More praise for *Cosmology's Century*

"A century of big ideas and powerful instruments has led us to the current model of our universe, with its inflationary beginning, cosmic structure built by the gravity of dark-matter particles, and accelerated expansion caused by dark energy. *Cosmology's Century* is a firsthand account of that remarkable period by Jim Peebles, who led this grand adventure with his manifold contributions and broad influence. A must-read for any serious student of cosmology."

—Michael S. Turner, Kavli Foundation and University of Chicago

"Jim Peebles has surely contributed more to the history of our understanding of the large-scale structure and evolution of the universe than anyone else still in a position to write about it; so written about it he has, and magnificently!"

—Virginia Trimble, Former President, International Astronomical Union, Division of Galaxies and the Universe

"An inspiring history of cosmic ideas."

—Joseph Silk, author of *The Infinite Cosmos: Questions from the Frontiers of Cosmology*

"Peebles offers a broad and deep description of cosmology, presenting the history of the field as well as many of the side turns, dead ends, and wrong paths that researchers explored along the way. I really enjoyed reading this book."

—David W. Hogg, New York University

Published by Princeton University Press
41 William Street, Princeton, New Jersey 08540
6 Oxford Street, Woodstock, Oxfordshire OX20 1TR

press.princeton.edu

ISBN 978-0-691-19602-2
ISBN (e-book) 978-0-691-20166-5

British Library Cataloging-in-Publication Data is available

Editorial: Jessica Yao and Arthur Werneck
Production Editorial: Brigitte Pelner
Jacket/Cover Design: Chris Ferrante
Production: Jacqueline Poirier
Publicity: Matthew Taylor (US), Katie Lewis (UK)
Copyeditor: Cyd Westmoreland

Jacket Image: Planck's Cosmic Microwave Background (CMB) Map, 2013, ESA and the Planck Collaboration

This book has been composed in Miller

Printed on acid-free paper ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

about it. And with each advance in science, we see an addition to the evidence that there is an objective physical reality and that we are probing ever more deeply into its nature. There is nothing new in all this, but I think the examples to be drawn from the history of modern cosmology are particularly clear and informative, because the subject is relatively simple.

I have already presented portions of this story. What happened in George Gamow's research group in 1948 is considered in detail in Peebles (2014). The work in Bob Dicke's Gravity Research Group that was so important to the development of experimental gravity physics—and led to the recognition of the sea of thermal radiation remnant from the hot early universe—is reviewed in Peebles (2017). Recollections of research in the 1960s by those who were involved in the identification and interpretation of this fossil thermal radiation are in the book *Finding the Big Bang* by Peebles, Page, and Partridge (2009).

References to the research papers I consider important to the story are indicated by authors' names followed by the year and are listed in the References section at the end of the book. The list is dismayingly long, but it has to be: although this has been a relatively small science, its development took a lot of work. I have selected samples of the pioneering contributions and apologize to colleagues who have different opinions about this subjective matter. Page numbers following references indicate where the papers are cited in the text.

Some of the quotations in this book are taken from the literature, and the sources are so indicated. Where the quote is in French or German, I add my translation, sometimes condensed and aided by Google. This book offered an excellent opportunity to ask for recollections from those who have long memories of research in this subject. Quotes drawn from them for the purpose of this book are marked by the author's name and "personal communication." I have profited also from the advice of younger people, and from the wonders of the Internet. I am particularly thankful for NASA's Astrophysics Data System Bibliographic Services archive, a most useful tool for tracking down research papers from times past.

Figures that illustrate data can be influential, and the evolving nature of these figures is a part of the history. I am grateful to colleagues who gave me figures they made and own; their names are mentioned in the captions. The figures I made for this book, or made in times past but never published, have no references in the captions. Captions state sources of the many figures that have been taken from the literature, and the copyright holder can be traced through the reference to the publication. Copyright holders have a broad variety of prescriptions for statements of permission to reproduce, and their conditions for permission range from casual statements that reuse of figures is OK to payments required to reproduce two of the figures in this book taken from the publication of an otherwise respectable scholarly society. I take this confusion of permissions to be a consequence of the natural desire of publishers to keep some control over their content while the ease of taking figures from the

literature for use in lectures can readily spill over into publications. I apologize for any permissions to reproduce I may have improperly stated or overlooked, and if notified will make amends in later printings.

The color plates that appear within chapter 9 are a sample of the actors in this history; I mean them to be reminders of the people behind those equations and measurements. I apologize to valued colleagues whose photos could have been included if space in this book and the energy to collect them had been more freely available. The text accompanying the photographs is my opportunity to comment on the stories behind the images, the analog in print of teachable moments.

My choice of units follows customs that tend to differ in different lines of research. In some parts of cosmology, the units usually are chosen so the velocity of light is unity. These equations look odd to me when the symbol c is entered, a matter of conditioning of course, but I follow tradition, which seems appropriate, since this is a history. In other places, Planck's constant $\hbar$ is unity, or Newton's constant G is unity. The old centimeter, gram, second units are being replaced by meters, kilograms, seconds. I suppose this is a sensible move, but the change is slow, and again I follow the history in staying with the former.

The index lists only a few of the pioneers of cosmology. This is a subjective choice, as is whatever else is deemed appropriate for an index. It would make no sense to place in the index the many appearances in the text of the word "redshift," so I index only the definition that appears early in the text. The word "inflation" appears a lot, too, and I enter the first significant commentary about the concept and later page numbers in which cosmological inflation is particularly relevant. But such algorithms are only of limited help with so many decisions.

This account may seem overly centered on the small town of Princeton in the small state of New Jersey. That is inevitable, in part because I have been a member of Princeton University since arriving here as a graduate student in 1958, but inevitable in even larger part because a good deal of the story happened here. My role in this story was aided by sabbatical leaves at the California Institute of Technology; the University of California, Berkeley; the Dominion Astrophysical Observatory in British Columbia; the University of Cambridge; and on two occasions, the Institute for Advanced Study in Princeton. I learned a lot at these places.

I have benefited from advice from and recollections of many colleagues: Neta Bahcall, John Barrow, Dick Bond, Steve Boughn, Michele Cappellari, Claude Carignan, Ray Carlberg, Rick Carlson, Robin Ciardullo, Don Clayton, Shaun Cole, Ramanath Cowsik, Marc Davis, Richard Dawid, Jaco de Swart, Jo Dunkley, John Ellis, Wyn Evans, Sandra Faber, Kent Ford, Ken Freeman, Carlos Frenk, Masataka Fukugita, Jim Gunn, David Hogg, Piet Hut, David

Kaiser, Steve Kent, Bob Kirshner, Al Kogut, Rocky Kolb, Andrey Kravtsov, Rich Kron, Malcolm Longair, Gary Mamon, John Mather, Adrian Melott, Liliane Moens, Richard Mushotzky, Kieth Olive, Jerry Ostriker, Lyman Page, Bruce Partridge, Will Percival, Saul Perlmutter, Mark Phillips, Joel Primack, Martin Rees, Adam Riess, Brian Schmidt, Jerry Sellwood, Joe Silk, David Spergel, Ed Spiegel, Paul Steinhardt, Matthais Steinmetz, Michael Strauss, Alex Szalay, Alar Toomre, Rien van de Weygaert, Hugo van Woerden, Steve Weinberg, Rainer Weiss, Cyd Westmoreland, Simon White, Ned Wright, Jessica Yao, and Matias Zaldarriaga. I surely have forgotten to mention some; my sincere apologies.

COSMOLOGY'S CENTURY

1.1 *The Science and Philosophy of Cosmology*

The starting assumption for cosmology, as in all branches of natural science, is that nature operates by kinds of logic and rules that we can discover by careful examination of what is observed, informed by past experience of what has worked. The results are impressive; I urge any who might disagree to consider the rich fundamental physics employed in the construction and operation of their cellphones. But despite the many demonstrations of its power, physics, along with all the rest of natural science, is incomplete. Maybe discoveries to come will make the physical basis for science complete, revealing the final rules by which nature operates. Or maybe it's successive approximations all the way down.

The standard and accepted methods of science must be adapted to what can be done, of course. In physical cosmology and extragalactic astronomy, we can look but never touch. In cosmology, we cannot run the experiment again; we must instead resort to what can be inferred from fossils of times past. We find some fossils relatively nearby, as in the rocks on Earth and the stars in our galaxy and others, all of which have their own creation stories. Our past light cone offers us views of times past, because radiation detected here has been approaching us at the speed of light: the greater the distance of an object, the earlier in the evolution of the universe it is observed. Our light cone integrated through human history captures an exceedingly thin slice of what has been happening, but it reveals the way things were over a long range of time in a large universe that offers a lot to see and to seek to interpret.

The research path to where we are now in cosmology is marked by debates on open questions, as is usual in natural science. But the issues in cosmology have been defended and criticized with considerably more vigor than might have been expected from the modest weight of the evidence at the time. This was in part because observations that might settle questions in cosmology have tended to seem just out of reach or perhaps just barely possible. And I think an important factor has been the tendency to take a personal interest in the nature of our world. Is the universe really evolving, or might it be in a steady state? If evolving, how might it all end, in a big crunch or a big freeze? And where did it all come from? Such debates are quieter now, because we at last have a theory that passes an abundance of tests, but they continue.

Research in cosmology in the twentieth century usually was done in small groups, often an individual working alone or maybe with a colleague or a student or two. In the twenty-first century, ongoing research in cosmology grew richer and called for larger groups to develop special-purpose equipment for data acquisition, which in turn called for groups of comparable size to reduce the data and interpret it. Big Science has become important to this subject: We have to get used to gathering data in vast amounts, analyzing these data, and employing massive numerical simulations that help bridge the gap between

theory and observation. But Big Science best takes aim at well-motivated and sharply defined questions. The main considerations in this book are about how small groups working on seemingly independent lines of research found their results coming together in a cosmology that looked good enough to call for the demanding tests afforded by Big Science. I date this revolutionary convergence to a credible theory to the half decade from 1998 to 2003.

Research certainly continued to be active and productive after the revolution; the difference is that the community had agreed on a paradigm, in Kuhn's (1962) terms. (This is what the majority was thinking, of course; not all agreed.) An example of the adherence to the normal science of cosmology is the study of how the galaxies formed and evolved, which builds theories of galaxy formation on the standard and accepted theory of the evolution of the universe. Normal scientific research of this sort may uncover anomalies that point to a still better underlying theory. This is a point of particular interest in cosmology, because the theory is at the same time well and persuasively tested and particularly incomplete.

Our present normal science of cosmology includes an excellent case for the presence of dark matter that interacts weakly if at all with ordinary matter. There are tight constraints on the properties of dark matter, but no clear evidence exists of detection of this substance other than the inference from the effects of its gravitational attraction. Some argue that dark matter will remain only hypothetical until there is more evidence of it than that: maybe detection in the laboratory, maybe indications of what it is doing to galaxies apart from holding them together. Others argue that the case for dark matter already is so tight that it is abundantly clear that the dark matter really exists. The same applies to Einstein's cosmological constant, Λ. It has gained a new name: dark energy. But that is a poor disguise for a fudge factor that we accept because it serves to unify theory and observations so well. There are other fudge factors, hypotheses to allow the theory to save the phenomena, in the present standard science of cosmology and in all the other branches of natural science. Research in the sciences continues to improve tests of our theories that, whether intended or not, may lead to better theories that inspire new tests. And they might on occasion replace fudge factors with unified theories in paradigms that bring parts of this enterprise closer together. It happens.

The physical cosmology that is the subject of this history is an empirical science, that is, it is based on and tested by what can be observed or measured by detectors, such as microscopes and telescopes and people. But we must pay attention to the role of theory, and intuition, and what Richard Dawid (2013 and 2017) terms "nonempirical theory assessment." The prime example in this history is that during most of the past century of research in cosmology, the community majority implicitly accepted Einstein's general theory of relativity. Few pointed out that this is an enormous extrapolation from the few meager tests of general relativity that we had in the 1960s. By the 1990s, as

research in cosmology was starting to converge on a well-tested theory, there were demanding checks of the predictions of general relativity on scales ranging from the laboratory to the solar system, probing out to length scales of about 10^{13} cm. But the application to cosmology on the scale of the Hubble length, about 10^{28} cm, extrapolates from the precision tests by some fifteen orders of magnitude in length scale. This was not often mentioned, in my experience, and when mentioned, it tended to make some scientists a little uneasy, at least temporarily. In the first decades of the twenty-first century, the parts of general relativity that are relevant to the standard cosmology have passed an abundance of demanding tests. In short, the theory Einstein built on laboratory experiments was seriously tested only by the orbit of the planet Mercury. (The test of the prediction of the gravitational deflection of light by the mass of the sun, led by the people pictured in Plate III, was heavily cried up but in retrospect, their evidence seems marginal.) We find that this theory successfully extrapolates to applications on the immense scales of the observable universe. It is a remarkable result.

General relativity is an elegant extension of electromagnetism in flat spacetime; it has been said that it is a theory waiting to be found (though that is easier to say in hindsight). The faith in its extrapolation exemplifies the powerful influence and very real successes of nonempirical theory assessment. Of course, influential nonempirical assessments can mislead: Consider that in the 1930s through the 1990s, few objected to the assertions by respected experts that Einstein's cosmological constant, Λ, surely may be discarded. The evidence now is that Λ, under its new name—dark energy—is an essential part of our well-tested cosmology.

The practice of nonempirical assessments is sometimes termed "post-empiricism," but I have not found this term in Dawid's writing. Dawid (in a personal communication, 2018) states instead that

> non-empirical assessment as I understand it crucially depends on the ongoing collection of empirical data elsewhere in the research field and on the continued search for empirical confirmation of the theory under scrutiny. In a "post-empirical" phase where no substantially new data comes in any more, non-empirical assessment would get increasingly questionable and eventually would come to a halt as well.

This is consistent with what I understand to be normal practice in the physical sciences. That is, I have in mind the kind of nonempirical assessments we have been practicing all along without thinking much about it.

I take account of three other kinds of assessments: personal; community, though some may disagree; and pragmatic. The first two speak for themselves. I take examples of the third from cosmology. The usual practice has been to analyze data and observations in terms of general relativity. This surely has been due in part to the beauty of the theory, and in part to respect for Albert Einstein's magnificent intuition. But it was important also that the use of a

common theory allowed comparisons of conclusions from independent analyses of the same or different data on a common fundamental ground. I do not imagine much thought has been given to this point, but I believe the implicitly pragmatic approach in cosmology (and I suppose in other branches of natural science) has helped reduce the chaos of multiple theories.

The pragmatic approach to science, if carried too far, could waste time and resources by directing research along a path as it grows increasingly clear that something is wrong. And even if the popular and pragmatically chosen path proves to be leading us in a useful direction, it can be important to have well-defended alternatives to standard ideas to motivate careful evaluations of approved ideas and observations. It may reveal corrections large or small that point toward a more profitable path. For example, a stimulating proposal in the mid-twentieth century was that textbook physics may have to be adjusted to include continual spontaneous creation of matter. The brave souls who argued for this steady-state cosmology were not always gently treated, but from what I saw, they gave as good as they got in debates over the relative merits of the general relativity and steady-state world views, arguments that were more intense than warranted by the evidence for or against either side. The idea of continual creation in the universe as it is now is no longer seriously considered in cosmology, but it had a healthy effect. New ideas can inspire defense and attacks that stimulate research, while a pragmatic defense of the old ways may help keep research from degenerating into confusion.

An important example of an implicitly pragmatic assessment is the general acceptance of Einstein's proposal that the universe is homogeneous in the average over local irregularities. Prior to the 1960s, there was scant evidence of this. Maps of distributions of the galaxies across the sky suggested instead that the galaxies are moving away from one another into space that is asymptotically empty or close to it, as in a fractal galaxy distribution. But whether by accident or design, this quite pertinent thought was put aside for the most part, and the main debate kept more sharply focused on the concepts of evolution or else a steady state of a nearly homogeneous universe. The first serious evidence for homogeneity came a half century after Einstein, from research for other purposes in the 1960s, as will be discussed in Chapter 2. Whether by good luck or good taste, the community was not much distracted by the elegant but wrong idea of a fractal universe.

It is not always easy to see why some issues receive much more attention than others; I suppose such things are to be considered eventualities. We do have reasonably clear standards for rejecting an apparently interesting idea. For example, the steady-state cosmology introduced in 1948 is elegant, but its predictions clearly violate the later accumulation of empirical tests. I do not know of a clear prescription for a move in the other direction, namely, the promotion of a working model to a standard theory. We might use the term "community opinion" to describe such decisions.

In 1990, general relativity usually was taken to be the appropriate basis for the study of the large-scale nature of the universe, but as argued above, it was an implicitly pragmatic assessment that the theory was serving well as a working basis for research. In 2003, after the revolution, the cosmological tests gave weight to the community opinion that the universe actually is well described by general relativity applied to the set of assumptions in what became known as the ΛCDM cosmological model. The introduction of these assumptions, including Einstein's cosmological constant Λ and the hypothetical cold dark matter, is reviewed in Section 8.2. Some disagreed, to be sure, but to most the accumulation of evidence (reviewed in Chapter 9) had become tight enough to have emboldened talk of what "really happened" far away and in the remote past, based on the ΛCDM theory. The notion of reality is complicated, so a more secure statement would be that whatever happened—and we assume something did happen—left traces that closely resemble those predicted by ΛCDM. And the traces are abundant and well enough cross-checked that the community opinion, including mine, is that this theory almost certainly is a useful though incomplete approximation to what actually happened.

1.2 An Overview

I have sorted this history of cosmology into lines of research that operated more or less independently of one another through stretches of time in the twentieth century. I consider the developments in each of the lines of research roughly in chronological order, but because different lines of research were at best only loosely coordinated, there have to be references back and forth in time as different lines of research started to interact. This outline is meant to explain how I have arranged the presentation of the research and how it all fits together, at least roughly, apart from the wrong turns taken.

I begin in Chapter 2 with considerations of Albert Einstein's (1917) proposal, from pure thought, that a philosophically sensible universe is homogeneous and isotropic: no preferred center or direction, no observable edges to the universe as we see it around us. That of course is apart from the minor irregularities of matter concentrated in people and planets and stars. Einstein's homogeneity is essential to the thought that we might be able to find a theory of the universe as a whole rather than of one or another of its parts. It was an inspired intuitive vision or maybe just a lucky guess; Einstein certainly had no observational evidence that suggested it. The history of how Einstein's thought was received and tested exemplifies the interplay in science between theory and practice, sometimes reinforcing each other; sometimes in serious tension; and, as in this case, sometimes aided by unexpected developments. Because I have not found a full discussion elsewhere, I consider in some detail the development of the evidence that supports what became known as Einstein's cosmological principle.

The subject of Chapter 6 is the astronomers' discoveries of apparent anomalies in the measurements of masses of galaxies and concentrations of galaxies. Other accounts of the exploration of these phenomena are in Courteau *et al.* (2014) and de Swart, Bertone, and van Dongen (2017). Fritz Zwicky was the first to recognize the phenomenon: He saw that the galaxies in the rich Coma Cluster of galaxies seem to be moving relative to one another too rapidly to be held together by the gravitational attraction of the mass seen in the stars in the galaxies in the cluster. One way to put it is that the mass required to hold this concentration of galaxies together by gravity seemed to be missing, always assuming the gravitational inverse square law of gravity (in the nonrelativistic Newtonian limit of general relativity). It was later seen that mass also seemed to be missing from the outer parts of spiral galaxies, based on the measurements discussed in Section 6.3 of circular motions of stars and gas in the discs of spiral galaxies. Much the same conclusion came from the studies described in Section 6.4 of how galaxies with prominent discs acquired their elegant spiral patterns. By the mid-1970s, it had become clear that understanding this is much easier if the seen mass is gravitationally held in near-circular motion in the disc with the help of the gravitational attraction of less-luminous matter that is more securely stabilized by more nearly random orientations of the orbits.

These observations pointed to a key idea for the establishment of cosmology: the existence of "dark matter," the new name for what was variously known as "missing," "hidden," or "subluminal" mass. The idea came almost entirely out of pursuits in astronomy, not cosmology, and for this purpose, the subluminal component need not be very exotic: low-mass stars would do, though they would have to be present in surprising abundance relative to counts of the more luminous observed stars. But in the 1970s, another key idea for cosmology was growing out of particle physicists' growing interest in the possible forms of nonbaryonic matter. Gas and plasma, people, planets, and normal stars are all forms of what is termed "baryonic matter." Most of the mass of baryonic matter is in atomic nuclei; the accompanying electrons are termed "leptons," but they are also counted in the mass of baryonic matter. The neutrinos are leptons that we now know have small but nonzero rest masses. Thus they act as nonbaryonic dark matter that contributes to the masses of galaxies, but in the standard cosmology, this contribution is much smaller than the total indicated by the astronomical evidence. We need a new kind of nonbaryonic matter.

The thought that the astronomers' subluminal matter is the particle physicists' nonbaryonic matter and the cosmologists' dark matter was and remains a conjecture at the time of writing. The only empirical evidence of the new nonbaryonic dark matter is the effect of its gravity. It has been a productive idea, however, that passes demanding checks. The particle physicists' considerations of nonbaryonic matter reviewed in Chapter 7 takes into account

the condition that if this nonbaryonic matter were produced in the hot early stages of expansion of the universe, then its remnant mass density must not exceed that allowed by the relativistic big bang cosmological model (again, assuming the relativistic theory). But it is notable that cosmologists took over the notion of nonbaryonic dark matter before the particle physics community had taken much interest in the astronomers' evidence of the presence of subluminal matter.

The nonbaryonic dark matter most broadly discussed in the 1980s came in two varieties, cold and hot. The latter would be one of the known class of neutrinos with rest mass of a few tens of electron volts (Sections 5.2.7 and 7.1). The initially hot (meaning rapidly streaming) neutrinos in the early universe would have smoothed the mass distribution, and that smoothing would have tended to cause the first generation of structure to be massive systems that must have fragmented to form galaxies. The spurious indication in 1980 of a laboratory detection of a neutrino mass appropriate for the hot dark matter picture certainly enhanced interest in the indicated formation of galaxies by fragmentation. This model was considered but had to be rejected: the observations show hierarchical growth of structure, from smaller to larger mass distributions.

The prototype for the nonbaryonic matter that is an essential component of the established cosmology was introduced by particle physicists in 1977. The idea occurred to five groups who published in the space of 2 months. These papers do not exhibit much interest in the astronomers' subluminal mass phenomena, but the considerations certainly were relevant to subluminal matter. Was this a curious coincidence or an idea that somehow was "in the air?" This is considered a little further in Sections 7.2.1 and 10.4.

Sections 8.1 and 8.2 review why in the early 1980s cosmologists co-opted the astronomers' subluminal mass and the particle physicists' nonbaryonic matter in what became known as the standard cold dark matter, or sCDM, cosmological model. The letter "s" might be taken to mean that the model was designed to be simple (as it was) but it instead signified "standard," not because it was established but because it came first. It was meant to distinguish this version from the many variants to be considered in Section 8.4. A large part of the cosmology community soon adopted variants of the sCDM model as bases for exploration of how galaxies might have formed in the observed patterns of their space distribution and motions (Section 8.3), and for analyses of the effect of galaxy formation on the angular distribution of the sea of thermal radiation. This widespread adoption was arguably overenthusiastic, because it was easy to devise other models, less simple to be sure, that fit what we knew at the time. And it was complicated by the nonempirical feeling that space sections surely are flat. In general relativity that could be because the mass density is large enough to produce flat space sections, or because Einstein's cosmological constant, Λ, makes it so. The nonempirical reasons for

preferring flat space sections, preferably without resorting to Λ, are discussed in Section 3.5. These reasons were influential and long-lasting enough to have played a significant role in the confusion of variants and alternatives to the sCDM idea considered in the 1990s.

The reduction of confusion in the years 1998–2003 was great enough to be termed a revolution. It was driven by the two great experimental advances discussed in Chapter 9. The first is the measurement of the relation between the redshift of the spectrum of an object and its brightness in the sky, given its luminosity: the cosmological redshift–magnitude relation. Its detection had been a goal for cosmology since the 1930s; it was at last accomplished by two independent groups at the turn of the century (Section 9.1). The second is the detailed mapping of the angular distribution of the CMB radiation. Work on this began in the mid-1960s, and coincidently also produced demanding constraints on cosmological models at the turn of the century. These results from the two sets of measurements, together with what was already known, made a tight case for the presence of Einstein's cosmological constant Λ and the nonbaryonic CDM in the relativistic hot big bang ΛCDM theory. It was a dramatic development.

It was proper to have asked whether the introduction of two very significant hypothetical components, CDM and Λ, along with all the other assumptions that go into the choice of a cosmological model, might only amount to adjusting the theory to fit the measurements. That line of debate did not become very prominent, because the ΛCDM cosmology that fit the two critical measurements brought together so many other lines of evidence in a tight network of empirical tests. This is the topic of Section 9.3.

By the year 2003, the community had at last settled on a respectably well-supported theory of the large-scale nature of the universe. Skeptics remained, as is appropriate, for this theory is an immense extension of the reach of established physics. Indeed, the 2003 theory has been modified to fit later measurements, but these changes amount to fine adjustments of parameters, not challenges to the basic framework of the theory. It is the nature of science to advance by successive approximations, and it would not be at all surprising to find that there is a still better theory than ΛCDM. But we have excellent reason to expect that a better theory will describe a universe that behaves much like ΛCDM, because ΛCDM passes an abundance of empirical tests that probe the universe in so many different ways.

I cannot think of any lesson to be drawn from this story of how cosmology has extended the boundaries of established science that cannot be drawn from other branches of natural science. This is no surprise, because cosmology operates by the methods of natural science. But I think there are lessons to be drawn with greater clarity in the relatively uncluttered historical development of this subject. My offerings are given in Chapter 10.

CHAPTER TWO

The Homogeneous Universe

MODERN COSMOLOGY GREW out of Albert Einstein's search for how his general theory of relativity might apply to the large-scale nature of the universe. Einstein's (1917) thought was that a philosophically reasonable universe is the same everywhere and in all directions, apart from minor irregularities, such as the observed concentrations of matter in planets and stars. This is a distinct departure from the tradition of research in natural science, which is to select for examination a level in a hierarchy of structure. It may be the examination of molecules; the atoms in molecules; the nuclei in atoms; the nucleons in nuclei; or the quarks and gluons in nucleons. One can examine structure on larger scales: the vast complexity of interactions of atoms and molecules in condensed matter, chemistry, and on up to biophysics; or the natures of planets around stars, stars in galaxies, or galaxies in groups and clusters and superclusters of galaxies. Einstein's thought was that this hierarchy of structures ends in something new to modern science: large-scale homogeneity. (Although not stated explicitly at first, the thought includes large-scale isotropy. That is, the universe is assumed to be invariant under rotations as well as translations.)

Einstein's homogeneity assumption allows us to consider and test the possibility of a theory of the universe as a whole, rather than a theory of a particular level in a hierarchy. If the universe is homogeneous in the large-scale average, then observations from our position may inform the theory of what the universe is like when observed from any other place. But we need evidence that this approximation is useful.

2.1 Einstein's Cosmological Principle

Einstein's (1917) original argument for the picture of large-scale homogeneity is difficult to assess. He argued against the idea that the material content of the universe might be confined to a single concentration, an island universe in otherwise empty space. If this were so, and the escape velocity were finite, then

stars would evaporate, escaping the island universe. This behavior would be contrary to his implicit assumption that the universe is in a stationary state. If the escape velocity were arbitrarily large, then statistical relaxation would produce the occasional star moving with arbitrarily large speed. This might be taken to be contrary to the observation that the velocities of nearby stars are much smaller than the velocity of light. Both points would make some sense if the universe were not evolving and the stars had had time to approach statistical equilibrium. Einstein does not seem to have paused to consider that if energy is conserved, then the stars must eventually stop shining. And if stars nevertheless shine forever, then his homogeneous universe would be full of starlight. This is Olbers' paradox, and is certainly an unacceptable situation.

The argument that may be closer to what Einstein was thinking in 1917 is stated in *The Meaning of Relativity*, the publication of his lectures at Princeton University in 1921 (Einstein 1923). He pointed out that his general relativity allows solutions in which there is a single mass concentration outside of which spacetime is empty and asymptotically flat, or as Einstein put it, quasi-Euclidean. Motions of matter in this mass concentration would have the usual properties of acceleration, such as the flattening of a gravitationally bound rotating galaxy. But in a nonrelativistic mass concentration, this rotation would be relative to empty spacetime. Thus Einstein (1923, 109) wrote: "If the universe were quasi-Euclidean, then Mach was wholly wrong in his thought that inertia, as well as gravitation, depends upon a kind of mutual action between bodies."

A similar sentiment, expressed in Einstein (1917), is that (in an English translation): "In a consistent theory of relativity there can be no inertia *relative to "space,"* but only an inertia of masses *relative to one another.*" He went on to point out that in his general relativity, a single particle of mass in otherwise flat spacetime would have inertia, contrary to his stated view of relativity.

Einstein (1923, 110) argued that it is "probable that Mach was on the right road" in the relativity of inertia, and cited three examples:

1. The inertia of a body must increase when ponderable masses are piled up in its neighborhood.
2. A body must experience an accelerating force when neighboring masses are accelerated, and, in fact, the force must be in the same direction as the acceleration.
3. A rotating hollow body must generate inside of itself a "Coriolis field," which deflects moving bodies in the sense of the rotation, and a radial centrifugal field as well.

With all respect to Einstein's genius, we must observe that the first example, if meant as a local measurement, may follow from Mach's principle, but it is not true in general relativity. This theory predicts that an observer confined to a

factor. With $l \propto a(t)$, the rate of change of the physical distance $l(t)$ between any two galaxies at separation l is

$$\frac{dl}{dt} = v = \frac{\dot{a}}{a}\ell(t). \tag{2.3}$$

The dot means time derivative. We see that Hubble's constant in equation (2.1) is

$$H_0 = \frac{1}{a}\frac{da}{dt}, \tag{2.4}$$

evaluated at the present epoch, at expansion time $t = t_0$.

The departure of a galaxy velocity from the mean value set by Hubble's law at that position is said to be the galaxy peculiar velocity. Peculiar velocities usually may be attributed to the gravitational pull of the growing clustering of mass in galaxies and concentrations of galaxies, but nongravitational forces produced by explosions may be important, too.

At nonrelativistic recession speeds, the cosmological redshift is defined as $z = v/c$, where c is the speed of light. This is a first-order Doppler shift. The distance at which Hubble's relation between distance and recession velocity extrapolates to the speed of light, $r_{\rm H} = cH_0^{-1} \sim 10^{28}$ cm, is the Hubble length. Consideration of the relativistic correction to equation (2.3) for galaxies at this great distance begins in Section 3.2.

2.2 *Early Evidence of Inhomogeneity*

In the 1930s, the cosmological principle passed an important empirical check: The prediction from homogeneity of the redshift-distance relation in equation (2.1) was shown to fit the tight tests discussed in Section 2.3. But homogeneity was not suggested by maps of the galaxy distribution. Charlier (1922) presented a map of the distribution across the sky of the known nebulae. Among the objects in Charlier's map are clusters of stars in our galaxy, and regions where starlight is reflected by clouds of dust, but most are extragalactic nebulae, that is, other galaxies of stars. Charlier pointed out that the map brings to mind hierarchical clustering: galaxies appear in clumps that are present in clumps of clumps, and so on, perhaps to indefinitely large scales. This was later named a "fractal universe."

A decade later, Harlow Shapley and Adelaide Ames at the Harvard College Observatory presented a catalog of the 1,249 known galaxies brighter than $m = 13$ (a measure of the brightness in the sky). Their maps of the angular positions in the two hemispheres of our galaxy are shown in Figure 2.2 (Shapley and Ames 1932). The left-hand panel shows the galaxies in the North hemisphere of our galaxy, the right-hand panel those in the South galactic hemisphere. The near absence of galaxies near the plane of our Milky Way galaxy is due to absorption of light by interstellar dust lying close to the plane

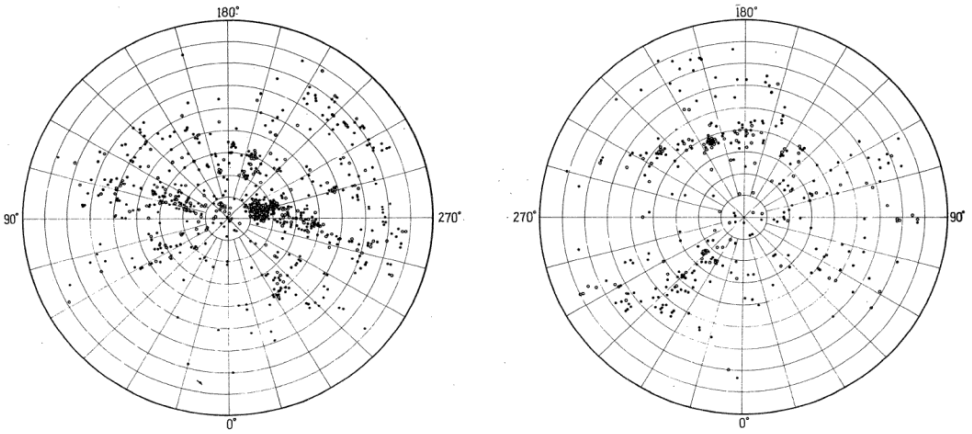

FIGURE 2.2. The Shapley and Ames (1932) map of galaxies brighter than apparent magnitude 13. Courtesy of the John G. Wolbach Library, Harvard College Library.

of our galaxy. The sky is clearer above and below the plane. The northern hemisphere on the left in the figure shows the many galaxies in the prominent concentration in and around the Virgo Cluster of galaxies. (The cluster is named for its position in the sky, near the stellar constellation Virgo.) De Vaucouleurs (1953 and 1958a) named the Virgo Cluster and the broad concentration of galaxies around it the Local Supercluster. This distinctly inhomogeneous distribution of the nearby galaxies is well established.

Willem de Sitter (1917a,b) presented discussions of Einstein's thoughts about the structure of the universe. Since de Sitter was a knowledgeable astronomer, he could have told Einstein about the nebulae, the thought that most are extragalactic, and the evidence that these extragalactic nebulae are not at all close to uniformly distributed. But I have not seen any indication that Einstein considered this observation and if so, whether it affected his thinking.

The possibilities in 1917 were that obscuration by dust is quite patchy even well away from the plane of our galaxy, or else that the observed distribution of galaxies does not at all resemble the homogeneity of the cosmological principle. Not much had changed by the 1950s except that the dust option was ruled out. The situation was recognized in the influential and informative book, *The Classical Theory of Fields* (Landau and Lifshitz 1951, the English translation of the 1948 Russian edition). It presents an admirable exposition of the special and general theories of relativity, but there is little mention of data in this book or in the others in their series on theoretical physics. A rare exception is the comment about Einstein's homogeneity assumption in Landau and Lifshitz (1951, 332):

> Although the astronomical data available at the present time give a basis for the assumption of uniformity of this density, this assumption

> can of necessity have only an approximate character, and it remains an open question whether this situation will not be changed even qualitatively as new data are obtained, and to what extent even the fundamental properties of the solutions of the equations of gravitation thus obtained agree with actuality.

As we see from Figure 2.2, this was a sensible remark, though from an empirical point of view, one might have expected another caution about the scant tests of general relativity. The situation in gravity physics was quite different from the empirical situation in the first part of their book, on the very well tested and broadly applied theory of electromagnetism.

In a report to the eleventh Solvay conference, *La structure et l'évolution de l'univers,* Oort (1958) began with the statement that "One of the most striking aspects of the universe is its inhomogeneity." As evidence, he showed the Shapley and Ames (1932) map in Figure 2.2. He could have added that Abell's (1958) catalog of the more-distant rich clusters of galaxies shows them scattered across the sky in a clumpy fashion, as in superclusters of clusters. But the distribution of clusters in Abell's map (1958, Figure 7) does look distinctly less clumpy than the distribution of the much closer galaxies in the Shapley-Ames map.

2.3 Early Evidence of Homogeneity: Isotropy

There remained the possibility that the galaxies are uniformly distributed in the average over larger volumes than Shapley and Ames had sampled. Hubble (1926 and 1934) introduced a test, the variation of the counts of faint galaxies as a function of position across the sky. Away from the areas obscured by interstellar dust close to the plane of the Milky Way, Hubble (1934) typically found about 100 galaxies per square degree (reduced to standard observing conditions) to a limiting redshift he estimated to be about $z = 0.1$. This is deep, 10 percent of the speed of light, and is about ten times the distance sampled in the Shapley-Ames map. Hubble's counts at low galactic latitudes, plotted as the lower strings of data in Figure 2.3, are smaller than at high latitudes and show a systematic variation across the sky. Both are effects of obscuration by dust in variable amounts along lines of sight near the directions of the plane of the Milky Way. The upper strings of data are counts at 40–50 degrees above the plane, plotted as filled circles in the north galactic hemisphere and open circles in the south. The counts are similar in the two hemispheres and do not show a systematic tendency to vary with position across the sky. Hubble (1934, 62) concluded that

> On the grand scale, however, the tendency to cluster averages out. The counts with large reflectors conform rather closely with the theory of

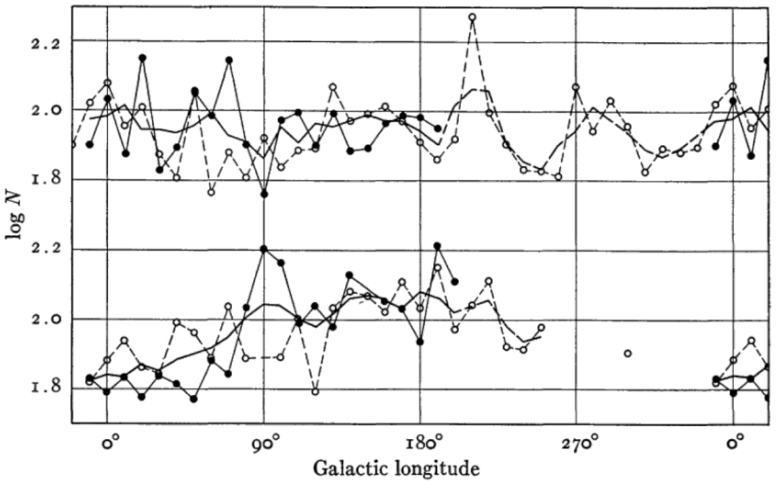

FIGURE 2.3. Hubble's (1934) counts of galaxies at high galactic latitudes in the upper curves, and at low latitudes in the lower curves. © AAS. Reproduced with permission.

> sampling for a homogeneous population. Statistically uniform distribution of nebulae appears to be a general characteristic of the observable region as a whole.

Bok's (1934, 8) considerations led him to the opposite conclusion:

> Different lines of evidence all indicate that the available material points to the existence of a widespread non-uniformity in the distribution of external galaxies, and that this tendency toward clustering is probably one of the chief characteristics of the part of the Universe within the reach of modern telescopes.

Bok was at the Harvard College Observatory, and he emphasized the clumpy distribution of galaxies in the Harvard Shapley-Ames map that came out of this observatory. He referred to Hubble (1934) but did not mention Hubble's Figure 4, which is reproduced here in Figure 2.3. Hubble took it to be an indication of approach to uniformity; Bok does not seem to have been convinced.

Hubble's interpretation seems to be the more reasonable to me, and I count it as the first indication that in the average over large enough volumes, the galaxy distribution approaches isotropy. That is easier to see now, of course. And it is easier to see that if we may take it that our position among the galaxies is not special, then the indication from this figure, though certainly preliminary, was that the galaxy distribution approaches homogeneity on large scales.

Another line of evidence opened in the 1950s with the ability to probe the universe at radio wavelengths and soon after that by X-ray and microwave detectors. Figure 2.4 shows the distribution of radio source positions across

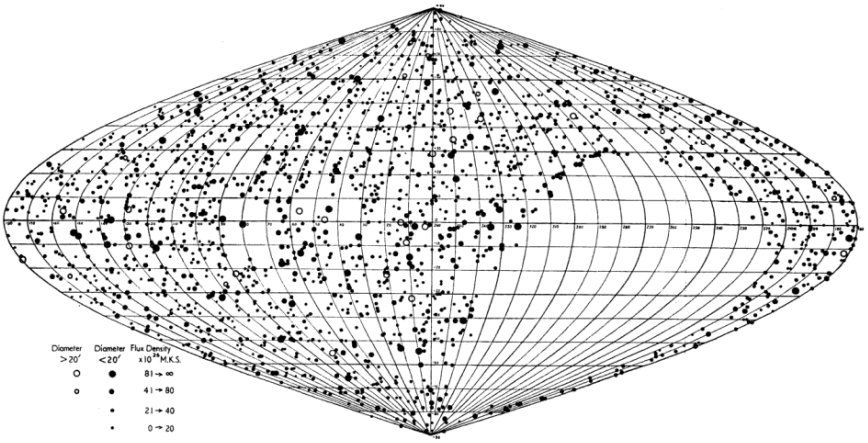

FIGURE 2.4. The Second Cambridge Catalog of Radio Sources (Shakeshaft, Ryle, Baldwin, et al. 1955).

the part of the sky surveyed in the *Second Cambridge Catalog of Radio Sources, 2C,* by Shakeshaft et al. (1955).[1] The sources were suspected then and are now known to be in galaxies. The catalog lists 1,936 sources at wavelength 3.7 meters (82 MHz). A few are close to the plane of the Milky Way and likely are in our galaxy. Others are spurious detections of sources in sidelobes, and some real sources are missing. The brightest extragalactic radio source in the sky, Cygnus A, is on the equator in this map and a quarter of the way in from the left-hand side. It is so bright in the radio that it obscures sources close to it in the map, accounting for the empty region around this object. (The large empty region to the lower right was not observed, because it always is below the horizon at the telescope.)

Optical identifications and redshift measurements of a few of these sources had suggested that many have redshifts large enough that the observations might show a detectible departure of the count of sources as a function of the radio flux density from what would be expected in the flat spacetime of special relativity. This is the cosmological test to be discussed in Section 3.4. Its application here was frustrated by spurious source detections and omissions. This systematic error has a less serious effect on the angular distribution of sources, however, and we see that the constant-area map of sources in Figure 2.4 does look about as expected in a homogeneous universe: no indication in any direction that the observations encounter an edge to the distribution of these objects.

We are in seas of X-ray and microwave radiation. The latter, later termed the "cosmic microwave background" (CMB), is the subject of Chapter 4. A 6-minute rocket flight gave the first evidence of the former, a sea of X-rays

1. This is the figure between pages 148 and 149 in Shakeshaft et al. (1955).

departures from Hubble's law to probe the gravitational effect of departures from a mass distribution that is homogeneous in the mean.

The other probe Bondi mentioned is the count of galaxies as a function of their brightness in the sky. If the space distribution of galaxies is homogeneous on average over the volumes sampled, and if we can neglect the relativistic corrections that are important at great distances, then the count of galaxies brighter than the received energy flux density f varies with f as

$$N(>f) \propto f^{-3/2}. \tag{2.5}$$

To see this, consider the inverse square law: A galaxy with luminosity L produces starlight energy flux density $f = L/(4\pi r^2)$ at distance r (neglecting obscuration and relativistic corrections). The galaxies with luminosity L that are observed to be brighter in the sky than f thus are observed at distances $r < \sqrt{L/(4\pi f)}$. In a homogeneous distribution, the count of galaxies with luminosity L that are brighter in the sky than f is proportional to the volume within distance r, which is proportional to $r^3 \propto f^{-3/2}$. This power-law scaling applies to galaxies in each class of luminosity, so it applies to the counts summed over galaxies of all luminosities, resulting in equation (2.5).

The astronomers' measure of energy flux density f is the apparent magnitude[3]

$$m = -2.5 \log_{10} f + \text{ a constant}. \tag{2.6}$$

The count-magnitude relation for a statistically homogeneous distribution of galaxies is then

$$\log_{10} N(<m) = 0.6m + \text{ another constant}. \tag{2.7}$$

The simple but valuable relation in Equations (2.5) and (2.7) was first applied to star counts. It revealed the limited extent of our Milky Way galaxy of stars by the departure from this relation.

Hubble (1926, 366) compared the relation to counts of galaxies and concluded that

> The agreement between the observed and computed log N over a range of more than 8 mag. is consistent with the double assumption

3. The astronomers' measure of intrinsic luminosity, L, is the absolute magnitude M defined by

$$M = -2.5 \log_{10} L + \text{ another constant}, \quad m - M = 5 \log_{10} d/(10 \text{ pc}).$$

The distance is d, and the distance modulus is $m - M$, normalized to $m = M$ at 10 parsecs distance, where 1 parsec is roughly 3 light years. Measurements of apparent and absolute magnitudes specify the window of wavelengths in which the radiant energy is measured, the correction for obscuration by the atmosphere and dust in the Milky Way and the source, and the shift of wavelengths if the distance is large. But that calculation should be left to more able hands.

> of uniform luminosity and uniform distribution or, more generally, indicates that the density function is independent of the distance.

This early recognition of evidence of homogeneity from galaxy counts is impressive, but the case was based on heterogeneous samples. The more systematic compilation of counts in Hubble (1936, 186) reaches impressively large distances, to recession velocities of about 40 percent of the speed of light (in the estimate by Peebles 1971a, 37). These counts increase with decreasing energy flux density f a little less rapidly than the $f^{-3/2}$ law. It could mean that the universe at great distances is slightly less dense than nearby, or that Hubble had a modest systematic error in his apparent magnitude scale, or perhaps that he had detected the relativistic correction. But we can conclude that the counts did not offer any indication that Hubble's observations of distant galaxies were reaching an edge to the realm of the galaxies.

The $f^{-3/2}$ law for counts, and the redshift-magnitude relation, assume space between the galaxies is fully transparent. Zwicky (1929) asked whether light passing through the great distances of intergalactic space might suffer a friction of some sort that causes photons to lose energy, a concept that became known as "tired light." With Einstein's expression for the energy of a photon, $\varepsilon = h\nu$, the tired light picture would indicate that photon wavelengths increase as they travel great distances and lose their energy ε. Might this have produced the redshifts of the galaxies? And might the friction also make free space slightly opaque? Hubble and Tolman (1935) proposed a test of the first, and implicitly the second, from the variation of galaxy surface brightnesses[4] with redshift, modeled as

$$i \propto (1+z)^{-r}. \tag{2.8}$$

In the standard theory, redshifts are the result of the expansion of the universe. This produces index $r = 4$ by the considerations discussed in Section 4.1: one power of $(1+z)$ comes from the loss of energy of each photon as it is redshifted, one power from the decrease in the rate of reception of photons, and two from the Doppler aberration of solid angle. And if free space in the expanding universe were not fully transparent, it would make $r > 4$. In a static tired-light universe, only the first effect operates, meaning $r = 1$, assuming space is transparent.

This elegant surface-brightness test is unaffected by space curvature, but its application is complicated by the difficulty of modeling the evolution of galaxy intrinsic surface brightnesses as stellar populations evolve. But we have a demanding test from another direction: the sea of microwave radiation

4. Radiation surface brightness is the net flux of energy integrated over frequency per unit area, time, and solid angle. In a static situation, the surface brightness along the path of a light ray is constant. This is Liouville's theorem applied to light modeled as a gas of photons. A Doppler shift in flat or curved spacetime produces the index $r = 4$ in equation (2.8).

discussed in Chapter 4. The close-to-thermal spectrum shown in Figure 4.7 shows that surface brightness evolution closely agrees with the Doppler effect with $r = 4$. Scattering by free space, not absorption, would not disturb the thermal spectrum of the sea of radiation, but it would tend to smooth the radiation anisotropy. The tests for this effect reviewed in Chapter 9 indicate that as much as a few tens of percent of the radiation from high redshift, back to the dark ages, may have been Thomson scattered by free electrons in intergalactic plasma (as first indicated in Spergel et al. 2003). In the standard big bang model, this means galaxy counts increase with decreasing flux density f less rapidly than expected in equation (2.5), apart from the effects of galaxy evolution, and redshifts increase with increasing apparent magnitude less rapidly than expected from equation (2.1). But the effects of this Thomson scattering are small at redshifts less than unity.

Let us note also that relativistic corrections are important for the modern deeper and more precise observations discussed in Section 9.1, but not for what could be done in the 1930s. The impressive thing is that there were observations in the 1930s that probed to galaxies distant enough that their redshifts indicate they are moving away from us at near the speed of light, and the observations did not encounter an edge to the realm of the galaxies.

2.5 The Universe as a Stationary Random Process

Following Jerzy Neyman (1962), a more formal statement of Einstein's cosmological principle is that the universe is assumed to be a realization of a stationary (statistically homogeneous and isotropic) random (stochastic) process. The stationary condition means that expectation values are independent of position and direction; only relative positions matter. The concept can only be empirically useful if the realization of the process that is our observable universe offers a close to fair sample, from which we can find estimates of the statistical measures of the galaxy distribution that usefully approximate these measures in the idealized, infinitely realizable process. The test is the check of reproducibility of statistical estimates that sample different parts of the sky at different ranges of distance.

Jerzy Neyman and Elisabeth Scott at the University of California, Berkeley, introduced a pioneering program of statistical analyses of the galaxy space distribution. They fitted counts of galaxies in cells in the sky to a model of galaxies in clusters of ν members, where the random number ν may assume the value unity, and the clusters may be in superclusters containing random numbers of clusters. They constrained model parameters by second moments of the counts. This program was motivated at least in part by Donald Shane's observational program at the nearby Lick Observatory. He was leading the cataloging of the million or so brightest galaxies in the sky (Neyman, Scott,

and Shane 1954; Shane and Wirtanen 1954 and 1967). The Neyman et al. program clarified the philosophy of statical analyses of extragalactic objects, and it foreshadowed the halo occupation distribution program that became a useful tool for analyses of galaxy distributions in the twenty-first century (e.g., Berlind and Weinberg 2002). But their program is not well suited to probe for a large-scale approach to homogeneity. That was achieved by measurements of galaxy N-point position correlation functions.

Limber (1953 and 1954), Rubin (1954), and Totsuji and Kihara (1969), introduced the use of the two-point statistical measure of the galaxy distribution. All used or mentioned the ongoing Lick counts. Limber's (1954, 656) description of how he estimated the two-point angular correlation of galaxy counts in cells is worth recording here:

> The number of nebulae per square degree for each degree along such a parallel was recorded separately on each of two strips [of paper]. In order to obtain $\overline{NN}_\phi$ for this parallel, one strip was displaced ϕ degrees with respect to the other, and then the values on the two strips which were adjacent after the displacement were multiplied together, and the mean value of these products was obtained.

The quantity $\overline{NN}_\phi$, after normalization and subtraction of shot noise, is an estimate of the angular two-point correlation function at separation ϕ; it has come to be a widely used statistic in extragalactic astronomy. But fuller use of the data from the Lick survey and other catalogs awaited the advances in computation in the 1970s that replaced Limber's labor-intensive method. I took advantage of this in the program of statistical analysis with colleagues at Princeton University. The results are summarized in Peebles (1980).

The N-point correlation functions represent the structure of the universe by a distribution of point-like particles: perhaps galaxies, perhaps mass elements. The probability that a particle is found in the volume element dV is

$$dP = n\,dV. \tag{2.9}$$

This defines the mean particle number density, n. In the assumed stationary process, n is independent of position. The probability that a particle is found in the volume element dV at distance r from a particle is

$$dP = n(1 + \xi(r))\,dV. \tag{2.10}$$

This defines the reduced two-point correlation function, $\xi(r)$ (where "reduced" simply means removal of the first term in parentheses and removal of the factor n). Under the assumption of statistical homogeneity and isotropy, this two-point statistic can only be a function of the separation r of the two points. If the realization we observe has presented us with a good approximation to a fair sample, then the estimate of $\xi(r)$ from the observations is a good approximation to the function in the idealized random process. The higher-order

reduced N-point functions, $N > 2$, are similarly defined, as discussed at length in Peebles (1980).

The mean (expectation value or ensemble average) number of particles within distance r of a particle is, from equation (2.10),

$$\langle N(<r)\rangle = n\,V + n\int_0^r 4\pi r^2\,dr\,\xi(r), \tag{2.11}$$

where V is the volume within distance r (and distances are small compared to the Hubble length, so we can think in term of flat space).

In a stationary random Poisson process, each particle position is assigned independently of where the other particles are. In this case, $\xi = 0$, and the mean number of neighbors is the usual product of the number density n with the volume V within radius r. To avoid confusion, note that in a stationary Poisson process, the volume V randomly placed contains nV particles on average, but the volume placed on a randomly chosen particle biases the count of particles contained to $nV + 1$ particles on average.

The second term in equation (2.11) is the mean number of neighbors in excess of the Poisson distribution. It can be negative, if particles tend to avoid one another. If the two-point function is positive and a power law, which is close to what is observed for the galaxies, we have

$$\xi(r) = (r_0/r)^\gamma, \quad \langle N(<r)\rangle = n\,V + \frac{4\pi n}{3-\gamma} r_0^\gamma r^{3-\gamma}. \tag{2.12}$$

Another way to write the second expression is the fractional difference between the mean number of neighbors within distance r of a particle and the mean number $N = nV$ expected if positions were uncorrelated,

$$\frac{\delta N}{N} = \frac{\langle N(<r)\rangle - nV}{nV} = \frac{3}{3-\gamma}\left(\frac{r_0}{r}\right)^\gamma. \tag{2.13}$$

The parameter r_0 is a measure of the clustering length in this power-law model. At $r \ll r_0$, the positions are distinctly clustered: A typical particle has many more neighbors than expected if positions were unrelated. At $r \gg r_0$, the mean departure from an uncorrelated Poisson distribution is a small fractional perturbation to the count.

Measuring $\xi(r)$ and the higher order functions requires finding a way around the relatively large uncertainties of galaxy distance measurements. The approach is indirect: Infer $\xi(r)$ from estimates of the angular two-point correlation correlation function $w_d(\theta)$ in maps of angular positions of galaxies that have distance estimates d in some chosen range of values. The errors in the galaxy distance estimates are assumed to be uncorrelated, though one can devise corrections for that. The probability distribution of distance errors is supposed to be reasonably well understood. And we have the powerful assumption that the distribution is statistically isotropic.

count N of galaxies in a randomly placed sphere of radius r. For the power-law correlation function, this latter statistic is[6]

$$\left(\frac{\delta N}{N}\right)^2 = \frac{\langle (N - \langle N\rangle)^2\rangle}{\langle N\rangle} = J_2 \left(\frac{r_0}{r}\right)^\gamma, \quad J_2 = 1.82 \text{ for } \gamma = 1.77. \tag{2.17}$$

A measure of the transition from nonlinear clustering on small scales to small departures from homogeneity on large scales is the sphere radius $r_{\rm cl}$ at which the galaxy counts fluctuate from the average by the root-mean-square fractional amount unity:

$$\frac{\delta N}{N} = 1 \text{ at clustering length } r_{\rm cl} = 7.6 h^{-1} \text{ Mpc}. \tag{2.18}$$

The small value of this characteristic clustering length compared to the Hubble length, $H_0 r_{\rm cl}/c \sim 0.003$, is an indication that the observable universe presents us with many different patches of clustering that allow many probes of the galaxy distribution that may be expected to yield fair and secure statistical measures of this random process. Patterns are seen in galaxy maps on considerably larger scales than $r_{\rm cl}$, but they are small fractional fluctuations in the counts of galaxies averaged over larger scales.

The scaled two-point angular correlation function in Panel (b) in Figure 2.5 breaks below the power law at large separation. Because the angular function is a convolution of the spatial function over the range of distances sampled, the spatial function $\xi(r)$ rises slightly above the power law at $r \sim$ 10 Mpc and then falls below it. This is demonstrated by Soneira and Peebles (1978, Fig. 6). The break from a power law is shown with greater precision in Efstathiou, Sutherland, and Maddox (1990) and Zehavi et al. (2011, Fig. B22).

These statistical measures apply to the common large galaxies, such as the Milky Way, that contribute the bulk of the cosmic mean luminosity density. Their characteristic luminosity is written as L^*. The more numerous galaxies with $L \ll L^*$ have close to the same clustering parameters as $L \sim L^*$ galaxies. The rare giants with $L \sim 10L^*$ are more strongly clustered. For example, Masjedi et al. (2006) find that the Luminous Red Galaxy (LRG) sample from the Sloan Digital Sky Survey (SDSS) has clustering length about twice that of $L \lesssim L^*$ galaxies. This is consistent with the tendency of the most massive galaxies to appear preferentially in the most massive clusters of galaxies, because the cluster positions are more strongly clustered than are common $L \sim L^*$ galaxies (Peebles and Hauser 1974, eq. [47]; Bahcall and Soneira 1983). Kaiser (1984) made the excellent point that massive concentrations of galaxies are

6. This ignores the shot-noise term in equation (60.3) and uses the analytic expression for J_2 in equation (59.3) in Peebles (1980).

expected to be more strongly correlated than are the common $L \sim L^*$ galaxies in a positively correlated Gaussian random process, as observed. This is discussed in Section 3.5.3.

2.6 A Fractal Universe

Bondi (1952, 14, 15) mentioned another kind of statistical homogeneity: a clustering hierarchy, or what later became known as a fractal galaxy distribution. For example, imagine that particles, perhaps galaxies, are placed in clusters, clusters are placed in second-order clusters, second-order clusters are in third-order clusters, and so on, perhaps continuing to indefinitely large scales. In a scale-invariant clustering hierarchy, or fractal, the mean number of galaxies (or the mean amount of mass) within distance r of a particle (or mass element) is the limit of equation (2.12) as $n \to 0$ and $r_0 \to \infty$. This amounts to

$$\langle M(<r)\rangle \propto r^{3-\gamma} = r^D, \quad D \equiv 3-\gamma. \tag{2.19}$$

In astronomers' units, this is

$$\log\langle M(<m)\rangle = 0.2Dm + \text{constant}. \tag{2.20}$$

The distribution is said to have fractal dimension D, with $0 < D < 3$ in three dimensions. If the distribution is spatially homogeneous, then $D = 3$, as usual. If $D < 3$, the distribution may be homogeneous in another sense: that is, each element of mass finds itself in statistically the same hierarchy of clusters within clusters and on up. But if $D < 3$, the mean mass density averaged over arbitrarily large scales is arbitrarily close to zero.

The Newtonian gravitational potential energy of a mass M concentrated within radius r is on the order of $U \sim GM/r$. In a fractal mass distribution with dimension D, the potential energy on the scale r thus varies as $U \propto r^{D-1}$. A pure scale-invariant fractal in three dimensions with $D = 1$ thus has gravitational potential that diverges only as the logarithm of the length scale on arbitrarily small and large scales. If kinetic energy scales like potential energy, then velocities would would be safely below the velocity of light over a broad range of scales. This could be an elegantly arranged universe, but it is not ours.

The galaxy distribution on scales less than about 10 Mpc approximates a fractal with dimension $D = 1.23$ (equation (2.16)). The three- and four-point correlation functions also agree with a simple fractal hierarchical clustering pattern with this dimension (Groth and Peebles 1977; Fry and Peebles 1978). In a gravitationally bound and stable clustering pattern, the relative velocity dispersion of particles scales as the square root of the mean gravitational potential difference at their separation. Since the small-scale galaxy

distribution has D slightly greater than unity, one might expect the galaxy relative velocity dispersion to increase slowly with increasing length scale. This is observed. But the departure downward from the power law form on a scale of about 20 Mpc is a well-established departure from scale invariance.

Pietronero, Gabrielli, and Sylos Labini (2002) make the interesting point that the ratios of depths of the catalogs in Figure 2.5 are scaled from the mean angular densities as $d \propto \mathcal{N}^{1/3}$, which assumes large-scale homogeneity. This argument is circular if we are seeking to check that the galaxy density has a nonzero mean. The circularity is mitigated by the fact that the galaxy distribution is sampled on length scales large compared to the clustering length in equation (2.18). And we have now an independent check: the weak lensing distortion of background galaxy images caused by the masses concentrated around foreground galaxies yields the galaxy-mass cross-correlation function $\xi_{g\rho}(r)$. Sheldon et al. (2004) find that $\xi_{g\rho}(r)$ is a good approximation to a power law at the range of separations $0.04 \lesssim r \lesssim 12$ Mpc, with $\gamma = 1.79 \pm 0.06$ and $r_0 = (5.4 \pm 0.7)h^{-1}$ Mpc. Within the uncertainties, these values agree with the parameters in equation (2.16) based on the scaling test for the galaxy-galaxy function.

In his books, *Les objets fractals*, Benoît Mandelbrot (1975 and 1989) reviews earlier discussions of clustering hierarchies and names them "fractals." He presents ample examples of fascinating fractal patterns, including mathematical constructions and pragmatic considerations, such as the measurement of the length of the coastline of Brittany, which behaves as a fractal because the length depends on the spatial resolution at which it is measured. Mandelbrot's stroke of genius forced attention on many interesting and practical applications of fractals. Perhaps it was inevitable that he should consider the idea that the galaxy distribution is fractal.

Others were thinking along similar lines. I noted in Section 2.2 Charlier's (1922) argument that the galaxies seem to be arranged in a hierarchal clustering pattern. Indeed, this is now well established at the distances Charlier could observe. Carpenter (1938) argued that the galaxy distribution fits $\langle M(<r)\rangle \propto r^D$ with dimension $D = 1.5$ in Mandelbrot's notation. Along with Oort (1958) and Abell (1958), de Vaucouleurs (1970) pointed out that maps of the galaxy spatial distribution probing out to the greatest distances that could be reliably surveyed offered no hint of convergence to homogeneity. Oort was willing to consider that the universe approaches homogeneity on still larger scales, on the basis of Hubble's deep galaxy counts, but de Vaucouleurs proposed that Carpenter's scaling relation extends to much smaller and much larger scales, in a "universal density-radius relation" with fractal dimension he put at $D = 1.3$.

Hubble's (1936) deep galaxy counts shown in Figure 16 in his book, *The Realm of the Nebulae*, fit fractal dimension $D = 2.6$. Reconciling this with large-scale homogeneity, $D = 3$, and neglecting relativistic corrections

requires the postulate that a systematic error exists in Hubble's distance scale. Reconciliation with de Vaucouleurs' $D = 1.3$ requires a larger systematic error in the other direction. Gérard de Vaucouleurs certainly considered this point: He told me that he examined Hubble's photographic plates for the deep galaxy counts but could not check the magnitude calibration, because the plates had faded. The systematic error needed to reconcile Hubble's counts with de Vaucouleurs' fractal dimension certainly was worth considering, but to be considered also is the line of evidence from the clustering scaling test in Figure 2.5. De Vaucouleurs' $D = 1.3$ is quite close to the measured correlation function power law on smaller scales, $D = 1.23$, but the measurements reviewed on page 30 indicate a break downward from the power law at separation ~ 20 Mpc: a break from scale invariance.

A qualitative point to be considered is the appearance of maps of angular positions of galaxies within a given distance, d, for different values of d. In a scale-invariant fractal with dimension $D < 3$, the number of particles in the map increases with increasing distance d, but the fractional fluctuations $\delta N/N$ of particle numbers across the sky are statistically independent of d. This follows from the scale invariance of the fractal distribution: If the fluctuations $\delta N/N$ grew smaller with increasing distance d, it would define a characteristic distance $d_{\rm nl}$ at which $\delta N/N$ at a chosen angular scale decreases through unity to small, linear, fractional fluctuations across the sky. But a scale-invariant fractal does not have a characteristic length $d_{\rm nl}$. For early examples of this test, compare the Shapley-Ames map of the angular positions of relatively nearby galaxies in Figure 2.2 (which shows the large number of galaxies near the north relative to the south galactic poles) to Hubble's (1934) deep counts of galaxies across the sky at higher galactic latitudes in the two hemispheres in Figure 2.3, and to the angular positions of distant radio galaxies in Figure 2.4. We see convergence toward isotropy that is contrary to a scale-invariant fractal behavior. In addition, if the galaxy distribution were fractal, the reduced correlation functions in Panel (a) in Figure 2.5 would not decrease with the increasing depths of the three samples.

At the Bern Conference on the *Jubilee of Relativity Theory* (Jubilee for the 1905 special theory; 40 years for the general theory), Oskar Klein (1956) discussed yet another world picture: Perhaps the galaxies are drifting apart into empty flat space after an explosion of a local concentration of matter. Klein considered that the total mass M_0 in the galaxies, and the radius R_0 in which matter was concentrated prior to the explosion, might satisfy $GM_0 \sim R_0 c^2$. This is a relativistic concentration at about the Schwarzschild radius. It could be in accordance with the near-relativistic expansion indicated by Hubble and Humason's observations of galaxy redshifts that are not far below unity. Klein had some good arguments. Velocity sorting would put more rapidly moving galaxies farther away, approaching Hubble's relation $v = H_0 r$. Explosions are familiar; why not consider a particularly large one that scattered the galaxies?

Flat spacetime is familiar; why bother with the notion of spacetime curvature? The clumpy and irregular galaxy distribution that might be expected from an explosion is familiar; why bother with the homogeneity that was not seen in galaxy maps at the time? Klein's commonsense model was viable then and might have been expected to have attracted broader interest than it did. Klein (1966) continued the argument for this picture, in a paper with the title "Instead of Cosmology." Hannes Alfvén (1965) added to Klein's picture a matter-antimatter universe, whose expansion was driven by the radiation pressure derived from annihilation of much of the matter and antimatter. But by this time, the observations had seriously challenged the notion that the observable universe has an edge.

2.7 *Concluding Remarks*

In the 1950s, there was some observational evidence for the cosmological principle, large-scale homogeneity, from Hubble's demonstration of the approximate isotropy of distant galaxy positions shown in Figure 2.3 and from the linearity of the Hubble and Humason redshift-distance relation. A nearly exact homogeneity and isotropy would offer the great convenience of an analytic solution to Einstein's field equation and Milne's elegant derivation of Hubble's law. But an alternative picture, a fractal universe, had the support of one of the best observational astronomers in the years around 1970, Gérard de Vaucouleurs, and it had the support of Benoît Mandelbrot's elegant examples of fractals in mathematics and physics. The fractal picture for the galaxy distribution rightly attracted attention and inspired debates and research. But a more vigorous promotion of the idea earlier, in the 1950s, could have been more productive, because by the 1970s, the totality of evidence reviewed in this chapter had made it clear that the fractal picture is not promising. The evidence in the 1970s instead favored the cosmological principle, which is best put as the assumption that our universe is a realization of a stationary random process. This assumption is implicit in the demanding cosmological tests reviewed in Chapter 9, and it continues to pass checks based on consistency.

Arguments from elegance and observation can instruct or mislead; we see examples of both in the history of the cosmological principle. We also see that, at least on occasion, confusion of this sort can be resolved. Cosmology in the 1970s could operate on the reasonably secure assumption of large-scale homogeneity, and the evidence of it has continued to grow more secure.

An abstract stationary random process has no edge. That may be true of our universe, or there may be edges where the universe is different from what we see. It would have to be far enough away not to have significantly disturbed the

$$ds^2 = dt^2 - a(t)^2 \left[\frac{dr^2}{1 - r^2 R^{-2}} + r^2 \left(d\theta^2 + \sin^2 \theta d\phi^2 \right) \right]. \qquad (3.1)$$

The spatial coordinates r, θ, ϕ are comoving (that is, fixed to the mean streaming motion of the material contents of the universe), and the factor R^{-2} is a constant (which, despite the notation, can be positive or negative). This expression is usually termed the Robertson-Walker form, after the recognition by Robertson (1929) and Walker (1935) that it follows from spatial homogeneity and isotropy in a spacetime described by a line element. It applies to the steady-state cosmology as well as the relativistic picture.

Comoving observers, at fixed r, θ, ϕ, keep the proper physical time t in equation (3.1). The expansion factor $a(t)$ in equation (3.1) appears in the discussion of the linear redshift-distance relation in Figure 2.1 and equation (2.3). Two events at the same world time t, at coordinate distance r from an observer who sees that the events are separated by the small angle $\delta\theta$, are at physical separation $\delta l = a(t) r \delta\theta$. If the constant R^{-2} is small enough to be neglected, the physical distance from the origin to coordinate radius r at given world time t is $l = a(t) r$, and the rate of change of the physical distance is

$$v = \frac{dl}{dt} = Hl, \quad H = \frac{1}{a}\frac{da}{dt}. \qquad (3.2)$$

This is Hubble's law in equation (2.1), with H evaluated at the present epoch, and ignoring space curvature and the relativistic correction for observation along the past light cone instead at fixed t.

In general relativity with Einstein's cosmological constant Λ, the expansion parameter $a(t)$ satisfies the Friedman-Lemaître equations:

$$\left(\frac{1}{a}\frac{da}{dt} \right)^2 = \frac{8}{3}\pi G \rho(t) - \frac{1}{a^2 R^2} + \Lambda, \quad \frac{1}{a}\frac{d^2 a}{dt^2} = -\frac{4}{3}\pi G(\rho + 3p) + \Lambda, \qquad (3.3)$$

with the expression for the local conservation of energy being

$$\frac{d\rho}{dt} = -\frac{3}{a}\frac{da}{dt}(\rho + p). \qquad (3.4)$$

The mean mass density, including the mass equivalent in radiation energy, is $\rho(t)$, and the pressure is p (with units chosen so the speed of light is unity). To understand the energy equation, recall that the energy $\varepsilon = \rho V$ in a container of volume V changes when the volume changes at the rate $d\varepsilon/dt = -p\, dV/dt$ when the pressure is p. The energy equation follows by setting $\varepsilon = 4\pi\rho a^3/3$ and working out the time derivative.

The factor R^{-2} in the first expression in equation (3.3) may be considered a constant of integration in the sense that, with the energy equation (3.4), the time derivative of $(da/dt)^2$ in the first expression is the second expression. But in general relativity, R^{-2} also defines the geometry of space at fixed t. If R^{-2} in equation (3.1) is positive, the spatial geometry is closed—the analog

of the surface of a sphere; if negative, it is the shape of a saddle. If $R^{-2} = 0$, the spacetime is said to be cosmologically flat, even though spacetime may be curved.

In the first step to modern cosmology, Einstein (1917) found the static solution to his field equation in general relativity for a universe that is homogeneous and isotropic, consistent with his thinking at the time about Mach's principle. Perhaps the condition that it is static seemed perfectly reasonable at the time. To get this solution, he had to modify his original relativistic field equation by adding what became known as the cosmological constant, Λ. Then, if pressure can be neglected, the conditions $da/dt = 0 = d^2a/dt^2$ in equation (3.3) require that the mass density ρ and the parameter R^{-2} representing space curvature in equation (3.1) satisfy

$$\Lambda = \frac{4}{3}\pi G\rho = \frac{1}{3a^2R^2}. \tag{3.5}$$

Here aR is the physical radius of curvature of space.

Eddington (1923, 166) noticed that equation (3.5) represents a curious situation: A dynamical variable, the mass density, must agree with a constant of nature, Λ. He wrote that "the question at once arises, by what mechanism can the value of λ [now Λ] be adjusted to correspond with M [a measure of the mean mass density]?" One might also wonder what happens if the mass is rearranged so this condition is locally violated. These questions offered an early hint that Einstein's static model is unstable. This is discussed in Chapter 5.

It takes nothing from Einstein's genius to note that a static universe does not make physical sense in conventional thinking: Recall the Olbers problem discussed in Section 2.1. The Russian Alexander Friedman showed the way out, by generalizing Einstein's solution to a homogeneous expanding or contracting model universe (Friedman 1922, 1924). Einstein's first judgments—that Friedman did not have a correct solution to the field equation in general relativity, and then that the solution is correct but unphysical—have been well reviewed, as by Goenner (2001) and Longair (2006). The Olbers problem with an unlimited buildup of starlight is removed in an expanding universe, because we can assume that the stars have limited lifetimes, in accordance with local energy conservation. And there is extra help from the expansion of the universe, which dilutes the energy density of the starlight. But I have seen no evidence that Friedman recognized he had solved Olbers' problem.

It is unfortunate that Friedman died before recognizing a possible connection between theory and observation: the observed tendency of galaxy spectra to be shifted to the red in proportion to their distance, as would be expected in a homogeneously expanding universe.

The Hungarian mathematical physicist Kornel Lanczos (1923) introduced a coordinate labeling of de Sitter's (1917b) solution for empty homogeneous

and isotropic spacetime with a positive cosmological constant. In the notation of equation (3.1), Lanczos' form (in his eq. [32]) is

$$ds^2 = dt^2 - \cosh^2\left(\sqrt{\Lambda}t\right)\left[\frac{dr^2}{1-\Lambda r^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right]. \tag{3.6}$$

This is Friedman's solution for closed-space sections with a positive cosmological constant in the limit of vanishing mass density. Lanczos did not take explicit note of the possible relation to the astronomical evidence, however. The Belgian Georges Henri Joseph Édouard Lemaître (1925, 192) also reported this solution for vanishing mass density, and he pointed out that the solution "gives a possible interpretation of the mean receding motion of spiral nebulae."

The American physicist Howard Percy Robertson (1928) reported another coordinate labeling of de Sitter's spacetime, this one cosmologically flat, with $R^{-2} = 0$ and expansion parameter $a \propto e^{\sqrt{\Lambda}t}$. He also pointed to the possible relation to the astronomers' redshift phenomenon.

We see from equations (3.2) and (3.3) that in de Sitter's solution, where pressure and mass density vanish, the physical distance from the origin of a radially moving test particle satisfies

$$\frac{d^2 l}{dt^2} = \Lambda l. \tag{3.7}$$

In the solution $l \propto \cosh\sqrt{\Lambda}t$, the particle falls toward the origin and then moves away, a behavior known in the 1920s as de Sitter scattering. In this solution, and in the solution $l \propto e^{\sqrt{\Lambda}t}$, the late-time behavior is the same: The velocity of recession of a particle becomes proportional to its distance from the origin independent of initial conditions. The same applies to velocity sorting of particles moving away from an explosion: When mass and Λ can be neglected, the end result is that more rapidly moving particles are farther away.

Friedman (1922 and 1924) found the evolving matter-filled solution to Einstein's field equation. Considering Einstein's skepticism in the early 1920s, it is interesting to see in Einstein (1931, 236) his positive attitude in the statement that

> Es ist von verschiedenen Forschern versucht worden, den neuen Tatsachen durch einen sphärischen Raum gerecht zu werden, dessen Radius P zeitlich veränderlich ist. Als Erster und unbeeinflußt durch Beobachtungstatsachen hat A. FRIEDMAN[1] diesen Weg eingeschlagen, auf dessen rechnerische Resultate ich die folgenden Bemerkungen stütze. Dieser geht demgemäß von einem Linienelement von der Form ... Bemerkenswert ist vor allem, daß die allgemeine Relativitätstheorie HUBBELS neuen Tatsachen ungezwungener (nämlich ohne λ-Glied) gerecht werden zu können scheint als dem nun empirisch in die Ferne gerückten Postulat von der quasi-statischen Natur des Raumes.

My condensed translation, aided by Google, is:

> Different researchers have attempted to do justice to the new facts by considering a spherical space whose radius is a function of time. A. Friedman was the first to have embarked on this path, unaffected by the observational facts. It is remarkable that the general theory of relativity seems to be able to cope with Hubble's new facts more easily (and without the λ component) than with the postulate of the quasi-static nature of space.

We see that Einstein agrees that his general relativity can do justice to the new facts, and without the Λ term (which he wrote as λ). Why the mention that Friedman was not influenced by observational facts? It can be taken to mean that the redshift-distance relation implicit in Friedman's solution is a prediction of what was later observed by Hubble, whom Einstein mentioned, and others.

Lemaître (1927) generalized his 1925 coordinate labeling to the solution for a matter-filled homogeneous spacetime.[1] Friedman found it first, but the evidence is that Lemaître's discovery presented in 1927 was made independently. Consistent with that is a footnote in Lemaître (1929) that thanks Einstein for telling him about the important work by Friedman. The papers Lemaître (1931a and 1950) also refer to Friedman's prior discovery.

For the purpose of this book, the important advance in Lemaître's 1927 paper is the demonstration that in the expanding matter-filled model, spatial homogeneity allows the redshift of a galaxy to remain proportional to its distance as the universe expands. The essential distinction is that earlier discussions ignore mass, so the linear redshift-distance relation does not require Einstein's homogeneity: Velocity sorting would do. The linear redshift-distance relation in a universe with matter follows from homogeneity; it does not require general relativity (as one sees from equation (2.2) on page 15). But all this is much easier to see now, of course.

Lemaître (1927 and 1931a) referred to earlier discussions of a possible linear relation between galaxy redshifts and distances in empty de Sitter spacetime by Lanczos (1922), Weyl (1923), and Lundmark (1924). The German mathematical physicist Hermann Weyl knew of the observations that galaxy spectra tend to be shifted toward the red, and he proposed that this is because matter has moved apart, based on what causality would suggest must be pictured as matter moving on geodesics diverging from a common origin in the asymptotic past. These would be orbits with $l \propto e^{\sqrt{\Lambda}t}$. We may consider this an early example of Klein's (1956) explosion picture, but aided by de Sitter scattering.

1. I am grateful to John Peacock for the argument that Friedman's solution was the more general: Lemaître considered only closed space sections, while Friedman (1922 and 1924) presented the solutions for the closed, open, and cosmologically flat cases.

Lemaître (1927 and 1931a) did not claim that there is empirical evidence for the linear relation redshift-distance. His impression of the observational situation is suggested by the comment in a footnote in Lemaître (1927, 56) that

> Certains auteurs ont cherché à mettre en évidence la relation entre v et r et n'ont obtenu qu'une très faible corrélation entre ces deux grandeurs. L'erreur dans la détermination des distances individuelles est du même ordre de grandeur que l'intervalle que couvrent les observations et la vitesse propre des nébuleuses (en toute direction) est grande (300 Km./sec. d'après Strömberg), il semble donc que ces résultats négatifs ne sont ni pour ni contre l'interpretation relativistique de l'effet Doppler. Tout ce que l'imprécision des observations permet de faire est de supposer v proportionnel à r et d'essayer d'éviter une erreur systématique dans la détermination du rapport v/r. Cf. LUNDMARK. The determination of the curvature of space time in de Sitter's world M. N., vol. 84, p. 747, 1924.

In brief,

> Attempts to find the relation between v and r have shown at best a weak correlation. Since the distance errors are comparable to the range of redshifts (300 km s^{-1} according to Strömberg), all the observations allow is to suppose v is proportional to r and to try to avoid systematic error in the determination of v/r.

His reference is to Lundmark (1924). The discussion of the redshift-distance relation in this paper may not have inspired confidence. Lundmark had a reasonable estimate of the distance to the nearest large galaxy, the Andromeda Nebula M 31, from Öpik's (1922) ingenious interpretation of its velocity of rotation.[2] But this galaxy is in the Local Group, and it has a negative redshift. Lundmark had only rough estimates of distances to a few other nearby galaxies from apparent magnitudes of variable stars compared to novae in the Milky Way. He supplemented this with redshifts and distances of globular star clusters, but they are much closer and are surely parts of the Milky Way galaxy. Lemaître (1927) did not refer to (and maybe did not notice) the later, more encouraging results reported in Lundmark (1925). This paper takes note of reasonably good distances to M 31 and its companions from Hubble's

2. In outline, let $r = \theta D$ be the radius of M 31 at the observed angular size θ and the wanted distance D, let v be the speed of rotation of M 31, and let $v_\odot$ be the speed of motion of Earth around the Sun with mass $M_\odot$ at distance $r_\odot$. Then the mass M of M 31 satisfies $M/M_\odot \approx (\theta D/r_\odot)(v/v_\odot)^2$, in Newtonian mechanics. The observed energy flux density from M 31 is $f \approx L/D^2$, where L is its luminosity. The combination yields the distance D in terms of the observables θ and f and the mass-to-light ratio M/L, which might be expected to be similar to that of the stars in the Milky Way if M 31 is another galaxy of stars. Öpik found $D = 450$ kpc, impressively close to the modern value, $D = 780$ kpc.

$$\begin{aligned} H_0 &\approx 630 \text{ km s}^{-1} \text{ Mpc}^{-1} \quad \text{Lemaître (1927)}, \\ H_0 &\approx 460 \text{ km s}^{-1} \text{ Mpc}^{-1} \quad \text{Robertson (1928)}, \\ H_0 &\approx 500 \text{ km s}^{-1} \text{ Mpc}^{-1} \quad \text{Hubble (1929)}. \end{aligned} \tag{3.8}$$

The calculation of H_0 in the first line of equation (3.8) is displayed in equation (24) in Lemaître (1927) and discussed in the footnote. Mario Livio's (2011) admirable detective work reveals correspondence that clearly demonstrates the removal of the 1927 footnote from the 1931 English translation was Lemaître's decision. Robertson's (1928) estimate in the second line is based on Slipher's redshift measurements and Hubble's Cepheid variable distances of six galaxies. Robertson did not offer an assessment of the evidence that the relation may be linear.

Hubble (1929) attended to the observations, not the theory. He had distances for six or seven galaxies from observations of Cepheid variable stars, 13 distances for galaxies on the assumption that the most luminous stars have a common intrinsic luminosity, and several galaxies in the Virgo cluster at a not very well understood distance. But Lemaître, Robertson, and Hubble had essentially the same information, which is why their estimates of H_0 in equation (3.8) are similar. My impression is that Lemaître (1927) was unduly pessimistic about evidence for the linear relation, Robertson just trusted the theory, and Hubble (1929) may have been overly optimistic. The redshift-distance plot in Figure 1 in Hubble (1929) seems suggestive of the linear relation but hardly convincing. But Hubble was on the right track, as he and Humason showed in the 1930s.

Why did Lemaître remove the footnote? He states there that there had been no convincing claim of the linear relation. The evidence reviewed by van den Bergh (2011) is that Lemaître and Hubble were aware of earlier indications of a relation between galaxy redshifts and distances, and we see that Hubble referred to Lundmark's (1925) not very encouraging discussion of the evidence. But the situation changed in 1929, when Hubble explicitly announced evidence of the linear relation. The straightforward guess is that Lemaître removed the footnote because Hubble had made it obsolete. But in any case, we can be quite sure the footnote had not been removed in a conspiracy to obscure a prior empirical discovery that Lemaître did not even claim.

Let us pause to notice that Hubble's measurements of galaxy distances were based on Henrietta Leavitt's discovery of the relation between the luminosities and periods of Cepheid variable stars. She reported the key result in Leavitt (1912, 2):

> there is a simple relation between the brightness of the variables and their periods. The logarithm of the period increases by about 0.48 for each increase of one magnitude in brightness.

Hubble's 1929 paper largely relied on Slipher's (1917) redshift measurements. Slipher was one of the excellent astronomers Percival Lowell brought to his observatory constructed to check the possibility of canals on Mars made by an advanced civilization. And in the 1930s, Milton Humason played an important role with Hubble in their considerable extension of distances of galaxies with measured redshifts (as in Hubble and Humason 1931) that greatly improved the case for the linear redshift-distance relation. All this is not meant to depreciate Hubble's contributions—to my mind, he had just the right instincts for the observations that would advance extragalactic astronomy at that time—but to note that Hubble had help in arriving at Hubble's law.

3.2 The Relativistic Big Bang Cosmology

Lemaître's (1927) solution traces the expansion of the universe back in time to Einstein's (1917) static world model. This demonstrates the instability of Einstein's model: A slight homogeneous disturbance to the static situation sets the universe expanding (in Lemaître's solution) or else collapsing. Lemaître (1931b, 706) turned to the idea that the expansion traces back to a dense state. Perhaps, as he wrote, "the world has begun with a single quantum." Lemaître (1931c, 706) termed this "l'atome primitif"; it is now known as the big bang.

McCrea and Milne (1934) showed that if the pressure p and cosmological constant Λ can be ignored, then the Friedman-Lemaître equations (3.3) follow from Newtonian physics. To see this, consider that in a homogeneous universe with mass density ρ, a sphere with radius $a(t)$ much less than the radius of curvature of space contains mass $M = 4\pi\rho a^3/3$. In Newtonian mechanics, the gravitational acceleration of the radius of this sphere is determined by the mass it contains, $d^2a/dt^2 = -GM/a^2$, independent of the spherically distributed mass outside $a(t)$. This is the first of equations (3.3) when p and Λ vanish. The second equation is the integral of the first, where the constant of integration is R^{-2}. This expresses the Newtonian conservation of kinetic plus potential energies. These results follow because general relativity in the limit of $\Lambda = 0$ and for small velocities, and applied to regions small compared to spacetime curvature, is Newtonian mechanics. In particular, the McCrea and Milne argument assumes the flat spacetime of Newtonian mechanics, but that is an arbitrarily good approximation to the relativistic model if we choose a small enough sphere radius. Another theory with this same limiting behavior would do as well for the purpose of this chapter, of course.

We see from the second expression in the Friedman-Lemaître equations (3.3) that the source of gravity tending to slow the expansion is $\rho + 3p$, which is to say that in general relativity, pressure acts as active gravitational mass density. Consistency of the Friedman-Lemaître equations requires it. A Newtonian analog of sorts for the dual role of the constant R^{-2}—conserved

kinetic plus potential energy in equation (3.3) and the measure of the curvature of space sections in equation (3.1)—is that the Newtonian energy of the matter in a sphere of comoving radius r is $U = -r^2/(2R^2)$, which is the departure from flat spacetime in equation (3.1).

Lemaître (1934, 12) introduced the thought that Λ may define the vacuum energy density. If Λ is nonzero then, he wrote,

> Everything happens as though the energy *in vacuo* would be different from zero. In order that absolute motion, i.e., motion relative to vacuum, may not be detected, we must associate a pressure $p = -\rho c^2$ to the density of energy ρc^2 of vacuum. This is essentially the meaning of the cosmical constant λ which corresponds to a negative density of vacuum ρ_0 according to
>
> $$\rho_0 = \frac{\lambda c^2}{4\pi G} \cong 10^{-27}\text{gr./cm.}^3 \tag{3.9}$$

His estimate of the vacuum energy density is too big, because he used Hubble's overestimate of Hubble's constant. The point was made, however: Λ sets the zero of energy, as far as gravity is concerned, and Λ acts as mass with homogeneous energy density and pressure (taking $c = 1$ as usual):

$$\rho_\Lambda = \frac{3\Lambda}{8\pi G}, \quad p_\Lambda = -\rho_\Lambda. \tag{3.10}$$

The sum of this effective pressure and mass density vanishes, as required to keep ρ_Λ constant in the energy equation (3.4) in an expanding universe. As we have seen, consistency of the two relativistic Friedman-Lemaître equations requires that pressure contributes to the active gravitational mass density in the amount $\rho_\Lambda + 3p_\Lambda = -2\rho_\Lambda$. And we can admire Lemaître's recognition that this vacuum energy density is not changed by a velocity transformation. One way to put it is that the relativistic stress-energy tensor $T'^{\mu\nu}$ of a fluid at rest with energy density ρ and pressure p is diagonal with components ρ, p, p, p. With $p_\Lambda = -\rho_\Lambda$, the stress-energy $T^{\mu\nu}_\Lambda$ is proportional to the Minkowski metric tensor and therefore is unchanged by a Lorentz transformation. That is, Λ does not define a preferred frame of motion. (But the idea discussed in Section 3.5.1 that the value of Λ may be evolving, decreasing to its "natural" value, zero, does introduce the preferred motion in which Λ has no spatial gradient.)

The effective mass density ρ_Λ has come to be termed "dark energy." The term first appeared in Huterer and Turner (1999). But the effective negative pressure is not to be associated with the negative pressure of a fluid, which is an unstable situation unless p is exactly the negative of ρ.

The cosmological redshift is defined in an expanding (or contracting) homogeneous and isotropic metric spacetime as follows. Let λ be the physical wavelength of a freely propagating photon, and more generally the de Broglie wavelength of a freely moving particle. This wavelength is to be measured