D space maps were analyzed in the beginning of the 1980's by statistical methods based on the small fluctuation hypothesis. Curiously, even though the structures appeared to extend to very large sizes, the uniformity scale was calculated to be small, a few megaparsecs. This mysterious result was interpreted in various ways, the leading one being to argue that the structures are large but their density over the average is small. So they were considered as compatible with the assumption of homogeneity. The situation was incomprehensible: the mathematical analysis of a truly strong inhomogeneity led to the strange conclusion that the galaxy distribution is uniform! This was the issue that raised the curiosity of Luciano Pietronero. In a letter to the authors, Pietronero recollects those times:

"In 1986 the first slices of the CfA galaxy catalogue became available and I noticed that they showed complex structures with apparent fractal properties. The voids extended at least to 50 Mpc while the clusters and filaments went up to 100 Mpc and more. I was quite surprised to see that the statistical analysis of these distributions led to a correlation length of only 5 Mpc. Note that the interpretation of this length within the mathematical method used was that it should mark the crossover between the scale of strong clustering and an essentially smooth distribution. I was not convinced by the argument of the small amplitude because fractal structures have intrinsically large amplitudes and at least voids cannot be considered as a small perturbation from an average density. So instead of being convinced by the claim that the irregular structures are compatible with the smooth distribution I reasoned in the opposite way: if a mathematical analysis gives this result then there must be something wrong with it."

17.12 The Great Fractal Debate

In December 1999 a hundred physicists and astronomers, half of them mature specialists and the other half young scientists, gathered together on the top of a mountain in Sicily, at an international school on cosmology and fractals, organized by Norma Sánchez from the Observatory of Paris, in the small picturesque town of Erice, said by some to be the oldest in Europe. Twenty years ago Pope John Paul II gave the keys of an old monastery to the European Physical Society for the purpose of discussions of the most important problems in physics.

Erice and the Observatory of Paris are places where the fractality of the galaxy universe has been ardently discussed in recent years, and where the young generation has had the possibility to learn about the new fractal methods in cosmology. Such conferences are one constructive outcome of the earlier debates which culminated in "The Dialogues" of Princeton.

Before 3-D astronomy, when only the blurred projection of the spatial distribution was seen, it was natural that superclusters divided opinions. But even the arrival of the space maps did not calm down the situation.

In 1996 Princeton University, the prestigious seat of learning in a small town not far from New York, hosted an international astronomical meeting with the title *Critical Dialogues in Cosmology.* No wonder that among the

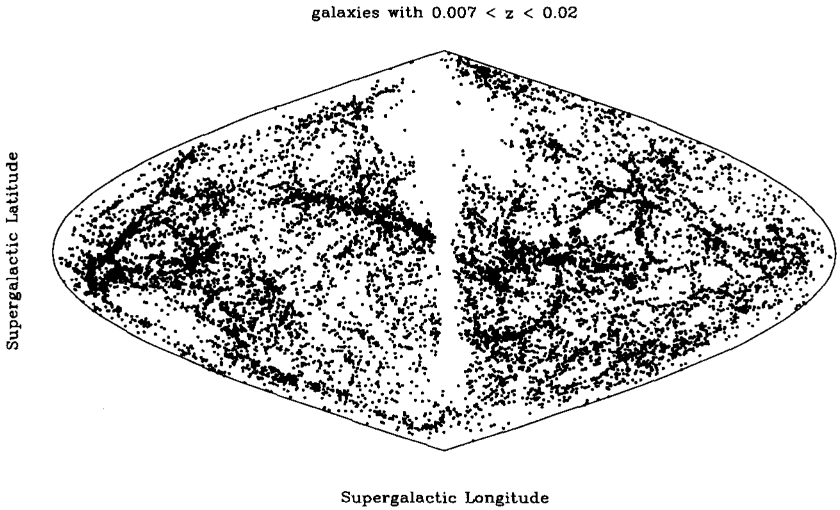


Fig. 17.6 A view of the galaxy sky offered by the LEDA extragalactic data base. We show galaxies with redshifts less than 6000 km/sec. The equator of the map coincides with the plane of the Local Supercluster. Even the projection on the sky reveals other superclusters connected by chains of galaxies. LEDA, created by Georges Paturel at the Observatory of Lyon, is a remarkable merging of data from galaxy catalogues, and provides us with a marvellous chance to study the structures of the local universe. Helen di Nella and Paturel, together with Pietronero's team analyzed a galaxy sample extracted from LEDA. They found that the galaxy distribution has a well defined fractal nature up to a scale of 250 Mpc, with fractal dimension around two.

highlights was the debate on the value of the Hubble constant, with Gustav Tammann defending the "low" value of 55 and Wendy Freedman from the Hubble Space Telescope Hubble constant project speaking in favor of the "high" value of 70. But who might have foreseen some years before that the first dialogue was about the question of whether the universe is uniform or fractal? And that the case for fractals was defended by Luciano Pietronero, who brought from statistical physics and the science of complexity new insights for astrophysics.

A lively debate ensued between the competing visions of two schools. The more conservative view was presented by Marc Davis, a respected specialist on structure formation in the expanding universe. He admitted that the galaxy universe has fractal properties, but only up to distances of about 10 Mpc, after which the cosmic scene levels off. Pietronero pointed out that the traditional statistical methods are inadequate for the observed

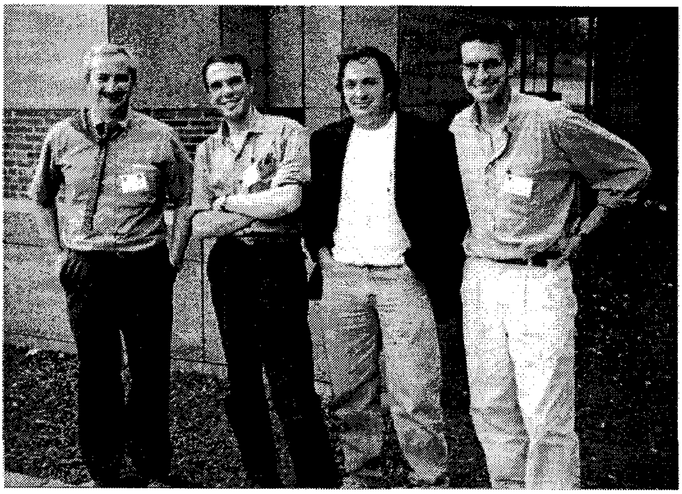


Fig. 17.7 Luciano Pietronero and his collaborators Francesco Sylos Labini, Paul Coleman and Marco Montuori in Princeton, taken during the conference Critical Dialogues in Cosmology, in 1996. This team has brought the fractal structure of the galaxy universe to more general awareness, utilizing new methods of analysis.

strong clustering. Then he argued that the new, more general methods naturally lead to a consistent picture pointing to the radical view that the fractal distribution observed on smaller scales continues to the limits of the 3-D maps (at the time about 100 Mpc). Whether the fractality exists on still larger scales was left open.

A result of the debate was a light-hearted bet between the two gentlemen. If in the next, deeper redshift survey the fractal properties really extend to scales larger than 15 Mpc, then Marc Davis will give to Pietronero a case of best Californian wine. In the contrary case, Luciano Pietronero will present Davis with best Italian wine...

The debate on the fractal properties of the galaxy universe has since continued for years, and the wines only have matured in their bottles. ¶ But what is the meaning of this debate? The main question is how to measure the extragalactic fractal.

¶The specific terms of the bet are found in the Proceedings of the Princeton Meeting. Beautiful images, opinions, scientific arguments, and latest news about the on-going debate on the "fractal versus homogeneous universe" may be found at the home-page of Luciano Pietronero's team: http://pil.phys.uniroma1.it/debate.html

17.13 The correlation function points at 5 Mpc

There are two main methods for analyzing the spatial distribution of galaxies. Usually astronomers apply the so-called *correlation function method.* This approach was initiated by James Peebles of Princeton, a leading theoretician in cosmology, whose comprehensive book *The Large-Scale Structure of the Universe*, in 1980, included a nice review of the early debate on homogeneity and clustering. The method had been applied by the Japanese H. Totsuji and T. Kihara in 1969 in their article "The correlation function for the distribution of galaxies". It was especially developed for investigating small fluctuations from an underlying *uniform* matter distribution.

If in some sample of galaxies the observed number of galaxy pairs is larger than expected from a random uniform distribution (given by the Poisson probability law), it is said that there is a correlation between galaxies. The correlation function measures how the excess number of pairs depends on their separation. (On a dance floor one expects a steep increase in the correlation below a separation of one meter or less!) Usually the correlation (i.e. excess number) is stronger on small separations (scales) and gets weaker towards larger scales. The distance scale beyond which the excess number of pairs is no longer large, is called the *correlation length.* In fact, the exact criterion is that the observed number of pairs is smaller than twice that expected from the random spatial distribution. On scales which are larger than the correlation length the distribution approaches uniformity.

The main result of the correlation function studies is a tendency towards an homogeneous distribution at about 10 Mpc. This method gives two numbers, around which the debate goes:

$$\mathcal{D} \approx 1.2 \textit{ and } R_{max} < 10 \textit{ Mpc}$$

The above description shows that the correlation function method presupposes uniformity already exists within the galaxy sample and knowledge of the average cosmic density. Both are needed in order to calculate the expected number of galaxy pairs. The method works splendidly when one characterizes the fluctuations on small scales in such a "fair" sample of the universe. However, if the galaxy distribution is far from uniform on all scales of the sample, then the correlation function method breaks down. Its application cuts away the large structures and leads to the erroneous conclusion that the scale of uniformity has been reached.

Indeed, puzzling things appear when one applies the correlation function analysis to the available galaxy samples. The derived correlation length is about 5 Mpc, the value which perplexed Pietronero. But in the same 3-D maps one can easily see structures with sizes of 50 Mpc by naked eye (see Fig. 17.5). Another surprise came when astronomers calculated the correlation function for deeper galaxy samples. The deviation from uniformity and the correlation length grew together with the increasing depth of samples. This should not happen, if the samples, both small and large, already embraced uniform parts of the universe

One more puzzle was offered by rich galaxy clusters, which are seen from larger distances than ordinary galaxies, and thus trace a deeper volume. In 1983 Russians Anatolij Klypin and Alexander Kopylov constructed the correlation function for Abell's clusters for the first time. The correlation length turned out to be surprisingly large, 25 Mpc! Are the clusters themselves more strongly clustered than galaxies?

17.14 The conditional density comes and finds fractality

The drawbacks of the correlation function are avoided with another kind of tool. The new approach of the *conditional density* was introduced by Luciano Pietronero in 1987. He is an expert of the complex structures which appear in many natural phenomena like the dielectric breakdown and aggregation phenomena. Here irregularities are the rule on all scales and new mathematical methods are necessary for their description.

Pietronero realized that a similar degree of complexity seemed to be present also in the spatial galaxy distribution. He likes to say that one of the best examples he knows of fractal pattern is the local galaxy universe. For him it is impressive that galaxy fractals are self-similar from about 0.1 Mpc to 100 Mpc, i.e. over a wide range of scales. In laboratory fractals the ratio of maximum and minimum scales is usually ten or one hundred.

The conditional density method is simple in essence. It does not measure the deflection from uniformity, as the correlation function does, but how the average density behaves when the scale increases. A sign of fractality is that the number density (number/volume) decreases around every point according to a typical law. In fact, the word 'conditional' means how to make the counts (c.f. the conditional cosmographic principle!).

First the astronomer makes a three-dimensional map of a part of the

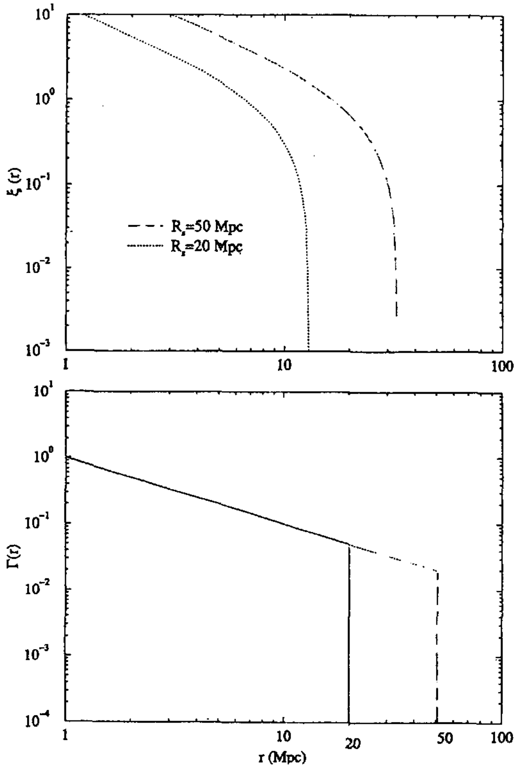


Fig. 17.8 The correlation function (above) and the conditional density (below) shown for two samples with depths of 20 and 50 Mpc within a fractal galaxy distribution. The amplitude and the correlation length increase, as the outer edge of the sample increases, whereas the conditional density just continues to decrease. The slopes of the linear parts of both functions ($\mathcal{D} - 3$) give the fractal dimension of the sample.

local universe around him. For this he must measure for each galaxy a) its position on the sky (direction), which is easy to do, and b) its distance, which is a lot more difficult.

In fact, one must put the surveyor on a galaxy to count the surrounding galaxies inside spheres of different radii: 1 Mpc, 2 Mpc etc... The observer writes down the results, and goes to another galaxy, where the counts are repeated. And so on, until all the galaxies in the sample have been visited. At the end of the journey, the surveyor takes the notes and calculates the arithmetic average of the counted number of galaxies inside the sphere of a given size. This he repeats for all radii. These average counts divided by the corresponding volumes give what is called the conditional density. If this density decreases with radius as a power law with exponent $\mathcal{D} - 3$, then the galaxy distribution is fractal and its dimension is equal to $\mathcal{D}$. [‖]

Fortunately, it is enough to make the counts from the map, without actually flying to other galaxies. To be fair, there are a lot of other practical problems. The maps are not complete, because we cannot see faint galaxies at large distances. The volumes of the maps are not spherical. They are usually narrow cones or slices, which hampers the counting procedure.

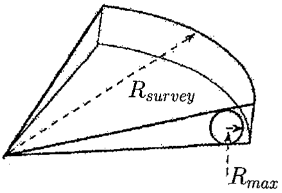


Fig. 17.9 A typical cone diagram for a narrow angle galaxy sample, showing how the maximum sphere fully contained in the surveyed volume is less than the survey depth.

Prompted by the 5 Mpc mystery, Pietronero decided to reconsider all the data on galaxies from the perspective of fractal structures. As a result of this effort, undertaken together with his collaborators (see Fig.17.7), a completely new picture appeared. It was, in his words, like constructing from the pieces of a puzzle the final true composition. The new perspective worked a miracle. Different observations which previously seemed in conflict with each other, were instead shown to be in good agreement. The results imply that the fractal clustering has a larger dimension and the maximum scale reaches the limits of the available maps, in brief:

$$\mathcal{D} \approx 2 \textit{ and } R_{max} > 100 \textit{ Mpc}$$

Another result was a simple explanation for the previous puzzling increase of the correlation length of galaxies and Abell's galaxy clusters. This change simply arises because in a fractal structure the average density insidiously decreases with increasing sample volume (see Fig.17.8).

[‖] The number of galaxies within a radius R increases as $N \propto R^{\mathcal{D}}$. The number density is $n = N/V \propto R^{(\mathcal{D}-3)}$ for fractal dimension $\mathcal{D} \in (0, 3]$.

That no uniformity was positively detected is interesting to compare with the quite diagonal conclusion by Edwin Hubble in the 1930's that "the uniform distribution extends out to the limits of our telescopes". The fractality found there where previously uniformity seemed to be established, illustrates how new observations can dramatically change our view of things.

17.15 To search for or to count on uniformity?

In their laboratories, physicists study samples of matter as external observers, experimenting with them at will. Such a luxury is only a Sunday dream for the astronomer who is doomed to inspect the galaxies from one point only, the Milky Way. Then, depending on the way of treating the data, the astronomer may derive surprisingly dissimilar results from the same observations – just recall the Hubble constant: 50 or 100! Such a thing also happens if a fractal distribution is probed with an unsuitable tool. Typically, there will be contradictory results and a feeling of confusion. This was precisely the situation with the correlation length. Now we understand why it is possible to infer such different results for the spatial arrangements of galaxies. The reason is that astronomers have used two different methods for measuring galaxy fractals, as we just described.

In a nutshell, the conditional density method *searches for* the crossover scale on which uniformity appears, whereas the correlation function method *carries within* itself the assumption of uniformity and, in a manner of speaking, cannot live without it. (Analogously, the method of normalized distances (Ch.16) searches for the sample of galaxies from which the Hubble constant may be safely determined.)

Of course, there are many different techniques on offer at the mathematical methods market which could be used to study the spatial distribution of galaxies. We have discussed only the two which first appeared in cosmology and which well illustrate why it is important to use a suitable method in the new conditions of fractality. Other methods should be carefully tested and applied to real data. Refusing a good mathematical tool for analyzing observations would be similar to declining Galileo's invitation to look through the telescope! A deeper understanding of fractality requires advanced methods which are sensitive to different properties of fractal distributions, such as lacunarity and what the physicists call the phase distribution of the Fourier transformation of density fluctuations. We know, for example, that the conditional density method cannot measure lacunarity.

What we have described above is a reminder of the importance of crossing the boundaries. When a change of perspective is needed, it is not so surprising that this is done by somebody coming from a different field. For Pietronero the study of galaxy clustering was an application of his speciality, statistical physics, to a new area. It is a good idea for scientists to keep their minds open to such interdisciplinary activities, even if these do not appear to conform to the usual approach in the field. Supporting original projects of this type may produce unexpectedly significant discoveries. **

17.16 Towards Einstein–Mandelbrot concordance

We do know of genuinely uniform components of the universe: the photon gas of the cosmic background radiation, the ocean of low-mass neutrinos, and maybe more importantly, the physical vacuum or dark energy. As the average density of the fractal matter decreases with increasing scale, there will eventually be a scale beyond which the density of the uniform component is larger than the density of the fractal component. Hence one may regard, after all, the universe as homogeneous on such scales. However, this is not due to the galaxy distribution, but because of the uniformity of the relativistic matter component!

As to the fractal galaxy distribution, there are two alternatives – a finite or infinite range of fractality. True, there are no scale limits to a pure *mathematical* fractal. The name 'fractal universe' is often linked with an infinite fractal. Such a universe would have zero average density. †† But real physical objects usually have lower and upper cutoffs between which the fractal properties are observed. Thus it is also expected that the fractal galaxy distribution itself changes its nature. One possibility is that it becomes homogeneous on some maximum scale R_{max}. Thus one may, as Mandelbrot did, allow for the possibility that the matter distribution

---

** We like to mention that in the case of Pietronero, he was generously supported by the Istituto Nazionale di Fisica della Materia (INFM) and by the European Community Network "Fractal structures and Self-Organization".

†† Note that a fractal is "asymptotically empty". "Zero-density" does not mean "devoid of matter" and there is no reason to ridicule such a world on that basis. The total mass of an ideal, infinite fractal universe is infinitely large! But the distribution of matter is such that increasingly large structures "require" more and more space in the way that the average density as calculated on large scales becomes less and less.

may become uniform on large scales, while being fractal on smaller scales. Such a universe would have a non-zero average density. Let us examine this point further.

For a fractal structure the average number density of galaxies inside a sphere around any galaxy increases as the radius raised to the power $\mathcal{D} - 3$. If the density comes to constant value beyond a certain sized sphere, then the fractal distribution has turned into uniformity. Now the observer has found the maximum scale of fractality (the upper cutoff distance R_{max}) beyond which the world is uniform in the large ($\mathcal{D} = 3$).

One may visualize such a situation with the help of the Menger sponge (Ch.14). Assume that the initial cube, out of which the regular fractal structure is "carved", is very large. For example, its size might be 500 Mpc. There is some lower level in the construction in which "galaxies" are encountered. Stop the construction at this level. Imagine such cubic "sponges" put side by side so that they fill the whole of infinite space. Loosely speaking, such a (rather artificial) universe would be "super-homogeneous" on scales larger than 500 Mpc, while fractal on shorter scales. [‡‡]

It is interesting to pick up Mandelbrot's feelings on this matter some thirty years ago. He thought that "the evidence is compatible with a degree of clustering that extends far beyond the limits suggested by existing models" and argued that "the distribution of galaxies and of stars includes a zone of self-similarity in which the fractal dimension satisfies $0 < D < 3$." Settling these questions is a major area in present-day astronomy.

At what distance does the decreasing fractal density drop below the constant density of one of the uniform substances and how does it compare with the cutoff distance R_{max} for galaxy fractals? The cosmic 3K radiation is a photon gas which uniformly fills space. Its density is about 10^{-33} g/cm^3. The physical vacuum, another recently suggested uniform component, could have about 10 000 times higher density. In order to calculate equal-density distances, one should know how much dark matter is attached to the galaxies and what the fractal dimension is. The current, still scanty knowledge on the different kinds of cosmic dark matter permits

---

[‡‡] A uniform distribution is not a regular net of points. It is formed when one scatters randomly points into space, so that the average number is proportional to the volume – this is the Poisson distribution. A regular net or lattice is a "super-uniform" distribution and has peculiar statistical properties, as was recently highlighted by A. Gabrielli, M. Joyce, and F. Sylos Labini in connection with the Harrison-Zeldovich spectrum of the primordial density fluctuations.

various distances from where uniformity begins. For example, if it is confirmed that the vacuum density really is close to the critical density, then the universe may begin to be uniform after just a few megaparsecs, long before the possible crossover of galaxy fractals to uniformity.

To our thinking, it is fascinating that both smoothness and roughness have their place in modern cosmological models. The intuitions of both Einstein and Mandelbrot appear to have grasped fundamental features of the universe.

17.17 Everything we know about the cosmos?

In science the collision of ideas does not necessarily result in a crash, but may rather serve to deepen our understanding of the universe. The fractal debate is such a case in which old questions may be seen in the light of new concepts and this may lead to a novel interpretation of the observations.

The cover of New Scientist's 21^{st} August 1999 issue announced: *Fractal Universe – Everything we know about the cosmos might be wrong.* With such a shocking headline the science journalist Marcus Chown expressed himself after following an acute dispute between a leading cosmologist Ofer Lahav and a younger generation astronomer Francesco Sylos Labini.

Lahav presented the view of big bang cosmology where the uniformity of matter is the foundation of the Friedmann model. So he, as most astronomers, believes that fractals stop after a few tens of Mpc. One of the arguments is that there is no dynamical theory which could have produced strongly fluctuating fractal structures on large scales from the very smooth initial state. Another reason is the isotropic sky distribution of distant X-ray sources and of the cosmic background radiation.

Sylos Labini replied that non-uniform fractals in space may look rather smooth on the sky, even if fractality continues up to hundreds of Mpc. This point deals with the lacunarity of a fractal distribution. Moreover he stressed that many of the standard results are based on the assumption of a huge amount of dark matter, which is not directly observed, but is invoked to explain differences in theory and observation. He pointed out that the observed fractal galaxy distribution creates a new and challenging enigma which we may be a long way from solving. But facing a hard problem is far more interesting than hiding it under the rug.

Both disputants agreed that crucial observations for solving the fractal

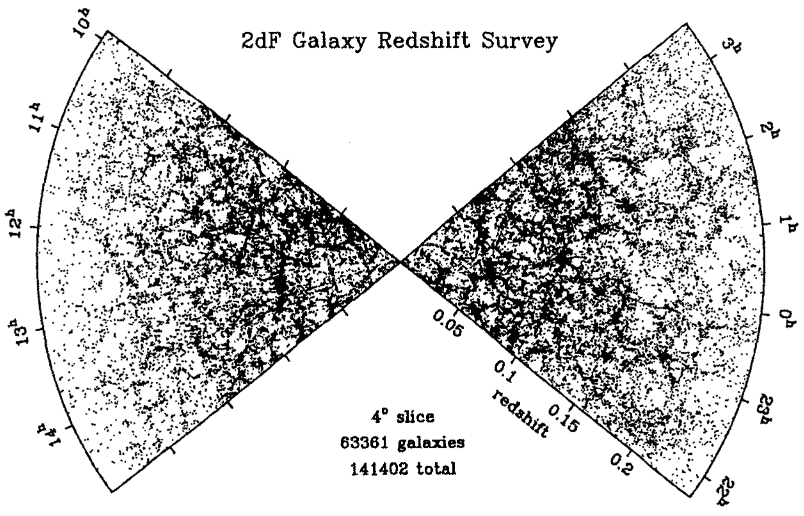


Fig. 17.10 The first results from the 2dF galaxy survey which extends into the depth of about one thousand megaparsecs. A question for the next decade: Do the fractals continue to the scales of gigaparsecs? (Courtesy of Matthew Colless and the 2dF team.)

debate will come from wide-angle redshift surveys of deep galaxy space. We recall about another important point. With uniformity, one should keep in mind that the universe is known to be comprised of several substances: the vacuum and other relativistic components, the dark matter, and the ordinary luminous matter. The vacuum is always homogeneous, whereas dark matter and galaxies may have a zone of fractality and a crossover to uniformity. As related above, the fractal galaxy universe and the uniform relativistic medium may live happily together. They both satisfy the Cosmological Principle, being without a center.

17.18 Opening the millenium: the race to a fair sample

The new century has started with a burst of remarkable scientific studies devoted to the largest galaxy structures ever seen. Three types of independent evidence have come from the deepest observations of galaxies, rich galaxy clusters, and quasars. These suggest that there is a continuous clustering hierarchy up to several hundreds of megaparsecs.

The first evidence came from the *2dF Galaxy Redshift Survey* which uses

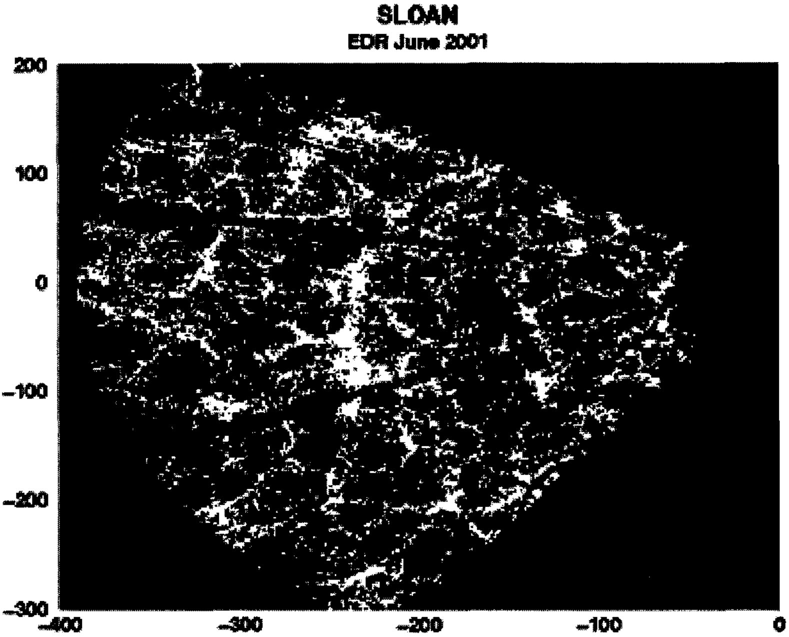


Fig. 17.11 The first release of the Sloan galaxy survey shows galaxy structures up to hundreds of Mpc.

the 4 meter Anglo-Australian telescope located in the mountains northwest of Sydney. Its name comes from the two degree wide field of view of the telescope. The survey will measure the redshifts of all galaxies down to a rather faint magnitude (19.5 mag) and consists of two parts. One in the direction of the North Galactic Pole covers an area of 75° × 7.5°. Its southern counterpart is 75° × 15°. When completed, 2dF will contain the impressive number of 250 000 galaxies having redshifts up to ≈ 0.15. The first analysis of about 100 000 positions of galaxies showed an increasing spectrum of density fluctuations up to 300 Mpc.

The second surprise was brought by new observations of rich clusters of galaxies. Astronomers use clusters to study the large scale structure because bright galaxies in their cores are easy to observe from large distances. Again as has happened before, the deeper survey showed a continuous growth of density fluctuations up to 600 Mpc.

The third news arrive from the deepest survey of objects ever in astronomy – the redshift survey of quasars, or the 2dF QSO survey. It uses the same telescope as the 2dF galaxy survey, but observes the most luminous objects in the universe, which give a chance to study the deepest regions of space. The first part of the survey now contains 10 000 quasars and the structure analysis reveals that the density fluctuations grow up to 900 Mpc.

These recent discoveries show that the last word on the space distribution of galaxies on the largest scales has not been uttered. We do not yet know where the fractal galaxy distribution turns into uniformity and how deep we must go in order to reach Hubble's fair sample, embracing at least one cell of uniformity. Only within such a volume can we calculate how small a fraction of the total cosmic mass is contained in galaxies, and measure the size of the biggest structures in the universe.

The decive step for understanding the nature of the distribution of galaxies is expected from the Sloan Digital Sky Survey. Funding for the project has been provided by the Alfred P. Sloan Foundation which was established in 1934 by Alfred P. Sloan, then President of the General Motors Corporation. This American-Japanese programme is measuring the redshifts of one million galaxies brighter than 19^{th} magnitude, using a special telescope situated at Apache Point Observatory in New Mexico. Its mirror has a diameter of 2.5 meters and its field of view is about 3 degrees in the sky. A state-of-the-art spectrograph attached to the telescope can measure simultaneously the redshifts of 640 galaxies in one exposure. During one succesful night the 30 silicon electronic light sensors of this very complex camera produce 200 gigabytes of data!

The Sloan survey covers over a quarter of the sky, which allows one to test the fractal structure in spheres having sizes comparable to the depth of the survey, about 500 Mpc! In July 2001 the first release of a small part of the Sloan survey became available. We give the wedge diagram for these data in Fig.17.11, definitely showing structures with sizes of about 100 Mpc. However, a rigorous analysis is a matter for the future.

In order to study gigaparsec (1000 Mpc) scales, one must go very deep into space. However, from very large distances one can detect only the most luminous and rare objects, such as quasars. This makes the investigation of the super large scale universe difficult, because one can still see only the "tip of iceberg". Matters are complicated by various cosmological effects which start to influence the observations and their interpretation. For example, the volume of the space cone we are filling with quasars can no longer be calculated from a simple linear relation between distance and redshift –

we have entered the realm of high redshifts and competing cosmological models.

The Sloan survey will collect from the sky about 100 000 quasars with the aim to study the structures on scales of about 1000 Mpc. The cosmic map made with quasars will help to answer the long standing question: is there really a crossover to uniformity on scales of gigaparsecs?

* * *

The first truly three-dimensional extragalactic maps have disclosed a rich variety in the structural forms of the galaxy universe. This richness is reflected in the descriptions which grace the pages of astronomical journals: binaries, triples, groups, rich, regular and irregular clusters, walls, superclusters, voids, filaments, cells, soap bubbles, great attractors, clumps, concentrations, associations... Of course, all such individuals are exciting subjects of study, but they can be also viewed as different appearances of one master entity called the fractal.

The success of fractal geometry in describing what Lundmark called the "very complicated structure in the doubtlessly gigantic universal system" is the necessary first step towards the understanding of the origin of large scale cosmic formations.

Chapter 18

The Origins of Megafractals

The megafractals – the cosmic continents, archipelagos and islands – were the news brought home by the modern explorers of the cosmos, exotic, but truths nevertheless about the worlds overseas. Even if their fractal dimension and maximum scale are still debated, megafractals cry for explanation. Their origin is the number one challenge for cosmological physics.

As the largest known structures are approaching the scale of the whole observed universe, fractal geometry is becoming a natural part of modern cosmology. This leads us to ask how cosmological models are built and of which building bricks? What kinds of models are compatible with megafractals? How do fractal structures emerge and evolve? How deep in the remote past are their seeds? What do they really tell us about our universe, the enigmatic unity of space, matter and energy?

* * *

18.1 The ladder of key discoveries

Astronomical observations play the role of experiments for cosmological physics. Three of them are especially crucial as cosmological key discoveries:

- *the Hubble law of cosmological redshift*
- *the blackbody 3K cosmic radiation*
- *the fractal structure of the galaxy universe*

Three generations of astronomers were needed to bring these unexpected aspects of the universe into our consciousness. The law of redshifts was discovered in 1929; the thermal ocean of photons was detected in 1965; and finally, at the verge of the new century, the fractality was unveiled.

The law of cosmological redshift was the first major discovery of extragalactic astronomy. A great majority of astronomers accept space expansion as a natural interpretation of the Hubble law. This view, based on many indirect evidences, is a solid part of the current cosmological paradigm. Allan Sandage has demonstrated, amidst difficult issues of evolution and selection effects, that high-redshift galaxies look as pale as predicted by the expansion. The value of the debated Hubble constant is now placed into the range 50 – 70.

Results of 20th century cosmology

Phenomenon	Observations	Derivations
Hubble law	Local galaxy universe Intermediate universe, SNIa	$H_0 = 50 - 70$ $\Omega_\Lambda \approx 0.7$, $\Omega_m \approx 0.3$ $T_0 \approx 16 \times 10^9$ years
Cosmic microwave background radiation	Blackbody temperature at different redshifts Anisotropy distribution	$T(0) = 2.725 \pm 0.001$ K $T(z=2) \approx 9$ K $\Omega_{total} = 1.01 \pm 0.02$
3-D galaxy distribution	Redshift surveys	$R_{max} = 10 - 100$ Mpc $\mathcal{D} = 1.2 - 2.5$

The cosmic background radiation was the next step on the ladder of key discoveries. It brought the Nobel Prize to Arno Penzias and Robert Wilson in 1978, and is of exceptional significance for cosmology and physics. The most accurate measurement in cosmology is that of the temperature of the cosmic radiation. Deeper in space the temperature does seem to be warmer, as expected from the expansion. The observations of its slight patchiness favors a universe with zero curvature.

The third main phenomenon is the fractal space distribution of galaxies, revealed by means of the first discovery, the Hubble law. The galaxies are clustered on a wide range of scales, and the clusters reveal self-similarity, which calls for the mathematics of fractal geometry.

Of course, many other observational facts also are important for cosmology, e.g. the chemical composition of visible matter, the evolution of galaxies in cosmic time, and collisions and mergers of galaxies.

18.2 The three whales of cosmology

Browsing through popular books and encyclopedias, one may find many different definitions of cosmology from "...myth or a system of belief to explain the workings of Nature..." to "the sum total of all the answers that we have found up to the present moment". For astronomer cosmology is something more special. Astronomy studies celestial bodies in the sky. Cosmology attempts to draw a picture containing everything in the world, also such parts which we never can observe. Its subject, the entire universe, is an extraordinary object, to say the least.

The exploration of reality modifies how we look at cosmology. After one thousand, nay, one hundred years, cosmologists may speak of matters very different from ours. But, perhaps the definition from The New Encyclopaedia Britannica will still be valid: "Cosmos, in astronomy, the entire physical universe consisting of all objects and phenomena observed or postulated." The aim of cosmology is to find a theory which explains and ties together observations believed to be cosmologically relevant at that time.

The science of the universe rests on the three "Whales of Cosmology": Principle, Observation, and Theory. They symbolize three intertwined aspects in any attempt to build a picture of the large scale world.

Principle is the fundamental assumption for cosmologists. Perhaps the most general one, since Giordano Bruno, is that the physical laws found on Earth are valid everywhere. Some may hide in the shadows, not even recognized to be principles, like the daring assumption that cosmology is possible at all. Some assumptions are rarely stated aloud, such as the conviction that ordinary, analytical mathematics adequately describes reality. Some principles have a special status: in order to bridge the known and unknown one has to postulate something called a cosmological principle, which is like extending our knowledge to regions from where we can never obtain information. Nowadays the cosmologists base their work on Einstein's principle of uniformity. Mandelbrot's cosmological principle of fractality has begun to raise interest, too. A universe based on fractals is not at all "unprincipled", as it is sometimes accused of by the adepts of uniformity. Fractality means order within chaos, which is the same for all observers, hence the Copernican principle is valid.

Observation, or empirical fact, may originate from simple visual exploration of our environment or from the sophisticated methods now used by astronomers. Observing is of utmost importance – the scholar may build elegant theories of the universe in his lonely cabinet, but only comparison with Nature will decide their validity. Principles and theories cannot live a true life without observations. What would we now do without the galaxy universe, the Hubble law, and the 3K cosmic radiation? With all our respect for principles and theories as an integral part of our science, we should not forget the good advice of Thomas Digges, from the Renaissance, when the new cosmology was emerging: Rather than progressing in inverse order from theories to seeking after true observations, one should proceed from observations and then examine the theories.

Theory is the physics known from laboratories, and extended by the aid of principles into the realm of the cosmos. * Cosmological models are based on classical, relativistic and quantum laws of modern physics. Models, as mathematical constructions, describe an ideal world imagined on the basis of principles, following the dream of Pythagoras who drew the first circles for the planets. In the usual spirit of science, models are confronted with the other side of reality, observations. As an ancient Roman Law states: *Audiatur et altera pars*, or "Both sides should be heard". Testing the predictions of the models is the cosmologists' daily bread.

18.3 The art of making universes

One may construct a variety of cosmologies, but it is clear that many of these will not live for even a day. A cosmological model should agree with known physics and old observations. In Richard Feynman's words: "The problem of creating something which is new, but which is consistent with everything which has been seen before, is one of extreme difficulty". Furthermore, new, unexpected observations may force one to adjust or even throw away a model.

Thomas Kuhn 1922-1996

Indeed, from time to time, established scientific views are shattered by novel evidence or fresh ideas, and the scientists have to change radically their lines of thought. This is called the change of *paradigm*. Introduced by the philosopher of science Thomas Kuhn, this concept means roughly the

---

*It is interesting to note that the word "theory" has its roots in the Greek "theoros", coming from *theos* (God) and *ora* (to look, to watch). One may think of a theory as an attempt to view God (or perhaps His Creation). In cosmology, theory and models are like mental antennae which Man pushes into the regions ever unreachable even with the largest telescopes. Though they are by no means mere flights of fancy – they must correctly describe the observable part of the universe.

Fig. 18.1 Cosmological models are cooked according to specific recipes. This kitchen with its cornucopia of cosmic provisions was imagined by Alexander Shneivais.

fundamental assumptions that the students are taught in the universities and that scientists base their work on. A paradigm may collapse if its groundwork is seriously challenged by new observations or theories, and it may be replaced by a new paradigm which then rules and inspires the scientific work of the following generations. This is how science advances, says Kuhn. Such a thing happened with the geocentric Ptolemy's system when the Copernican revolution transformed the universe with the unique center into our present abode without a preferred place.

Beside stage shaking jumps, there are also more peaceful periods in the evolution of science. The scientific heritage of the past is continuously growing. Old theories do not just evaporate without a trace, but may find their true home. For example, Newtonian gravity theory remains an excellent instrument in modern astronomy. Relativistic gravity shows where one should halt relying on the laws of Newton.

Our present cosmological thinking may be characterized by a few questions which were either unimaginable or intractable a century ago, but which now must be addressed by a reasonable cosmological model:

- *What is gravity and space–time?*
- *How is matter distributed in space?*
- *What is the nature of the cosmological redshift?*
- *What is the message of the thermal background radiation?*
- *What determines the arrow of time?*

Answers to these questions are the building bricks from which cosmologists make their "castles in the air". The big bang model has its particular recipe. General relativity is the gravity theory, applicable for the whole universe. Matter in all its forms is uniformly dispersed in space. The Hubble law appears on scales where the uniform matter distribution exists. The cosmic radiation originates in the hot beginning before the birth of the stars. The universe is evolving during the expansion of space.

18.4 Art is long, life is short

Fred Hoyle 1915-2001

The famous old rival of the big bang, the steady-state theory, also illustrates how a model uses the cosmological building bricks. The steady-state cosmology, some aspects of which remain fascinating, was invented in 1948 by Hermann Bondi, Thomas Gold, and Fred Hoyle working at Cambridge University in England. Their *Perfect Cosmological Principle* states that the universe is not only *uniform in space*, but also *uniform in time*, i.e. has the same appearance at all times.

The recipe of the steady-state cosmology differs essentially only in one respect from that of the big bang: no global evolution is accepted. The uniform distribution of galaxies is always the same. The redshift is due to space expansion. The gravity theory is geometrical, almost the same as general relativity, except for one extra effect. This "C-field" provides for the continuous creation of matter, which keeps the density constant, in spite of expanding space. This process has gone on eternally. The rate of matter creation is about one hydrogen atom per cubic meter every five billion years. Though each star sooner or later burns up, newly formed stars (and galaxies) are continually replacing those which have passed away.

As the density of matter remains the same, there is no beginning, no hot

Fig. 18.2 Fred Hoyle standing before the massive ancient stone circles of Stonehenge. This place tells about the interest in matters celestial thousands of years ago. The photo appeared in Hoyle's very nice autobiographical book *Home is where the wind blows.*

big bang, no primordial photons... instead the cosmic radiation is thought to have its roots in the energy released by nuclear reactions inside stars. The energy appears at surfaces of stars in the form of hot (6000 degrees K for our Sun) photons which stream out and fill space. The photons must be transformed into the cool microwave radiation at about 3 degrees K. It was proposed that this task is performed by cold dust particles in intergalactic space. The dust catches hot photons and gets a little warmed up. Then it radiates, as every warm body does, at a low temperature.

The discovery of the thermal cosmic radiation in 1965, so natural for the hot big bang, made the steady-state implausible for most astronomers. In spite of this blow the theory's development has continued "on the back burner". In 2000 Fred Hoyle, Geoffrey Burbidge, and Jayant Narlikar presented the results of their long term study in *A different approach to cosmology.* The new quasi-steady-state model allows for evolution in the form of discrete creation events: matter begins as Planck particles which decay in "fireballs" to baryons. The repeated fireballs are said to serve as sites for nucleosynthesis of light elements. However, it is a pity that the new model has lost some of the elegance of the original steady-state, which lived an interesting, brave life amongst the hostile observational facts.

18.5 Growth of large scale structures in big bang cosmology

According to big bang cosmology galaxies and their structures were gradually formed after the gravity of tiny primordial seeds of matter started to collect surrounding stuff. Such initial graininess is thought to have had its origin in the very early inflation epoch, as fluctuations of the repulsive quantum fields which were speeding up the expansion. The gravitational growth could start only after the matter became neutral and got rid of the grip of the radiation, some 300 000 years after the big bang.

But could the tiny grains grow into something so dense and big as galaxies? In fact, this requires that the initial density fluctuations were about 0.001 times the average density. And how to sweep away galaxies from large parts of space and leave behind huge voids? Such an heroic deed needs a lot of time. The observed galaxy streams, sped up by gravity, have speeds of 500 km/sec. Thus, within the age of the universe the galaxies may travel no farther 10 Mpc, much less than the observed 100 Mpc voids. Hence, the seeds for the large structures must already have existed at the time when the cosmic radiation originated. And as was pointed out by Joseph Silk and James Peebles in the 1960's, the seeds for the galaxies alone should have imprinted relative granularity of ±0.001 on the background radiation. [†] Why then is the sky one hundred times smoother?

Here dark matter again shows up. In order to marry the observed smoothness of the cosmic radiation and the present-day lumpy fractality of the galaxy distribution, one has to assume the whole process has been controlled by dark matter which takes over the growth of the megastructures.

Yakov Zeldovich 1914-1987

As was first shown by Arthur Chernin, the cosmologist from Moscow University, the dark matter seeds are not only the initial reason for clustering of visible matter, they can even perform this trick without leaving fingerprints on the microwave sky. This fine idea was published in 1981 by Chernin. The energetic leader of the Moscow cosmology school Yakov Zeldovich, legendary for his strict attitude concerning the level of scientific work, in this case quickly recommended the publication of the article.

[†] The expansion of space and gravitational contraction make a density enhancement $\delta d/d$ grow like the scale factor S: $\delta d/d \propto S(t) \propto 1/(1+z)$. In order for a seed to become a galaxy, $\delta d/d$ should have grown to about 1 (at least) when galaxies are already observed around redshift $Z \approx 1$. Hence, when $Z \approx 1500$, $\delta d/d$ should have been about 1/1500. As these seeds were accompanied by roughly similar temperature variations $\delta T/T \approx \delta d/d$, one would now expect to observe temperature fluctuations of around $\Delta T/T \approx 10^{-3}$.

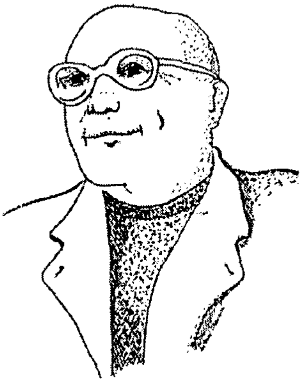


Fig. 18.3 Yakov Zeldovich as seen by Arthur Chernin during a seminar in Moscow in the 1970's. At the time he was developing his "pancake" theory of large scale structure formation. In this scenario flat walls (or pancakes) form first and only afterward do galaxies congeal within them. An alternative picture was proposed by James Peebles, in which small stellar systems are first created and only then do galaxies and larger structures appear as a result of gravitational growth.

Thus the structure formation theories hide the fluctuations in the background radiation by assuming two things: firstly, the usual baryonic matter, tied to radiation, was initially uniform, and secondly, there was dark non-baryonic matter through which radiation easily slipped. Even if this dark matter is non-uniformly distributed, it does not leave its signature on the cosmic radiation. The dark matter is envisioned as making the invisible seeds around which all large scale structures in the universe are formed.

To study the process of structure formation, cosmologists use supercomputers and extensive N-body simulations of gravitational interactions between millions of particles. These advanced calculations have largely replaced the earlier approximative theories developed by the cosmological schools of Zeldovich and Peebles.

The N-body models can produce fractal structures on small scales (< 10 Mpc) in which the clustering is "non-linear", i.e. the density peaks become higher than the average smooth density and local gravity is the dominant factor. On larger scales the expansion of space dominates and it is rather the initial conditions which determine the resulting structures. This means that it is critical for the structure formation computations to have full control over the initial positions and velocities of the particles at the starting point

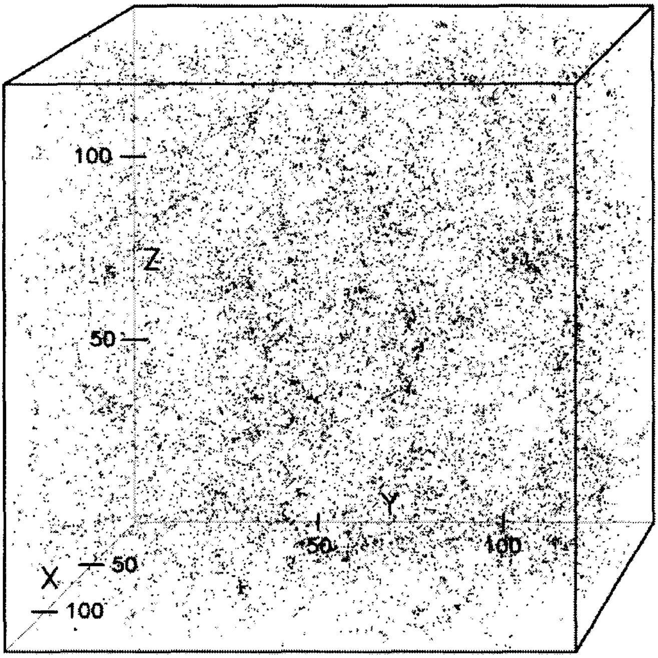


Fig. 18.4 A simulation of the large-scale structure within the Big Bang model with cold dark matter ($\Omega_{CDM} = 0.3$) and dark energy ($\Omega_{\Lambda} = 0.7$). This computation was made by the VIRGO consortium, and the picture was designed by Mikko Hanski. The cube is about 200 Mpc across and contains 15 000 galaxies, the end results of the gravitational interaction between several millions of dark matter masses.

of the evolution. As a whole, the clustering in big bang N-body simulations is a result of several effects: space expansion, the initial conditions, the properties of dark matter, and the self-gravity of the initial density seeds.

In order to compare the theoretical results with the observed clustering of galaxies one must introduce yet one more effect, "biasing". The real producer of the lumpy distribution is the dark matter, but galaxies do not necessarily follow the dark matter. Biasing quantifies the difference. For example, galaxies may have preferably formed on the shallow peaks of the more smoothly distributed dark matter, leaving vast volumes with low density empty of galaxies.

Performing the N-body simulations and comparing the results with observations of the galaxy universe is a very complicated field of astronomy, which is still evolving. But even on the face of the various sources of uncertainty and the possibility to add dark matter into the recipe, tune the initial conditions, and adjust biasing, it is still impressive that this theoretical apparatus has produced landscapes which much resemble the real ones "out there". It might have happened that the universe is too young and the background radiation too smooth for such structures to come out from any calculations based on gravitational instability in an expanding universe.

18.6 The smooth Hubble law ignores local roughness

"Uniformity implies the Hubble law." This explanation of the Hubble law has always been regarded as a great success of the expanding uniform universe. And as has been noted during the debate on fractals, if the universe were non-uniform, one would not expect to see a uniform expansion, but an irregular pattern of velocities of galaxies, resulting from the matter lumps. Arthur Eddington, in his 1932 book The Expanding Universe, even applied his ever-useful balloon to this issue too, writing that that "the inflation is only uniform if the density is uniform" and in a roughened or pimply sphere "the roughened parts do not expand at the same rate as the smooth intervals between them".

The conditional density analysis has revealed the fractal distribution of galaxies on scales starting from 1 Mpc and going up to at least 100 Mpc. In this same distance range (and beyond) we find the Hubble law. Thus there seems to be two cosmological relations, the Hubble law of redshifts and de Vaucouleurs's fractal density law (Ch.17), both universal in the sense that they are valid for every observer.

The peaceful coexistence of the Hubble and de Vaucouleurs laws is very puzzling, because deep inside the very inhomogeneous fractal structure the Hubble law is expected to break down as a velocity–distance relation. The gravity of matter lumps pulls the galaxies off their permanent positions in expanding space, gives them extra velocities in one direction or another, and thus distorts any regular expansion law.

One may calculate how unexpected this coexistence of fractals and the Hubble law is. Suppose that all cosmic matter is in galaxies and the maximum scale of fractality is 100 Mpc, beyond which the universe is uniform. In such a model the Hubble law is strictly valid only for distances larger than 100 Mpc. Now its fate in the near universe will depend on the density. If the density parameter is equal to 1 (inflationary universe without dark

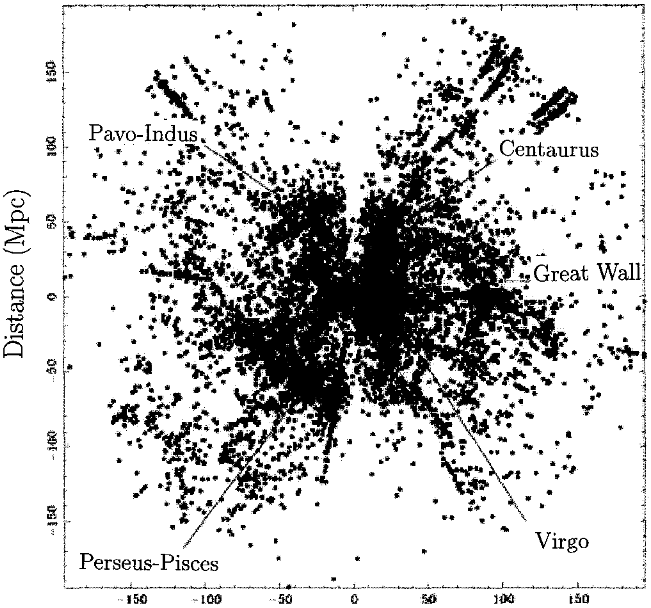


Fig. 18.5 Megafractals revealed by the LEDA galaxy database. Such a local map up to 300 Mpc will be the basis for more detailed study of the structures and galaxy streams.

energy), then there should be no cosmological redshift closer than 25 Mpc. But it was at these small distances where Hubble first found his law! [‡] We do see the good Hubble law much closer than the Virgo cluster, down to the outskirts of our Local Group. What does this paradox tell?

First, the paradox would be resolved if the average density of all the matter in the universe were much less than the critical one, for example with the density parameter of not 1 but 0.01. But, such a low density falls short of the density value 0.3, into which different cosmological tests appear to converge. Another possible explanation is a uniform dark substance. Its mass should dominate on all scales in which the linear Hubble law is

---

[‡] Such calculations were first made by J. Wertz and M. Haggerty in 1972. More recently we have, with Alexander Gromov, used Tolman-Bondi models for spherically symmetric masses. Space expands according to the Hubble law on large scales where the uniform matter distribution is reached. On small scales where the average mass density is larger because of fractality, the expansion is slower, and on still smaller scales any regular expansion is destroyed by gravitational braking.

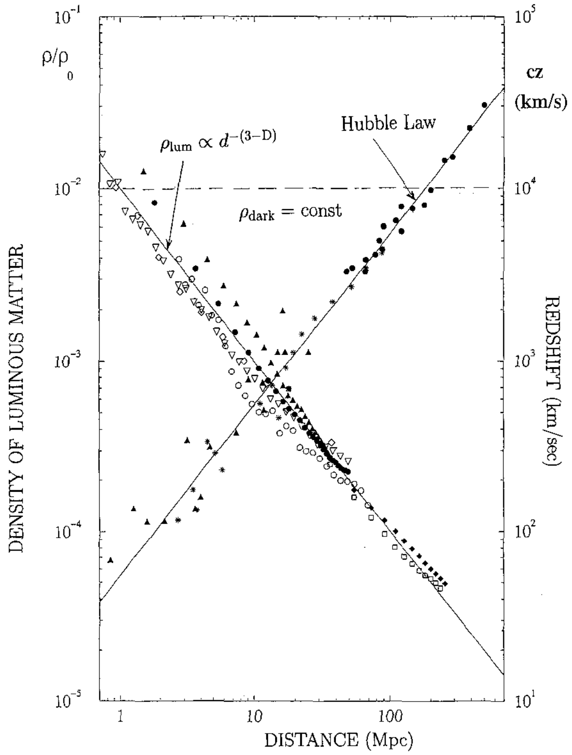


Fig. 18.6 The Hubble – De Vaucouleurs paradox or the peaceful coexistence of two main cosmological laws: the Hubble law of redshift (increasing to the right) and the fractal law of decreasing average density. The uniformly distributed dark energy is shown as the horizontal dashed line. The lumpy galaxy structures of larger and larger scales sink deeper and deeper into the pacific ocean of dark energy.

observed, starting from the neighborhood of the Milky Way. The required amount of dark smooth matter is about 100 times the visible matter in galaxies. Though such an ocean of dark substance, within which galaxies are sailing as rather insignificant pieces of baryonic pollution, may be at first sight a rather tempting view, there is little evidence for it. On the contrary, observations of gravitational lensing suggest that dark matter rather follows galaxies. However, if dark matter cannot do it, there is the fascinating possibility that the recently discovered dark energy, accelerating the universe, may be the missing piece in the puzzle!

18.7 Gravitational redshift inside a fractal structure

The Hubble – de Vaucouleurs paradox may point to two opposed possibilities, depending on which law is primary for cosmology. From the view point of the standard model, the Hubble law is fundamental. In this case, as explained above, there must be a uniform substance producing the linear Hubble law. Then the observed fractal structure would be confined to the sparse luminous matter and hardly affect the dynamics of the universe.

But perhaps it is the fractal law which is primary, encompassing all matter! Then the reason for the Hubble law must be something other than uniform matter. Could the Hubble law have some deep link to the fractal law? And if uniform matter is not needed for its chief function, to explain the Hubble law in expanding space, should we question space expansion as well? Let us play the heretic for a moment... If there is no space expansion, then redshift could be caused either by ordinary motion or by gravitation. The latter possibility is especially interesting because the gravity of a fractal structure may appear as a redshift in the light from distant galaxies.

Sir Hermann Bondi showed, in 1947, in his classical article about the inhomogeneous cosmological model (later called the Tolman-Bondi model) that the total redshift of a galaxy may be regarded as containing two parts: firstly, a velocity shift and secondly, a global gravitational shift due the mass distribution between the galaxy and the observer.

The mass in a sphere around a distant galaxy increases as the square of the radius, if the fractal dimension is two. If one applies the formula of gravitational redshift (Appendix A.5) for this case, one sees that the gravitational redshift should follow the linear Hubble law:

$$\textit{gravitational redshift} = \textit{constant} \times \textit{distance} \qquad (\mathcal{D} = 2)$$

The observed value of the Hubble constant is close to the value of the "gravitational Hubble constant", if the average matter density within the Hubble radius is not far from the critical density of Friedmann models. Then why is the gravitational redshift not a widely accepted interpretation of the Hubble law? Alas, this mechanism demands much more mass in the nearby universe than has been observed. A huge mass of about ten thousand galaxies within a typical cluster of radius one Mpc would be required to explain the value of the Hubble constant as a pure gravity effect! This paradoxical end result again illustrates how dark matter stubbornly appears from various cosmological reasonings.

18.8 A Friedmann universe with fractal galaxy distribution

The three key discoveries – the Hubble law, the cosmic background radiation, and the large scale fractals – have been recently united in a new kind of Friedmann model. Here the role of the uniform substance for the model is played by the cosmic radiation, while the fractality of visible matter achieves scales where its average density is less than that of the photon gas. This simple sounding, but novel and imaginative model was built by the team of astronomers Michael Joyce, Marco Montuori and Francesco Sylos Labini, and physicists Philip Anderson and Luciano Pietronero. Philip Anderson, a Nobel-prize winner from Princeton and an advocate of the science of complexity in different fields of physics, now has entered the physics of cosmic fractals.

The major property of a fractal distribution – that the average density decreases for increasingly large scales – makes it possible that the matter density becomes less than the radiation density. The density parameter of the cosmic radiation at the present epoch is $\Omega_{CR} = 7 \cdot 10^{-5}$. For fractal structures spreading over a few thousand megaparsecs the average density of clustered matter becomes less than the very small density of the uniform photon sea. This permits a Friedmann model that contains the fractal structure extending beyond the Hubble distance (5000 Mpc).

An interesting feature of this model is that it describes the universe without the dominating dark matter. It contains a new scenario of large scale structure formation, in which the fractal seeds are sown in the very early universe. But one prediction is possibly not confirmed by recent observations of the fluctuations of the cosmic background radiation. The patches are expected to be a few angular minutes across, which is much less than the one degree observed in the Boomerang experiment.

18.9 Dark energy drives the remote and the local universe

As originally presented, the low density fractal Friedmann model did not include a cosmological vacuum which nowadays is an essential part of cosmology. What happens if one adds the high density vacuum to the fractal model, making a universe with the critical density ($\Omega = 1$)? Now the dominant component on large scales is not the radiation, but the vacuum or, more generally, dark energy. All attractive features of the open fractal

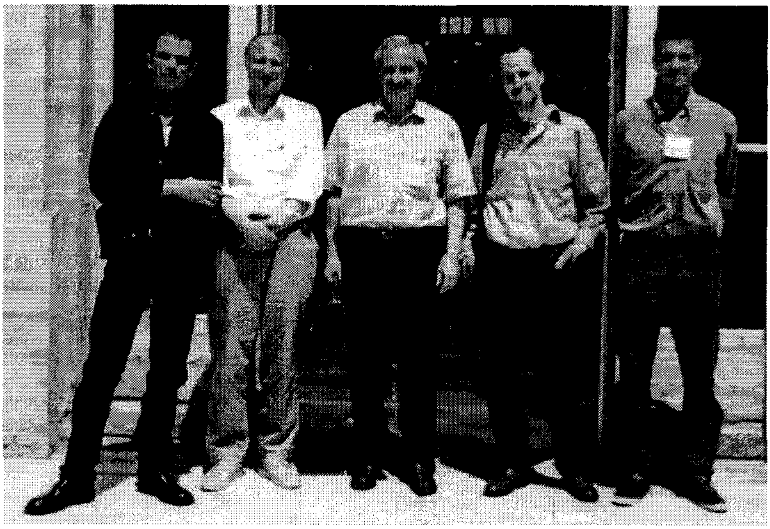


Fig. 18.7 Francesco Sylos Labini, Yurij Baryshev, Luciano Pietronero, Michael Joyce, and Marco Montuori in Paris Observatory during the Sixth Paris Cosmology Colloquium in 1999. Here the origins and properties of the megafractals were lively discussed.

model remain in the vacuum-dominated world. Furthermore, this model predicts the one degree sized patches in the sky as detected by Boomerang and, of course, is in better agreement with distant supernovae.

Also the enigma of the local Hubble flow, existing deep within the fractal distribution of galaxies, may have its solution in the frame of the fractal Friedmann model if one takes into account the vacuum or dark energy. Then one can calculate the distance of the "zero-mass surface" around the Local Group. Inside this sphere with a radius of 1.5 Mpc, as we mentioned in Chapter 11, the normal gravity rules, while outside it the vacuum determines the expansion of space and produces the linear, undistorted Hubble law. Between the matter concentrations there are vacuum-dominated regions, "pacific oceans", where fractal structures are carried on by the Hubble flow, with the expansion rate (the Hubble constant) the same everywhere. Such an explanation of the smooth local Hubble flow has emerged from an original idea by Arthur Chernin, developed in collaboration with the authors. One may imagine that Arthur Eddington, the ardent advocate of Einstein's cosmological constant, might have been happy with this solution of the enigma.

18.10 Early work around fractal dimension one...

The protofractal world designed by Fournier and then Charlier and Selety may have different fractal dimensions. Their motive for the particular value of one was the desire to avoid the infinite gravity and blazing sky paradoxes of Newton's universe, coming from the uniform star distribution in infinite space. For $\mathcal{D} = 1$ the escape velocity from an element of any hierarchy level is finite. [§] It is much less than the speed of light, if the size of an element is much larger than its gravitational radius. And the fact known to Fournier that the velocities of stars indeed were small, inspired him to expect that such mass distribution is valid for the universe as a whole.

Fred Hoyle, a pioneer in many fields of modern astrophysics, developed, in 1953, the first model of hierarchical formation of stars and galaxies. In the article "On the fragmentation of gas clouds into galaxies and stars" he considered a cascade process starting from a uniform hydrogen gas cloud. Due to Jeans's instability this cloud will start collapsing and subelements will emerge within it. The number and size of the elements, which determine the fractal dimension of the hierarchy being born, depend on the cooling processes in the gas. Hoyle's hierarchical fragmentation process naturally generates structures with fractal dimension $\mathcal{D} = 1$.

Within the steady-state cosmology, Hoyle and Narlikar considered a hierarchical model with size ratio of the elements $R = 3$ and number of subelements $N = 8$, which gives $\mathcal{D} = 1.9$. This 1961 study was especially concerned with the counts of radio sources, which at the time were seen as the main argument against the steady-state. In the new quasi-steady-state model, the large scale structures have a non-gravitational origin (this mechanism is much less studied than gravitational clustering). Mini-creation events, with the probability of emergence of new matter higher around existing matter, are expected to lead to a fractal-like spatial distribution.

18.11 ...and intriguing aspects of fractal dimension two

The modern cosmic maps suggest that the space distribution of galaxies has fractal dimension close to two, at least on the scales from 1 Mpc to about 100 Mpc. Is this fractal dimension just an accident? Perhaps there is special physical significance in $\mathcal{D} = 2$?

[§] The square of the escape velocity is $v_{esc}^2 \propto \phi_N \propto GM/R =$ constant for $\mathcal{D} = 1$.

In the 1920's Knut Lundmark was inclined to a hierarchy which follows Charlier's second criterion, corresponding to $\mathcal{D} = 2$. The scarce observations did not provide any convincing evidence, but Lundmark liked of the thought that "Nature has ventured the utmost in order to reach the best possible results. Just as the temperature of the human body is close to the temperature of the coagulation of the albuminous substance, the world is constructed in such a way that it is not far from collapse on account of the total attraction being near one of the limits. Anyhow, the three systems of different orders we now know, the stars, the galactic, and the metagalactic systems are so arranged that they fulfill the above found condition."

One of us (Yu.B.) advanced arguments for why the galaxy distribution could be fractal with dimension $\mathcal{D} \approx 2$ in 1981. In fact, at that time the concept of the fractal was not yet well known in the Soviet Union and the book by Mandelbrot was practically impossible to obtain. The available articles by Wertz and de Vaucouleurs utilized regular hierarchical structures à la Charlier, characterized by the dilution parameter (which is simply related to the fractal dimension). Both theory and observations made the dilution corresponding to $\mathcal{D} = 2$ an attractive possibility.

Theoretically interesting was the gravitational redshift within fractal structures, which in the case $\mathcal{D} = 2$ produces a linear redshift-distance relation, similar to the Hubble law (as we discussed above). Observations included the galaxy counts and the virial mass versus size relation for galaxy systems from binary galaxies to superclusters. In particular, the virial masses (which included both the luminous and dark matter) were based on then unique data on "The average characteristics of galaxy systems and the problem of the existence of hidden virial mass" published in 1968 by Igor Karachentsev, who was a pioneer in observational studies of dark matter (and, by the way, a student of Victor Ambartsumian). It turned out that the hidden mass was closely proportional to the square of the size of a system, or in the language of fractals $\mathcal{D} = 2$. Recently one of us (P.T.) analyzed the modern extensive data on binary galaxies collected by Karachentsev and showed that in order to make consistent the orbits of binary galaxies with their massive haloes, one again has to assume a size–mass relation with $\mathcal{D} \approx 2$.

Another context where $\mathcal{D} \approx 2$ appeared was, in 1986, L. Schulman's and P. Seiden's study of galaxy formation. In their model, galaxies were born in a way analogous to the way star formation is thought to propagate inside a galaxy. The birth of one galaxy stimulates the birth of nearby galaxies from the primordial cosmic gas. If this process occurs near its percolation threshold then a hierarchical structure is created. Recall from Chapter 14 that the critical fractal dimension for the percolation process is about 1.95.

Hannes Alfvén 1908-1995

The Swedish physicist Hannes Alfvén, the Nobel Prize Laureate in 1970, proposed a cosmological model in which a highly ionized gas (i.e. plasma) and magnetic fields play a central role. "Because in the beginning was the plasma". ¶ Though there are arguments against this model as a whole, it is fascinating that an essential feature of Alfvén's cosmology is the inherent lumpiness of the matter distribution. In plasma phenomena inhomogeneities and their hierarchies occur spontaneously. When discussing galaxy formation in plasma cosmology, Eric Lerner showed in 1986 that a plasma gives birth to magnetic filaments with a universal mass–size relation, again corresponding to $\mathcal{D} \approx 2$.

An especially important feature of the fractal dimension two, which distinguishes it from $\mathcal{D} = 1$, is that $\mathcal{D} = 2$ is a critical value as regards the projection on the sky. For all larger dimensions the fractal will isotropically fill the celestial sphere. Such fractals with a low lacunarity can also easily represent the observed galaxy distribution in space, overcoming the old objections to the first fractal models, based on $\mathcal{D} = 1.2$.

18.12 The fractal state of many gravitating particles

Looking at the hierarchic structures formed by gravitating particles, one is reminded of some passages in Arthur Koestler's *The Ghost in the Machine* from the "pre-fractal" epoch in 1967, in which he captivatingly wrote of hierarchic systems in biology and psychology. He coined the word "holon" (from the Greek word for "whole") to denote sub-wholes or elements of complex hierarchies which "behave partly as wholes or wholly as parts, according to the way you look at them". Koestler likes to speak about the "Janus effect". The elements of hierarchy, like the Roman god Janus, all have two faces staring in opposite directions. One face looks "down" at the subordinate levels, as a self-appointed master. The other gazes "up", being a dependent part, a servant. In fact, every master is also a servant, and every servant is also a master. As a servant, an element feels the gravity of

---

¶The standard cosmology also has plasma in the early universe, but it is considered that after the recombination epoch, when matter became neutral, gravitation took the dominating role and magnetic fields were important only locally.

the other elements on the same level, while as a master it is supported by the internal gravity of its own servants. But could such a hierarchic order live for long?

The problems of the evolution, stability and preferred fractal dimension of a large system of particles moving under their mutual gravity force are not yet solved. Only a few studies have been made, but these suggest that the fractal dimension 2 appears to be a theoretically preferred value in a self-gravitating gas.

Jean Perdang suggested in 1990 that for a fractal formed by gravitating mass points a stable configuration is possible, but only if its fractal dimension has a critical value which is equal to or less than 2. A few years later (in 1996) Daniel Pfenniger, when he studied the cloudy interstellar medium, constructed a hierarchical model where on every level the elements have reached virial equilibrium. He found that systems dominated by long-range forces prefer to adopt a fractal dimension which is lower for a longer force range. For gravity the fractal dimension must be less than two, while for short-range forces the state can not be hierarchical and the dimension must be close to three.

A team of French astronomers, Hector de Vega, Norma Sánchez, and Françoise Combes published in 1996 in *Nature* an interesting result which seemed to agree with Perdang's computations. In a theoretical study of a gas of self-gravitating particles, they concluded that at statistical equilibrium the gas is not uniform, but forms structures with fractal dimension equal to either 2 or 1.59. Their work applies both to the local galaxy universe and the interstellar medium inside galaxies. [‖]

Then an important simulation study of clustering in gravitating N-body systems was made by Pietronero's team at Rome University. In order to separate the effect of mutual gravity from the effect of cosmological space expansion they considered static Euclidean space and about 30 000 particles, using a supercomputer. The process of structure formation was found to start from small initially present random density fluctuations, which grew into self-similar granular structures. They developed from small to large scales eventually pervading the whole volume, as may be seen in

[‖] More exactly, they showed that a self-gravitating ensemble, in a quasi-isothermal regime, can be viewed as being in a critical state, in the sense of phase transitions: there exist multiple fluctuations on all scales, which correspond to the structures observed, either interstellar clouds or galaxy groups. Renormalization group theory was applied to predict their fractal dimension.

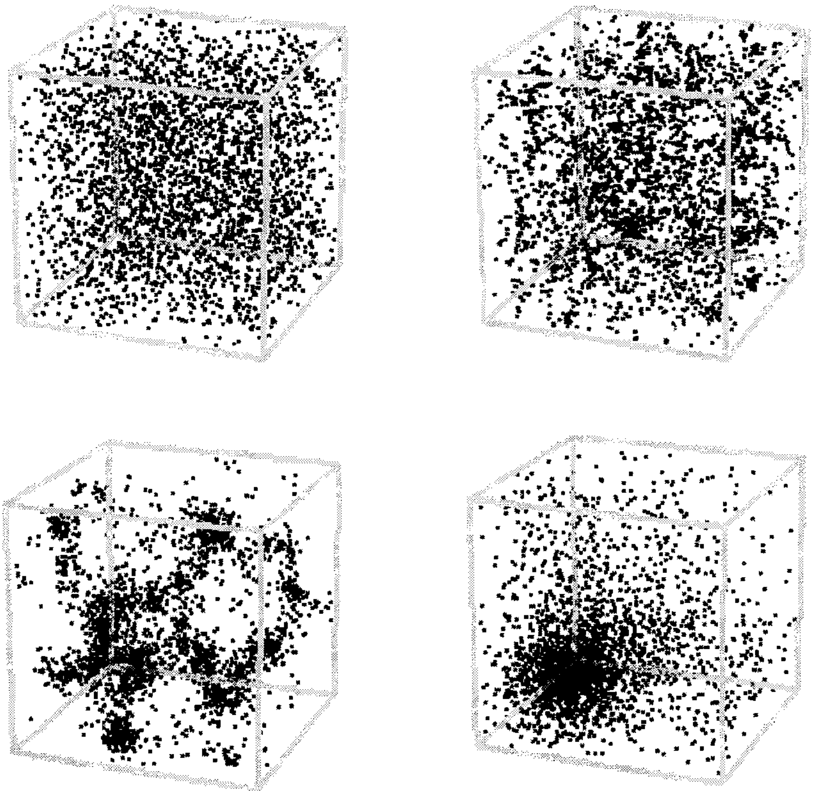


Fig. 18.8 N-body simulations of the structure growth under Newton's law of gravity. Randomly placed gravitating particles (upper left) begin to produce structures on all scales, as time progresses (courtesy of M. Bottaccio, M. Montuori, L. Pietronero).

Figure 18.8. Though the system was not found to approach a definite statistical equilibrium, the forming structure quickly started to follow the fractal law with $\mathcal{D} \approx 2$. Due to various computational limits, in particular the finite size of the simulated volume, it is still an open issue of how long such structures could exist in an unbounded universe.

The magical number two may be qualitatively understood by realizing that the Newtonian gravity force in such a fractal system is constant for all distance scales. If the fractal dimension is larger than two, then the gravity force from large-scale fluctuations will dominate over the force from small clusters. In the opposite case ($\mathcal{D} < 2$) short-scale fluctuations dominate.

Hence, $\mathcal{D} = 2$ is the critical fractal dimension for which all clusters "democratically" influence all particles equally. ** The conjecture of Lundmark that fractal dimension $D = 2$ is a kind of critical dimension for the universe may now be better appreciated.

We note that the old argument in favor of $\mathcal{D} = 1$ is no longer needed in the Friedmann model in which there has not been enough time to speed up galaxies. Even in a static universe, as was shown by Holtsmark (Ch.3) for uniformly dispersed matter, the velocities of galaxies remain low, thanks to the isotropy of the mass distribution.

18.13 Cosmological questions within quantum field gravity

What about the quantum field gravity theory and cosmology? The forthcoming observations of gravity waves and compact relativistic objects will show how deeply this alternative to general relativity could be involved in cosmology. For the moment, one may just deliberate upon some interesting questions. If one considers fractal structures with $D = 2$ then the energy density of the gravity field (virtual gravitons) causing the gravity force will be distributed uniformly like a gas in an equilibrium state. †† This may be a way to understand why self-gravitating gas tends towards a distribution with $D \approx 2$. It should be emphasized that such an analogy is only possible within a theory in which the energy of the gravitational field exists.

The global expansion of space and the Hubble law follow elegantly from general relativity, but pose remarkable problems for the static Minkowski space of field gravity. The relation between these two gravity theories, in Feynman's words, is not something readily explainable, and remains so. But it is tempting to speculate about a possible synthesis of the geometrical and field approaches. The possibility to pass over from geometry to field and vice versa (in the weak field conditions) vividly reminds us of the near-identity between the Friedmannian and Newtonian equations for expansion.

**For a fractal structure the mass within radius R is $M(R) \propto R^D$, hence for $D = 2$ the mass of a sphere around every galaxy increases as the square of its radius, while Newton's force *decreases* inversely proportional to the distance squared. Hence, the growth of force due to the increase of mass is compensated by the increase of distance, and the total force remains constant.

††The energy density of a gravity field is given by $\epsilon = \frac{1}{8\pi G}(\nabla\phi)^2$. In the case of a structure with fractal dimension $D = 2$, where $\phi \propto R$, it follows that $\epsilon =$ constant.

The interplay between the classical Newton, the relativistic Einstein, and the quantum-relativistic Feynman is a most puzzling triangle drama in modern cosmology.

18.14 The cosmic architecture of complexity

In 1972 the milestone article "More is different" by Philip Anderson brought to light the idea which has now evolved into a new branch of science called complexity. Its tenet is that "reality has a hierarchical structure in which at each stage entirely new laws, concepts, and generalizations are necessary, requiring inspiration and creativity to just as great a degree as in the previous one". This has given a quite fresh view of natural phenomena. The focus on the elementary bricks of which matter is made is not sufficient, when the bricks are put together in marvelous structures. Complexity is the study of such highly structured architectures, which have laws and properties which cannot be predicted from knowledge of the bricks themselves. It is a novel interdisciplinary field which includes topics ranging from physics to geology, astrophysics, economics, and biology.

A good example of complexity is seen in the problem of computing the orbits for a system of two gravitating particles and a system of three such particles. The two-body system has simple and easily imaginable solutions, while solving the system of three bodies has turned out to be an entire complex science in itself. As we discussed in Chapter 15, the modern theory of chaos emerged from Poincaré's deep analysis of the three-body problem, which cannot be grasped without a study of the fractal properties of the dynamical strange attractors. The complex fractality of the large scale structure of the universe appears to reflect the collective phenomena generated by huge numbers of gravitating particles.

Complexity, in the sense expressed by Anderson, means that even under the action of the same fundamental force, large systems display novel properties which have to be discovered. But the study of the universe is further complicated by the fact that such complexity occurs on the background of widely different scales with different dominating fundamental forces and laws. Furthermore, scientific understanding is gained at the cost of simplifying the picture of Nature so that one may apply to its phenomena a mathematical model. But after some time Nature herself will protest against, so to speak, this tight clothing. New observations and experiments lead to new theories and models. And sooner or later Nature again overflows the adopted mask – reality is slipping through your fingers. Indeed, a real object, be it a piece of stone picked from the street or a galaxy seen through a telescope, contains in itself the rich depths of physical reality.

It is tempting to see here a glimpse of the ancient Indian view of "infinite heterogeneity" as described by Konrad Rudnicki (Ch.17). In this context it would mean that one cannot catch the universe into one definite model: new, unexpected things appear when we study the universe deeper, the border of the unknown draws away and expands both into the macrocosm and the microcosm. It is more than just a metaphor to say that widening knowledge is accompanied by increased contact with the unknown. Experience shows that virgin landscapes are always encountered and novel questions, previously unuttered, appear, when science advances.

Nevertheless, the wish to understand the whole physical universe resides in the cosmologist's heart, though he may have a nagging feeling that it is more like a dream than a realistic aim (even with "whole" suitably defined!). Recall how Newton's world, based on old physics, suffered from many paradoxes. And with all its success, the infinite Friedmann model, filled with all those types of matter and dark energy, still contains that inexplicable singularity, a gate, as it were, into something much more. The new physics, with its relativistic and quantum laws will eventually shed light on the primordial arena in which the seeds of today's galaxies grew. The next step, in the unpredictable future, would take us even closer to what we call the singularity, shrouded by the misty Planck epoch.

At the frontier of knowledge, we like to imagine the things in the simplest possible way and are only later compelled to draw a more complex picture. In cosmology it does appear that the Friedmann model, smooth thanks to its relativistic components, well describes many current observations. But does this mean that we have finally reached a level of reality where ideal uniformity is the last word? Or are we still, as it were, imagining Plato's heaven of regular forms? The reasonings which led to the inflationary theory, carry within themselves the dream-like idea that what we call our universe is just one regular bubble among the myriads of other universes which have burst out of the primordial quantum soup. Such a fantastic superuniverse would be extremely patchy, infinitely heterogeneous!

18.15 What is the message of the megafractals?

The current picture of cosmic fractals has been expressed in a nutshell in a review article by Kelvin Wu, Ofer Lahav and Martin Rees in *Nature*: "The Universe is inhomogeneous – and essentially fractal – on the scale of galaxies and clusters of galaxies, but most cosmologists believe that on larger scales it becomes isotropic and homogeneous." Nevertheless, the field is so young that even in the latest astronomical encyclopedia one does not find the headword 'fractal'. For the present, the debate rages on the largest scales up to which fractality continues. Here we wish to highlight a few insights and implications springing from the studies of megafractals.

- It is generally agreed that from snowflakes on Earth to the cosmic honeycombs of galaxies fractals are an essential part of the picture. And megafractals show that the "competition" between smoothness and roughness in Nature continues on still larger scales. Remarkably, the Cosmological Principle can embrace both uniformity and fractality. The question posed by Johann Lambert on the possibility of having both "subordination" (hierarchy) and "democracy" (no governing center) can now be answered in the affirmative.
- The detection of megafractals is based on giant collections of galaxy positions in space and on the application of adequate mathematical methods, such as the conditional density analysis. The study of cosmic fractals is a superb example of the advantage of an interaction between different branches of science, which involves a joint effort of astronomers, physicists, and mathematicians.
- Gravity can explain the formation of galaxy fractals on scales from 1 to 100 Mpc. This is a remarkable achievement, as fractals appear in so many phenomena, but their complexity rarely allows a theoretical explanation in a pure form in terms of fundamental forces. This reminds us of the old prediction by Immanuel Kant that the origin of the whole present arrangement of the world edifice, will sooner be understood (on the basis of gravity) than the production of a single herb or of a caterpillar.

The fractal links observed phenomena and the theory which explains them. It is a descriptive notion, which inherently contains the principle of self-similarity, but says nothing about the physical mechanism building the fractal. In other words, in the spirit of Mandelbrot's view of Nature, one may say that the fractality of galaxies can serve both downstream and upstream. Downstream, fractality accounts for the impression of hierarchy. Upstream, one anticipates a deeper explanation for the fractality, for example, by the action of the force of gravity.

What new physical insight could galaxy fractals provide? William Saslaw ponders also this question in his fundamental book The Distribution of the Galaxies – Gravitational Clustering in Cosmology. He expresses the hope – so far only a hope – that "the multifractal description of galaxies may represent a strange attractor that is the nonlinear outcome of the dynamical equations of gravitational galaxy clustering".

- The observed Hubble law can no longer be looked at as a result of the uniform *galaxy* distribution – the fractal lumpiness of the visible (and part of the invisible) matter rather unveils some underlying, unknown smooth substance on which the Friedmann model is resting. It is fascinating that simultaneously with megafractals, astronomers have found evidence for the cosmological vacuum or dark energy which could provide such a longed for uniformity.
- Megafractals are no longer needed for explaining the darkness of night, but with their deep roots in the past they will eventually shed light on the Dark Age when galaxies were in gestation, and even on the era before the photons of the cosmic radiation were emitted. N-body simulation of gravitating particles is the chief instrument for the study of the growth of structures from primordial seeds. Gravitation, the cosmos, and fractals naturally appear together.

As we have seen, different processes may be involved in the building up of cosmic fractals. Depending on the physical state of the matter, the evolutionary phase of the universe and the scale of the structure, one or the other generating mechanism may dominate. If fractality continues to gigaparsec scales, such "gigafractals" might have their roots in the distribution of the primordial seeds themselves. We have only just started drawing the whole picture of cosmic fractals and their origins.

An example of the ideas springing from fractals is the interesting attempt, by the Indian physicist Burra G. Sidharth, to construct a "jagged coastline" universe from the smallest microcosm scales up to the largest cosmic scales. He argues that there is a universal relation between the size of a system and the number of its elements ($R \propto \sqrt{N}$), holding from the

Planck scale upwards. Sidharth's vision is in the spirit of John Wheeler's view of physics as regularity based on chaos, expressed by Wheeler as "I believe everything is built higgledy-piggledy on the unpredictable outcomes of billions upon billions of elementary quantum phenomena, and that the laws ... of physics arise out of this chaos by the action of a regulating principle, the discovery and proper formulation of which is the number one task..."

18.16 Through deeper observations to novel perspectives

A rich source of new observations and ideas, the lonely astronomical observatory under dark skies is our true link to deep space. The pioneering efforts by Edwin Hubble and Allan Sandage led to the observational approach to cosmology in which the part of the world which can be explored with existing telescopes becomes our entrance-hall to the universe.

Nowadays cosmology is often viewed as the study of what happened in the first minute after the creation of the universe. Such work, a faraway excursion into what cannot be directly observed, is attractive for the theoretical mind. Observational cosmology, in the spirit of which our story was told, rather starts from our vicinity in space and time. Observation gradually deepens our knowledge of the local universe and enlarges the small but precious region in which colossal world models make contact with reality.

Astronomy is truly getting three-dimensional, and it's happening both in extragalactic space and even closer to home. Because of the lack of good distance indicators inside our Milky Way, it has been impossible to make reliable 3-D maps of its stellar, gaseous, and dust populations. This will change radically during the next decade. The European satellite GAIA will measure precise parallaxes for millions of stars throughout the Milky Way, so that finally we can study our home galaxy with an accurate map.

During the coming decade the spatial distribution of luminous matter will be mapped with unprecedented accuracy from all-sky surveys of millions of galaxies and thousands of quasars. These surveys are made, not with the new telescope heavyweights, but with modest-sized new technology telescopes. The biggest are not always the best! But in order to penetrate the Dark Age, beyond redshifts of 5, and study the early evolution of galaxies and their structures, astronomers need very large telescopes on the ground and in space. The successor to the Hubble Space Telescope, the 8-meter Next Generation Space Telescope is scheduled for launch in 2009.

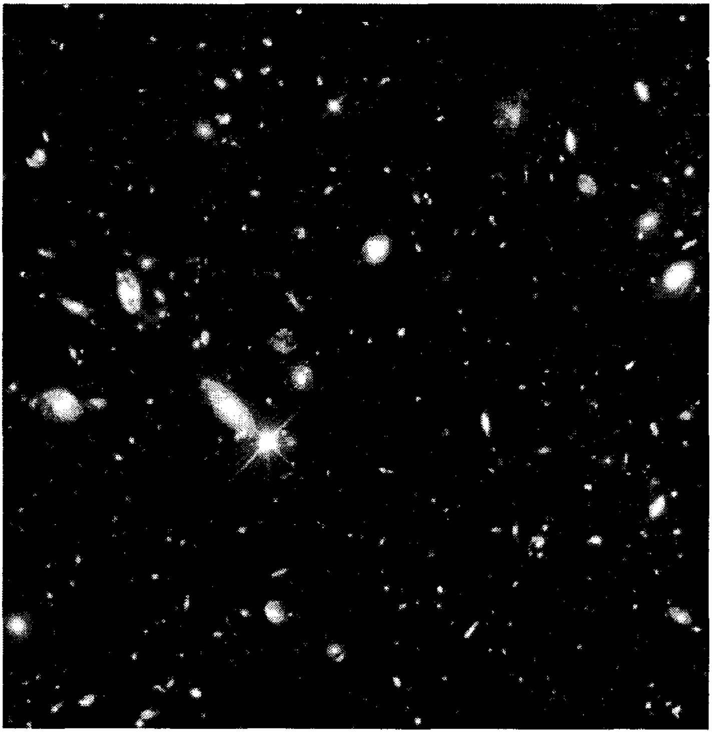


Fig. 18.9 This deepest ever view into the galaxy universe was offered by the Hubble Space Telescope. When we look at the sky, we no longer see the blue jasper-stone or the revolving crystal spheres carrying the stars. Our thoughts fly deep in space and back in time, beyond the galaxies to the enigmatic beginning which fills us with awe and wonder.

Hubble's and Sandage's programme of probing Friedmann world models has continued unabated with each new generation of ground-based and space observatories. The observations have brought both success and surprise. Supernovae, the best standard candles in deep space, have led to the revolutionary suggestion that dark energy rules the universe.

Does dark energy really exist or, will the pendulum swing again? The number of observed supernovae, used to detect the acceleration of the universe, will steeply rise from the present one hundred when supernova hunter telescopes start operating. For example, Supernova/Acceleration Probe (SNAP) could harvest 2000 supernovae in three years of operation! If new observations confirm its presence, dark energy with its antigravity will become the bread and butter of cosmological physics. Even the nearby space, close to the Local Group, is a natural laboratory in which all the cosmic ingredients, luminous and dark matter and dark energy may be studied.

Only a few hundred years ago, the world known to people was like a grain of sand compared to what we now know. The grandiose expansion of our astronomical knowledge, starting from our home planet, passing through the Solar System and the stellar fields of the Milky Way, and finally reaching the deep galaxy universe, has revealed a precious thing: on vastly different scales the same physical laws act. It is an incredibly long step from the red apple falling in your garden to the galaxy clusters gathering into megafractals. But it is this universal law at small and large, which is believed to be the builder of the large scale structures in the universe.

If megafractals are a result of gravitational growth, then three fields of cosmology should eventually find a concordance: space maps of luminous and dark matter, theoretical N-body simulations of the gravitational architecture, and studies of the patchiness of the cosmic background radiation. MAP and PLANCK missions will deliver unique information about the seeds which are thought to have grown into present day megafractals. Note, however, that our window into the past is not quite clean even for the millimeter radio waves. The tiny primordial ripples have to be extracted from amongst the feeble glow of cool cosmic dust, which is scattered along the line of sight all the way from deep space down to our home galaxy.

* * *

Astronomy and cosmology are going through major changes and what might even be called revolutions. The raw quantities of observational data are growing steeply: 3-D maps of galaxies are expanding to larger scales every year; the cosmic radiation is subject to intensive observational projects, and, occasionally, unanticipated discoveries – like that made with the aid of the supernovae – cause cracks in the familiar picture. This is a novel situation for a field which up to now has been characterized by many weird theoretical speculations and few observations to constrain them. And the flow of data – the exploration of the universe – brings fresh surprises, as it always has in the history of astronomy. Our story of cosmic megafractals has exposed one of these unexpected discoveries in which cosmological principles, the physics of gravitation, and the cosmic fractal structures appear together in the great design of our beautiful and enigmatic world.

Appendix A

A.1 Definition of the astronomical magnitude

In astronomy 'magnitude' is the measure of the light flux coming from a star or a galaxy. Curiously, the larger the magnitude, the fainter the star! The magnitude of the brightest star in the sky Sirius, $m = -1.5$, the faintest stars visible to the naked eye have $m = 6$, and the faintest objects detected by the Hubble Space Telescope have $m = 30$. This record faintness is like the glow of a cigarette on the Moon. In fact, magnitudes express the brightnesses of stars and galaxies in a logarithmic scale. If the flux ratio is f_1/f_2, then the magnitude difference is $m_1 - m_2 = -2.5 \log(f_1/f_2)$. Astronomers make flux measurements in different wavelength bands, which results in different useful magnitude systems.

A.2 The mass of the Milky Way

For a circular orbit around a mass M, Newton's 2nd law is $mV^2/R = GmM/R^2$ (acceleration caused by central force), or $M = RV^2/G = (R/1AU)(V/V_{earth})^2\ M_{\odot}$, using the Earth's orbit around the Sun as a comparison. With the Sun's distance from the Milky Way's center $R = 9000$ pc, the Milky Way's rotation velocity $V = 220$ km/sec, and the Earth's velocity $V_{earth} = 30$ km/sec, one calculates that the Milky Way's mass inside the orbit of the Sun is $M \approx 100 \times 10^9$ solar masses. Beyond the Sun's orbit the rotation speed remains roughly constant, hence this is a lower limit to the total mass. The (uncertain) total mass of the Milky Way, together with its dark halo, may reach 500 billion solar masses or even more.

A.3 A standard candle in the Hubble diagram

In the local universe the observed flux of light f from a galaxy follows the simple inverse square law $f \propto r^{-2}$ where r is the distance. Thus its magnitude $m = 5\log r + constant$. Because of the Hubble law $Z = H/c \times r$, identical galaxies lie in the Hubble diagram on a straight line $m = 5\log Z + constant'$. Such a relation is expected locally when the deviations from a static Euclidean space are minor.

More generally, standard candles in the Friedmann universe are predicted to have a magnitude–redshift relation $m = 5\log D_{lum}(z) + constant$, where D_{lum} is the so-called luminosity distance. In some cases it has a simple form: If 1) $\Omega_\Lambda = 0$, $\Omega_m = 1$, then the luminosity distance $D_{lum} = 2(1+z-(1+z)^{1/2})$; 2) if $\Omega_\Lambda = 0$, $\Omega_m = 0$, then $D_{lum} = z(1+z/2)$; 3) $\Omega_\Lambda = 1$, $\Omega_m = 0$, then $D_{lum} = z(1+z)$.

A.4 The classical electron and gravitational radiuses

The rest mass energy $E_{rest} = m_e c^2$ must be larger than the energy of the electric field around the electron $E_{field} = e^2/2R$ where R is the radius beyond which the energy is calculated (the energy density of the electric field is $\epsilon_{el} = (\vec{\nabla}\varphi_{el})^2/8\pi)$. These two energies are comparable for the radius $R_e = e^2/m_e c^2 = 2.8 \times 10^{-13}$ cm, the classical electron radius.

Similarly for any massive body one may define a classical gravitational radius based on the energy of the gravity field around the body $E_{field} = GM^2/2R$, which follows from the energy density $\epsilon_g = (\vec{\nabla}\varphi_N)^2/8\pi G$. The reasoning goes: $E_{field} < E_{rest} \implies R > R_{min} = GM/2c^2$

A.5 The cosmological gravitational redshift

The gravitational redshift within a fractal structure is calculated as $z_{grav} \approx \phi/c^2 = GM(R)/Rc^2 = 2\pi\frac{G}{c^2}\rho_0 r_0 R$, where $M(R)$ is the mass of the fractal structure within the scale R, and the last term is given for $D = 2$; ρ_0, r_0 are the radius and density at the zero-level of the hierarchy. Hence $z_{grav} \approx H_{grav}R/c$. Here $H_{grav} = 2\pi\frac{G}{c}\rho_0 r_0$ is the "gravitational Hubble constant". If we take a typical galaxy for the lowest building brick of the hierarchy, then $2\pi\rho_0 r_0 \approx 1$ g/cm^2, and $H_{grav} \approx 70$ km sec^{-1}/Mpc. But this would require much more dark matter locally than is observed (Ch.18).

Suggestions for Reading

For further reading and general background we recommend:

General astronomy and history

H. Karttunen et al. (eds.): Fundamental Astronomy (Springer 1994)
M. Hoskin: Stellar Astronomy (Science History Publications 1982)
T. Kuhn: The Copernican Revolution (Harvard University Press 1957)

Fractals

B. Mandelbrot: The Fractal Geometry of Nature (W.H. Freeman 1983)
H.-O. Peitgen & P. Richter: The Beauty of Fractals (Springer 1986)
M. Schroeder: Fractals, Chaos, Power Laws (W.H. Freeman 1991)

Large scale structure and cosmology

A. Fairall: Large-scale structure in the universe (John Wiley & Sons 1998)
E.R. Harrison: Darkness at Night (Harvard University Press 1987)
E.R. Harrison: Cosmology – the science of the universe, 2nd edition (Cambridge University Press 2000)
A.R. Sandage, R.G. Kron & M.S. Longair: The Deep Universe (Springer 1995)
W. Saslaw: The Distribution of the Galaxies – Gravitational Clustering in Cosmology (Cambridge University Press 2000)

Gravitation and general physics

P. Davies (ed.): The New Physics (Cambridge University Press 1989)
V.J. Ostdiek & D.J. Bord: Inquiry into Physics (West Publishing Co. 1995)
R. Feynman: Feynman Lectures on Gravitation (Perseus Books 1995)

Index