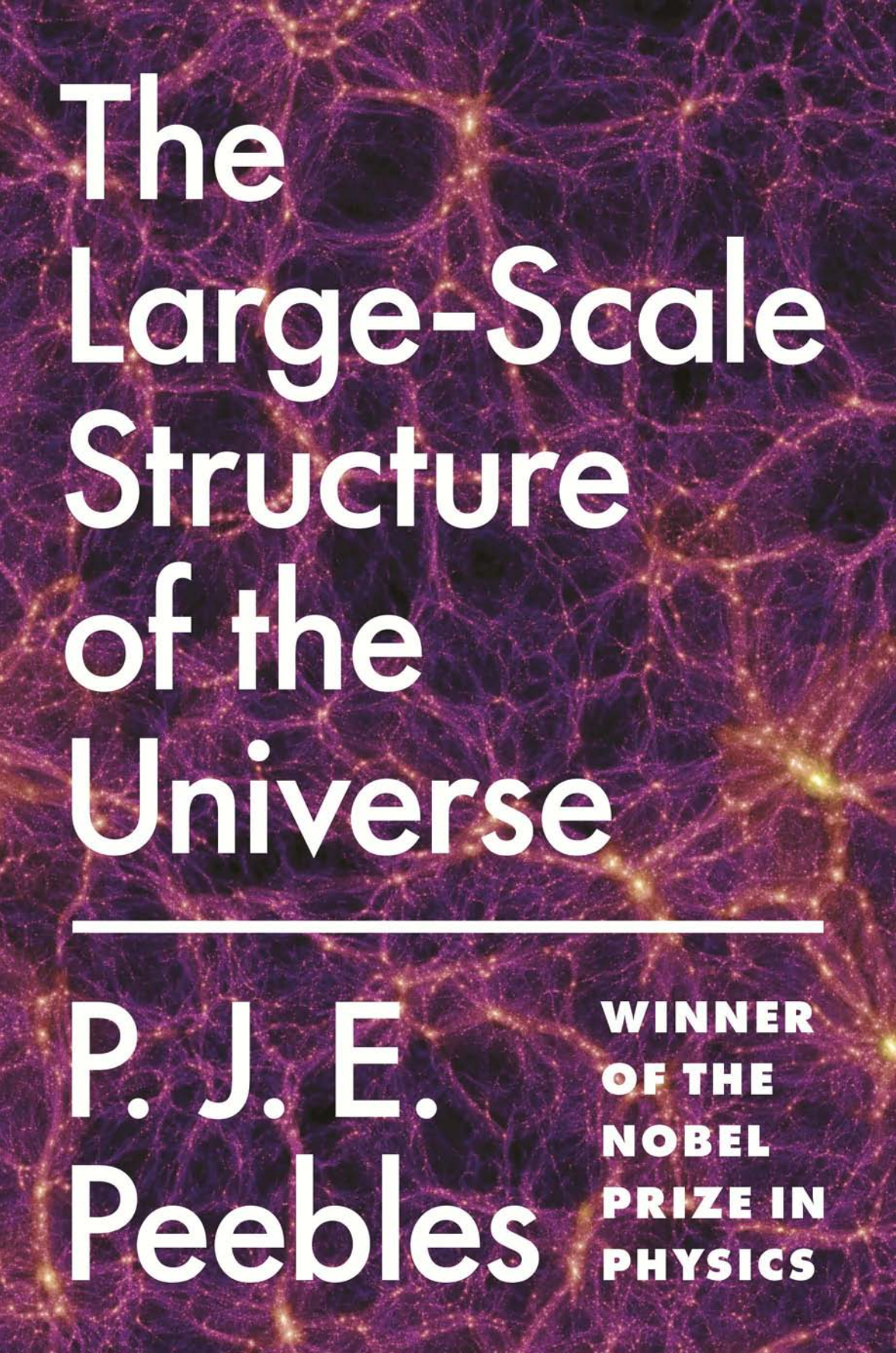
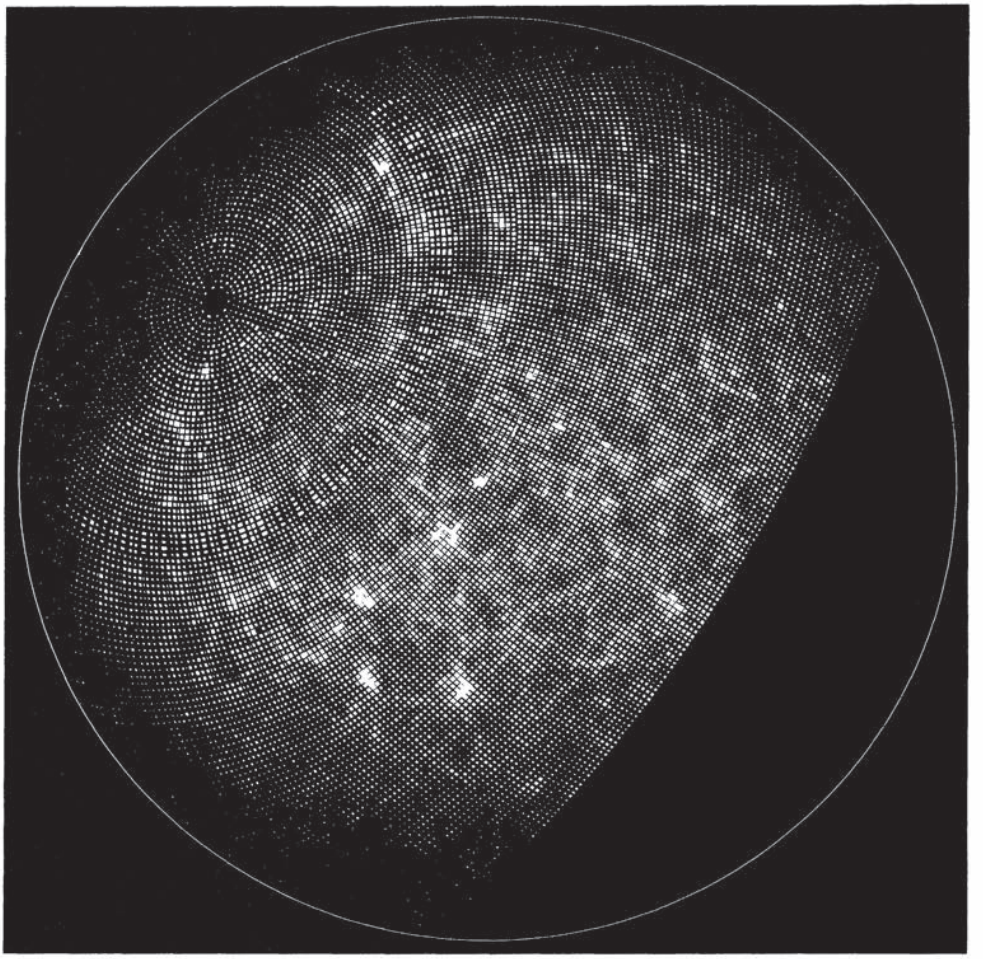

The Large-Scale Structure of the Universe
P. J. E. Peebles
WINNER OF THE NOBEL PRIZE IN PHYSICS

The Large-Scale Structure of the Universe



FRONTISPIECE: The large-scale pattern of the galaxy distribution. Each white square represents a sky cell about one degree by one degree in the Lick sample. The size of the white square is proportional to the number of galaxies brighter than 19th magnitude in the cell. The cells are arranged along lines of fixed right ascension and declination. The north pole of the galaxy is at the center of the map and the equator along the edge. (Map by J. A. Peebles and P.J.E. Peebles.)

The Large-Scale Structure of the Universe

by

P. J. E. Peebles

Princeton Series in Physics

Princeton University Press
Princeton, New Jersey

Cover image: Dark matter distribution in the universe, based on the Millennium Simulation, a very large cosmological N-body simulation (more than 10 billion particles). Courtesy of V. Springel, Max Planck Institute for Astrophysics, Germany

First printed 1980
New paperback printing, 2020
Paperback ISBN 9780691209838

This book was originally published in the Princeton Series in Physics.
For more information on the series, please visit
https://press.princeton.edu/series/princeton-series-in-physics

The Library of Congress has cataloged the cloth edition as follows:

Peebles, Phillip James Edwin.
The large-scale structure of the universe.

(Princeton series in physics)
Bibliography: p.
Includes index.
1. Galaxies. 2. Cosmology. I. Title.
QB857.P43 523.1'12 79-84008
ISBN 0-691-08239-1
ISBN 0-691-08240-5 (pbk.)

Printed in the United States of America

To Alison

CONTENTS

PREFACE

From the first developments of modern cosmology people have recognized that an important part of cosmology is the large-scale clustering of matter in galaxies and clusters of galaxies. The point was largely eclipsed by the debate over homogeneous world models, but in recent years there has been a considerable revival of interest in the large-scale mass distribution and what it might tell us about the nature and evolution of the universe. The purpose of this book is to review our present understanding of these subjects.

Chapter I is a history of the development of ideas on the large-scale structure of the universe. As is usual in science the story is a mixture of inspired guesses and rational progress with excursions down paths that now seem uninteresting. What makes it somewhat unusual is the slow rate of development that has left ample time for the growth of traditions that are more than commonly misleading, and so it seems worthwhile to examine the evolution of the ideas in some detail. This is a history in the rather loose convention of scientists, that is, it is based on what I could glean from published books and journals. The few conversations I have had with participants have left me only too aware of how limited that is and how much more could be done. On the other hand, the published record is what was readily available to most people who might want to work on the subject and might want to learn what has already been done, though the actual use of the record was just as erratic in the 1930s as it is these days. I have tried to give a complete account of the important developments since about 1927 and have added enough more recent references to serve as a guide to the literature.

Chapter II deals with the behavior of a given mass distribution in the Newtonian approximation. This is only a limiting case of the full relativistic theory, but it is discussed first and in detail because it is a good approximation for most practical applications and is much simpler than the full relativistic theory. There is a considerable variety of methods and results in the analysis of the Newtonian limit. I have collected all those that seem to be useful and interesting.

The statistical pattern of the galaxy distribution is discussed in Chapter III. The descriptive statistics that have proved useful and are analyzed in this chapter are n-point correlation functions (analogs of the autocorre-

lation function and higher moments for a continuous function). The general approach has a long history but it is only in the last several years with the application of fast computers to the large amount of available data that the technique has been extensively developed and applied. This chapter surveys the main theoretical results and observational methods.

The n-point correlation functions have proved useful not only as descriptive statistics but also as dynamic variables in the Newtonian theory of the evolution of clustering. This is discussed in Chapter IV. The functions are generalized to mass correlation functions in position and momentum, and the BBGKY hierarchy of equations for their evolution is derived. This yields a new way to analyze the evolution of mass clustering in an expanding universe. Of course, the main interest in the approach comes from the thought that the observed galaxy correlation functions may yield useful approximations to the mass correlation functions, so the observations may provide boundary values for the dynamical theory of evolution of the mass correlation functions. The test will be whether we can find a consistent theory for the joint distributions in galaxy positions and velocities. The subject still is in a crude state because adequate redshift data do not yet exist. I present some preliminary considerations on how the analysis of the data might proceed.

The full relativistic analysis of the evolution of mass clustering is presented in Chapter V. The important application is to the behavior of the early stages of expansion of the universe when the high mean density would have made even modest density fluctuations strongly relativistic.

The last chapter describes some of the attempts to trace the links between theory and observation showing how the character of the matter distribution we observe developed out of reasonable conditions in the early universe. This is the main point of the subject, but it is not treated at length because I think there are too many options, all apparently viable but none particularly compelling. It seems likely that the game of inventing scenarios will go through several more generations before a secure picture emerges. Perhaps the best we can hope is that the final answer will draw on significant elements of the theory and observations as we now think we understand them.

I have limited the range of the discussion to length scales no smaller than the nominal size of a galaxy or else redshifts no smaller than the epoch at which mass concentrations comparable to present day galaxies appeared, thus excluding the structure and evolution of galaxies. I have excluded a few topics relevant to other areas of cosmology, such as the effect of mass clustering on the standard cosmological tests, and some obviously important subjects where I could find nothing very useful to report, such as the question of intergalactic gas clouds. I have omitted all

discussion of the possibilities offered by nonstandard cosmologies not so much because I am sure the big bang picture is the most likely candidate as that I expect it is neither reasonable nor likely to expect that people will pay much attention to these alternatives until we have a much clearer picture of what the standard model has to offer and what it must deal with.

The choice of emphasis on topics within the boundary conditions, of course, reflects a personal judgment of what is promising. Perhaps the largest omission is the primeval turbulence picture. I have described its origins and some general and well-established results but have not discussed any specific scenarios. That seems reasonable because I doubt the merits of this picture, and there are others who can serve as better and more enthusiastic advocates.

I have provided a short guide to symbols and conventions in the appendix. It probably will prove best to look this over before reading much of the main text. I have given short summaries of concepts of cosmology as they appear in the text, but have left out details available in the standard books. References to my book, *Physical Cosmology,* are indicated by the letters *PC.*

ACKNOWLEDGMENTS

I list with special thanks the people who played the most direct roles in shaping this book: Charles Alcock, Marc Davis, Bob Dicke, Jim Fry, Margaret Geller, Ed Groth, Mike Hauser, Bernard Jones, Jerry Ostriker, Bill Press, Mike Seldner, Joe Silk, Ray Soneira, Juan Uson, Simon White, Dave Wilkinson, and Jer Yu. The process would have been considerably slower and the results less satisfactory without the skill and energy of Marion Fugill.

The first concrete steps toward this book were taken while I enjoyed the hospitality of the Physics Department at the University of California at Berkeley during the 1973–74 academic year. The first draft developed as course notes at Princeton University. I am grateful to John Bahcall for providing hospitality at the Institute for Advanced Study where the final draft was written. The work was supported in part by the National Science Foundation.

The Large-Scale Structure of the Universe

I. HOMOGENEITY AND CLUSTERING

1. Homogeneity and clustering

Modern discussions of the nature of the large-scale matter distribution can be traced back to three central ideas. In 1917 Einstein argued that a closed homogeneous world model fits very well into general relativity theory and the requirements of Mach's principle. In 1926 Hubble showed that the large-scale distribution of galaxies is close to uniform with no indication of an edge or boundary. In 1927 Lemaître showed that the uniform distribution of galaxies fits very well with the pattern of galaxy redshifts. The homogeneous model, when generalized to allow for evolution, yields a linear redshift-distance relation consistent with what Hubble was finding from his estimates of galaxy distances (as summarized by Hubble in 1929).

The evolving dynamic world model quickly won attention and in the following decades, before the idea became commonplace, it generated some lively discussions. The following sections trace the development of several questions. The first question is whether the universe really is homogeneous (after averaging over a suitable clustering length). Assuming it is, must we be content to say only that this happens to be a reasonable approximation to our neighborhood at the present epoch? Could the homogeneity of the universe have been deduced ahead of time from general principles? Or might it be a useful guide to new principles? The matter distribution in any case is strongly clumped on scales of stars, galaxies, and clusters of galaxies. This clustering is a fossil of some sort, a remnant of processes in the distant past as well as an on-going phenomenon. How does the clustering evolve in an expanding universe? What is its origin? What does it tell us about the nature of the universe?

2. Is the universe homogeneous?

In 1917 the phrase "the large-scale distribution of matter" was generally taken to mean the distribution of stars in the Milky Way galaxy. For example, the title of Eddington's (1914) book on the latter subject is *Stellar Movements and the Structure of the Universe*. It was considered well-established from star counts that the stars are concentrated in a flattened roughly spheroidal distribution, the Kapteyn system (after the

astronomer mainly responsible for the laborious accumulation and analysis of the star count data). If the distribution had been homogeneous, the number of stars brighter than f would have varied as[1]

$$N(<m) \propto f^{-3/2} \propto 10^{0.6m}, \tag{2.1}$$

where m is the apparent magnitude. The star counts are different in different directions in the sky and increase with decreasing f distinctly less rapidly than would be expected from equation (2.1). The implication is that we are seeing the edge of the system.[2]

It is not clear how much Einstein in 1917 knew or was influenced by these ideas. He wrote of the distribution of stars as possibly being uniform on the average over large enough distances. He did not mention the arguments (marshaled at the time by Sanford 1917) that the spiral nebulae may well be other "island universes," other galaxies of stars, though it seems likely Einstein knew the general idea because he had discussed with de Sitter how matter might be distributed in the universe (de Sitter 1916). Einstein rejected the idea that the universe of stars might be a limited island in asymptotically flat space because a star escaping from the system would move arbitrarily far from all other matter yet preserve its inertial properties, contrary to Mach's Principle. At first he proposed that the line element might become singular outside the realm of the matter, but then hit on a much more elegant solution, a homogeneous closed world.

De Sitter was an astronomer (and a student of Kapteyn) and well aware that the stars are not uniformly distributed. He was willing to accept the island universe hypothesis and to speculate that these systems might be uniformly distributed through space (de Sitter 1917). However, he mentioned no tests of the uniformity idea.

It was known at the time that there are many more faint spiral nebulae, that is, nebulae of small angular size, than bright ones, and that there are hundreds of thousands of very faint objects that might be just like the bright ones but so far away that it is not possible to make out the spiral structure (Fath 1914, Sanford 1917, Curtis 1918). Hubble (1926) was the

[1]For stars of fixed intrinsic luminosity, those appearing brighter than f are at distances $< r \propto f^{-1/2}$, according to the inverse square law. For a homogeneous distribution the number counted would vary as the volume $\propto r^3 \propto f^{-3/2}$. The sum over stars of different intrinsic luminosities affects the constant of proportionality but not the power law behavior.

[2]It is now recognized that the counts in the direction of the Milky Way are strongly reduced by interstellar absorption, so the size of the star system was substantially underestimated. The counts in directions well away from the Milky Way are little affected by absorption, so the estimates of the thickness of the disc were quite reasonable.

first to ask whether the counts of these nebulae are consistent with the assumption that they are uniformly distributed through space. He used Seares' (1925) estimate of the limiting magnitude, $m \approx 16.7$, for Fath's counts of faint nebulae. He found that the number of these faint nebulae agrees well with what would be expected from the counts at $m < 12$ extrapolated according to equation (2.1).

The success of Hubble's test is impressive, for there are 600 times as many galaxies in the deep survey as at $m < 12$. And this stood in sharp contrast to the familiar behavior of the star counts; the indication is that the observations of stars reach the edge of the local star system while the observations of galaxies give no evidence of an edge to Sanford's "realm of the nebulae." Hubble put it this way in a later paper: "There are as yet no indications of a super-system of nebulae analogous to the system of stars. Hence for the first time, the region now observable with existing telescopes may possibly be a fair sample of the universe as a whole" (1934, p.8).

It is now known that the excellent numerical agreement Hubble found for these data is in part fortuitous because the galaxies at $m < 12$ are not a fair sample: there is a substantial excess of bright galaxies due to the local concentration in and around the Virgo cluster. Indeed Hubble clearly recognized that his result was only a preliminary indication, and over the next decade he undertook an extensive program of deep counts in selected areas (to be compared to the program of star counts except that far fewer astronomers were directly involved). A preliminary report was published in 1931, and in 1934 Hubble discussed in some detail the counts at limiting magnitudes ~ 19.1 and 19.6 The ratio of counts agrees well with the $10^{0.6\,m}$ law, as do the ratios of these counts to the number of Shapley-Ames (1932) galaxies at $m \leq 13$ (though again because of the local supercluster this latter result is in part fortuitous). In 1936 Hubble discussed counts to 5 limiting magnitudes in the range $m \sim 18.5$ to 21. The counts increase with m less rapidly than the $10^{0.6m}$ law, the discrepancy amounting to a factor 1.8 out of an observed ratio of counts of 19 over this range of magnitudes. Hubble tentatively concluded that the discrepancy is larger than would be expected in any reasonable relativistic world model and that this might indicate the relativistic theory is incorrect. The present tendency is to suppose that systematic errors in magnitude estimates and K-corrections (correction for the shift of the galaxy spectrum toward the red and out of the range of sensitivity of the photographic plate) could account for this relatively small discrepancy. Of enduring interest is Hubble's first point: to the depth of his survey there is no pronounced evidence of an edge to the realm of the nebulae.

It is surprising that the number-magnitude test (and the equivalent relation $N(>\theta) \propto \theta^{-3}$) was first applied to the counts of spiral nebulae and

faint nebulae as late as 1926. It seems reasonable to suppose that Hubble was emboldened to try the test because he had just recently shown, by the identification of Cepheid variable stars of known absolute magnitudes, that the brightest spiral nebulae are galaxies of stars comparable to the Milky Way. He might also have been influenced by Einstein's and de Sitter's discussions of homogeneous world models, for he was at least familiar with de Sitter's attempt to guess at the mean mass density (and since Hubble had a much better distance calibration he found a much better estimate of ρ).

Hubble's results from 1926 to 1934 clearly were only preliminary though encouraging indications of homogeneity, but most theorists were quick to accept the evidence. Thus Einstein in 1933 wrote, "Hubble's research has, furthermore, shown that these objects [galaxies] are distributed in space in a statistically uniform fashion, by which the schematic assumption of the theory of a uniform mean density receives experimental confirmation" (1933, p. 107). Robertson, in his influential review of the Friedman-Lemaître cosmological models, said "we accept the data, due primarily to Hubble and Shapley, on the uniform distribution of matter in the large within the visible universe, and we extrapolate them to the universe as a whole" (1933, p. 82). In 1931 Eddington made the cautionary remark, " 'Lemaître's world' is also a model in that it represents the universe as a uniform spherical distribution of matter; there is no reason why the actual shape should not be highly irregular" (1931a, p. 415). But later in the same year he stated, "We no longer look for an end to the world in its space dimensions. We have reason to believe that so far as its space dimensions are concerned the world is of spherical type." (1931b, p. 447). It is perhaps not surprising that de Sitter was more cautious. He wrote in 1931, "It should not be forgotten that all this talk about the universe involves a tremendous extrapolation, which is a very dangerous operation" (1931, p. 708). And in 1932, he wrote "These wonderful observations [of galaxies from the Mount Wilson Observatory] have enabled us to make fairly reliable estimates of the distances of these objects and to say something about their distribution in space. It appears that they are distributed approximately evenly over 'our neighborhood' " (1932, p. 114).

In the 1930s there was a somewhat indirect running debate between Hubble and Shapley over the relative importance of departures from homogeneity. Both clearly emphasized that the galaxy distribution is strongly clumped on relatively small scales. For example, Hubble (1934) noted that the frequency distribution of nebular counts N found in different telescope fields is not Poisson, as would be expected if the galaxies were randomly distributed; the general clumping makes for a

considerably broader distribution of counts. (He also made the interesting observation that the distribution of log N is remarkably close to Gaussian.) However, to Hubble the main effect clearly was the uniform distribution on large scales as revealed by very deep counts averaged over many sample fields. Shapley emphasized the great irregularities in the galaxy distribution: "the irregularities are obviously too pronounced to be attributed to chance; they are rather a demonstration of evolutionary tendencies in the metagalactic system" (Shapley 1933, p. 3). Having smaller telescopes at his disposal, Shapley and his colleagues studied the galaxy distribution at lesser depth but in greater detail across the sky. He noted that there is a considerable difference in the numbers of Shapley-Ames galaxies, $m \lesssim 13$, in the northern and southern galactic hemispheres, and he suggested that this north-south asymmetry might still amount to as much as 50 percent at $m = 17$ (Shapley 1934). The data suggested also that at $m \approx 18$ (in a magnitude system roughly consistent with that of Hubble) the galaxy density might vary across the sky by a factor ~ 2 on scales $\sim 30°$ (though there were problems with this because there were practical difficulties in transferring magnitude standards across the southern sky; Shapley 1938b). This led Shapley (1938a) to question whether the galaxy distribution really is close to uniform even when averaged over large scales and to suggest that the deviation from the $10^{0.6m}$ law in Hubble's data might be the result of large-scale density irregularities, not a failure of relativity theory. (Hubble in 1936 had mentioned but rejected the idea of large-scale irregularities.)

Shapley's remarks did not attract much attention. McCrea (1939) did point out that large-scale irregularities would raise problems for observational programs to measure the parameters in the standard cosmological models, then a subject much discussed particularly in connection with the possible role of the 200 inch telescope. Eddington (1939) and Tolman (1949), apparently independently, suggested that large-scale inhomogeneity may account for Hubble's results, and Omer (1949), at Tolman's suggestion, devised an inhomogeneous relativistic model (spherically symmetric about our position; § 87 below) that he could adjust to fit the galaxy counts. However, by the 1950s the possibility of large-scale inhomogeneity was largely displaced in the minds of cosmologists by the debate over homogeneous world models—evolving versus steady state versus Milne's kinematic cosmology—and, in the relativistic models, the possible values of parameters such as the cosmological constant, Hubble's constant and the time scale, the acceleration parameter and the open versus closed models. An example is Bondi's (1952) book on cosmology where the suggestion of Eddington and Tolman is noted but rejected as unprofitable. A second example where one finds a more cautious view is

McVittie's (1961) *Fact and Theory in Cosmology*. All the cosmological models discussed are homogeneous, but McVittie does stress the observational problems in establishing homogeneity.

Though the subject has not been very popular one can find occasional recent discussions of the question of large-scale inhomogeneity. De Vaucouleurs (1960, 1970, 1971) and van den Bergh (1961) have joined Shapley in observing that the traditional evidence from galaxy counts as functions of magnitude or position in the sky is at best slim. In 1965 Omer rediscussed his spherical model for large-scale inhomogeneity. Rees and Sciama (1968) used the spherical model in a discussion of the suggestion by Strittmatter, Faulkner, and Walmesley (1966) that clustering scales for quasi-stellar objects might be comparable to the distance to the horizon. Bonnor (1974) and Silk (1977a) also discussed this model, and Silk pointed to the indications of large-scale matter currents found by Rubin, Thonnard, Ford, and Roberts (1976) from a systematic survey of redshifts of galaxies. Kristian and Sachs (1966) explored another approach based on the assumption that all properties of the universe out to redshifts $Z \sim 0.3$, for example, can be usefully expanded in a power series about our position. Wertz (1971), Haggerty (1971), and Wesson (1976), stimulated by de Vaucouleurs, considered the possible dynamics of yet another picture, where the hierarchy of clustering continues to indefinitely large scales (§ 62 below).

De Vaucouleurs has made the interesting point that if the universe really is close to homogeneous on the scale of the horizon cH_0^{-1}, it is a remarkable break with the state of affairs on smaller scales: from subatomic particles on up we deal with objects—localized structures. De Vaucouleurs noted that this tendency to clump continues to scales at least as large as the local supercluster (the concentration of galaxies around the Virgo cluster, distance $\sim 10h^{-1}$ Mpc, of which we are an outlying part), and he could cite as indications of irregularities on still larger scales the angular gradients found in the Harvard survey (Shapley 1938a, b) and the large-scale correlation of rich clusters found by Kiang and Saslaw (1969), both effects pointing to strong clustering on scales $\sim 100\ h^{-1}$ Mpc. The indication is that if the clustering does terminate it does so perhaps suspiciously close to the largest depth of reliable observations and close to the largest possible scale consistent with the assumption that the universe is accurately uniform on the horizon ($cH_0^{-1} = 3000\ h^{-1}$ Mpc).

Direct observations of the large-scale galaxy distribution still are beset with the problem of controlling systematic errors when galaxy densities are compared over widely separated parts of the sky or at very different apparent magnitudes. Modern deep galaxy counts (Brown 1974, Kron 1978, Tyson and Jarvis 1979, Ellis 1980) are found to vary with magnitude

roughly as $10^{0.45m}$ to depths comparable to cH_0^{-1}. The deviation from the $10^{0.6m}$ law is about what is expected from the K-correction. The best sample of the distribution across the sky is the Lick catalog (Shane and Wirtanen 1950; 1967). This gives counts to limiting magnitude $m = 19$ in 10′ by 10′ cells across two-thirds of the sphere. The effective depth of the sample is $\sim$ 200 h^{-1} Mpc. There are large-scale density gradients, amounting to rms fluctuations $\delta\mathcal{N}/\mathcal{N} \approx 0.10$ in the surface density smoothed over 10°. One cause is purely local, the variation of absorption across the sky. It is difficult to decide how much might be true large-scale fluctuations in the space density of galaxies.

One convenient measure of the irregularities in the space distribution is the dimensionless autocorrelation function

$$\xi(r) = \langle \rho(\mathbf{r}_1)\, \rho\,(\mathbf{r}_1 + \mathbf{r})\rangle / \langle \rho\rangle^2 - 1, \tag{2.2}$$

where the angular brackets signify an average over the position $\mathbf{r}_1$ within the sample. An upper limit on large-scale clustering within the Lick sample is (Peebles and Hauser 1974)

$$\xi(50h^{-1}\text{Mpc}) \lesssim 0.025, \tag{2.3}$$

and $\xi(50)$ is significantly less than this if variable absorption is important. One measure of the scale on which clustering is strong is the value of the lag r_0 at which the correlation function ξ is unity. In the Lick sample (Groth and Peebles 1977, § 57 below)

$$r_0 \approx 4h^{-1}\text{Mpc}, \qquad \xi(r_0) \equiv 1. \tag{2.4}$$

Since this is small compared to the depth of the survey, $\sim 200\, h^{-1}$ Mpc, the indication is that within the Lick sample the progression of clustering observed on small scales does blend into a nearly uniform background.

Equation (2.4) describes a mean over the distribution, and one certainly can find spots in the Lick sample where the density stays higher than the mean over distances larger than r_0. Examples are provided by Abell's (1958) catalog of rich compact clusters. Galaxies, of course, tend to concentrate around Abell's cluster positions. This can be measured by averaging the galaxy space density over all shells, radius r to $r + \delta r$, centered on all Abell clusters. One finds that this mean density $n(r)$ is twice the overall average density in the Lick sample at distance (Seldner and Peebles 1977a).

$$r_a \approx 14h^{-1}\text{Mpc}. \tag{2.5}$$

For the correlation among positions of Abell cluster centers, Hauser and Peebles (1973) estimate the clustering length

$$\xi_{cc}(r_s) \equiv 1, \qquad r_s \approx 30h^{-1}\text{Mpc}. \tag{2.6}$$

Kiang and Saslaw (1969) suggested r_s is closer to 100 h^{-1} Mpc from a reconstruction of the three-dimensional distribution using Abell's estimates of apparent magnitudes of brighter cluster members. This method has the advantage that it makes the apparent clustering larger (the angular correlation function, which must be unfolded to find $\xi(r)$, is smaller than ξ because of the overlapping of objects at very different distances), and it has the disadvantage that if the errors in Abell's magnitude scale vary systematically with distance, as seems possible, it will introduce spurious radial clustering.

The indication from equations (2.4) through (2.6) is that the clustering does blend into small fluctuations, $\delta\rho/\rho < 1$, well within the sizes of available samples. Of course, the samples are limited and deeper surveys are needed.

New methods of observation have provided some very deep glimpses into space and, indirectly, precise measures of homogeneity. Extragalactic radio sources, all or a fair fraction of which are galaxies, are distributed across the sky in a remarkably uniform way; the distribution of the 5000 4C sources (flux levels $S \geq 2$ Jy) is almost indistinguishable from random (Webster 1976b, Seldner and Peebles 1978). Because the number of 4C objects is much less than the number of Lick galaxies, the radio source data do not improve our upper limits on fluctuations in the density of objects across the sky at $\theta \lesssim 10°$. For example, the mean number of 4C sources found in a 3° by 3° cell is $\langle N \rangle \approx 2$, and the rms fluctuation in the number is close to Poisson, $\delta N/N \approx 0.7$. The rms fluctuation in the number of Lick galaxies is a factor ~ 3 smaller, $\delta N/N \approx 0.25$ (compared to the expected value $\delta N/N = 0.045$ if galaxies were randomly distributed). But since many of the sources are at distances $\sim cH_0^{-1} = 3000\ h^{-1}$ Mpc, we do have a strong new test of isotropy on large scales. The contrast with the distribution of bright galaxies is worth emphasizing; if the northern hemisphere were divided into two equal parts, the number of 4C sources in each would agree to $|N_1 - N_2|/(N_1 + N_2) \simeq 0.015$ (rms), while the number of Shapley-Ames galaxies would scatter by a factor ~ 2. A second important measure is the diffuse X-ray background (Wolfe 1970, Wolfe and Burbidge 1970; Fabian 1972). Since a substantial part of the flux comes from objects at modest redshifts—active galaxies and clusters of galaxies—this measures how the projected density of matter, integrated to the horizon, varies across the sky. The present limit on fluctuations in the

projected density is $\delta f/f \lesssim 0.04$ at $\theta \sim 5°$ (Schwartz 1979, Schwartz, Murray, and Gursky 1976). Finally, the microwave background radiation is isotropic to $\delta T/T \lesssim 0.001$ on angular scales from 10′ to 180°. This does not measure the matter distribution directly because the radiation is thought to be very weakly coupled to matter in the present universe. But since the radiation temperature varies inversely as the expansion factor, or more generally as the redshift from source to observer, it does indicate the large-scale motion has been isotropic about us to an accuracy better than 1 part in 10^3.

These three sets of observations show that the matter distribution and motion are quite accurately isotropic on scales $\sim cH_0^{-1}$. This is a strong test of the standard homogeneous and isotropic world picture, but of course it is not complete because it leaves open the possibility that the universe is inhomogeneous but isotropic about a point near us. However, the galaxies at high redshift look much like the ones nearby, and in such a model an observer on any one of the enormous number of distant galaxies would find the universe is much less isotropic than we do. The more reasonable presumption is that the universe would appear isotropic on a distant galaxy, so the visible universe is accurately homogeneous.

We certainly do not have definitive evidence of homogeneity, and further developments in the tests will be followed with great interest. On the other hand, the observational situation has improved many times over since the 1920s, and the results must be counted as a spectacular success for the vision of Einstein and Hubble.

3. PHYSICAL PRINCIPLES

A. Prediction of homogeneity?

Might the homogeneity of the universe have been expected from general arguments and physical principles? In a sense the answer is yes, for Einstein did hit on a homogeneous world model as a way to satisfy some general considerations. He rejected the idea of an infinite material Newtonian universe on the grounds that the potential and hence star kinetic energies would be arbitrarily large. In the 1917 paper he gave two arguments against the idea that matter is concentrated like an island in otherwise empty asymptotically flat space. The first argument was that, given sufficient time, the system would relax, part contracting to high density (we would now say, to a black hole), part escaping with positive energy. Since Einstein supposed the global properties of the universe must be unchanging, this was unacceptable. It is not clear how seriously Einstein weighed this, for the universe could not be eternal in any case; for example, the solar system, given sufficient time (and if the sun does not explode),

would relax in the way he envisioned for the island universe as a whole, and if energy is conserved, the stars must eventually stop shining. As mentioned in the last section, he did emphasize Mach's principle: a particle escaping this island universe would move into flat space, arbitrarily far from all other matter, but yet, according to relativity theory, its inertial properties would not change, contrary to the idea that inertia is generated by the matter in the universe. A discussion reported by de Sitter (1916) gives an interesting view of at least some aspects of Einstein's thoughts. Since flat space at infinity conflicts with Mach's principle, he considered the idea that the components of g_{ij} degenerate to singular values at the edge of the universe. Since observed objects give no evidence of strong space curvature, one would have to suppose, as de Sitter put it, that the g_{ij} become singular outside "hypothetical" masses that surround the known and ordinary realm of matter. The next year Einstein found a more elegant solution: replace the singular behavior of the g_{ij} at the boundary with the condition that the universe be closed—the three-dimensional analogy of the closed two-dimensional surface of a balloon. This universe is finite, with no flat exterior, no hypothetical masses, and indeed no edge. Einstein's brilliant argument from general principles thus led to a world picture that has stood the test of time and observation.

It is worth bearing in mind, despite this success, that such arguments tend to be matters of opinion. Many people have been attracted to another picture, an unlimited clustering heirarchy (§ 62 below). A review of such models and of the history of development of the idea is given by Mandelbrot (1977). In the scale-invariant clustering model that, according to Mandelbrot, can be mainly attributed to Fournier d'Albe (1907) the hierarchy scales so that the typical value of the mass within distance R of an observer varies as $M \propto R$. The size and mass of the universe are arbitrarily large, but the mean density M/R^3 converges to zero and the mean mass per unit area in the sky of an observer converges, so even if stars shine forever the surface brightness of the sky is not large and Olbers' paradox is avoided. The hierarchy is arranged so the virial velocity in clusters of size R is $v^2 \propto M/R$, independent of R. Thus although the mass is infinite, the peculiar velocities need not be high. This anticipated and countered one of Einstein's arguments against an infinite quasi-static universe. Also, it has been found that Fournier d'Albe's model gives a remarkably good approximation to the statistics of galaxy clustering on small scales. This is discussed in Chapter III below. Einstein (1922) felt that the hierarchical world picture (as rediscussed by Charlier 1908, 1922) was compatible with general relativity theory but not with his interpretation of Mach's principle. As discussed in the last section, the evidence from recent observations is that the hierarchical model in fact fails, and

Einstein's picture is a reasonable first approximation on scales larger than about 10 h^{-1}Mpc.

B. The cosmological principle

Milne (1933a) was the first to notice that although Hubble's law (recession velocity proportional to distance) was derived from the relativistic world model, it cannot be considered a very specific test of the model because it is the only functional form allowed by homogeneity and isotropy. (A particularly clear explanation is given by Milne 1934.) He proposed that this result might be extended and that one might be able to derive cosmology more or less complete by following such arguments from powerful general principles.

Milne referred to the homogeneity assumption by phrases such as "the extended principle of relativity" and "Einstein's cosmological principle," and soon fixed on the "Cosmological Principle." He clearly felt that this principle has a considerable *a priori* philosophical merit, perhaps even that it was logically necessary for what one means by the universe.[3] His program did not meet with much approval, though it was an important forerunner of the steady state model. His term, "the Cosmological Principle," was quickly taken up as an easy way to state and justify a central assumption. For example, it appears in the introductory comments in papers by Robertson (1935), Walker (1936), and, in a less positive way, de Sitter (1934). The cosmological principle is now firmly lodged in the lore of the subject.

The most interesting immediate reaction to Milne's ideas was that of Dingle (1933a,b). He and others objected to the idea that the cosmological principle is to be compared to a law of nature: "a principle coequal with the principle of relativity should be capable of universal application," at least, as he noted, within some substantial domain of phenomena (1933a, p. 173). Homogeneity could only apply in the average over many galaxies. Dingle pointed out that Einstein's field equations do admit strongly inhomogeneous solutions, and he took it to be "perfectly conceivable that an increase of telescopic power may reveal a variation of material density with distance." Dingle noted that the evidence of isotropy of the galaxy redshifts is far from complete because of the absence of observations in the Southern Hemisphere. Curiously, he did not mention the galaxy counts, though, since he had been in Pasadena, he should have been in a position to learn the status of Hubble's program.

[3]Milne 1933b, p. 185. In his book (Milne 1935), at least partly in reaction to Dingle's comments, Milne was careful to state the cosmological principle as an assumption or axiom, though he did argue it is necessary for an intelligible universe.

C. The instability of the universe

If the symmetry of the universe is not enforced by a principle then one might ask whether it always has been or will be as close to homogeneous as it is now. The history of ideas may be traced back to Einstein's 1917 paper.

Einstein had assumed as a matter of course that the universe is static. But the field equations of general relativity as originally formulated then indicate the pressure would have to be negative, $p = -\rho c^2/3$ (so the active gravitational mass density associated with pressure cancels that of ρ). To avoid this he modified the gravitational field equations, introducing a universal cosmic repulsion that varies in proportion to separation with strength determined by the cosmological constant Λ. With $|p| \ll \rho c^2$, the static model then requires (§ 97)

$$4\pi G\rho = \Lambda. \tag{3.1}$$

Nearly twenty years passed before it was clearly stated by Tolman that there is a serious problem with this—the model is unstable. The general point was sensed earlier by Weyl (1922) and Eddington (1924), who observed that a physical variable the density ρ is set equal to a constant of nature Λ. What happens if the matter is rearranged or if some of it is annihilated in stars thus changing $\langle \rho \rangle$?

In 1930 Eddington learned of Lemaître's (1927) work on evolving world models and recognized that it gives a partial answer. If an Einstein model were somehow perturbed so that the mean density is slightly less than $\Lambda/4\pi G$, the universe would expand, the density drop, and the expansion steadily accelerate. If the Einstein universe were perturbed so that ρ is slightly higher than $\Lambda/4\pi G$, the universe would collapse.

Lemaître and Eddington at first assumed the universe is expanding away from the Einstein model, and Eddington (1930) proposed that the balance of the initial Einstein world was broken, the expansion initiated, through the perturbation caused by the formation of galaxies "by ordinary gravitational instability." In the following several years there was rather extensive discussion of this, mainly by McCrea and McVittie (1931, and earlier references therein), who tried to decide whether this condensation into galaxies would inevitably produce general expansion rather than contraction. However, the topic soon went out of style as attention turned to models that do not trace back to the Einstein case.

McCrea and McVittie approximated a condensation as a distribution spherically symmetric about one point. This is a very convenient model because it permits a description of at least the rough outlines of a mass concentration like a galaxy while keeping the mathematics simple.

Lemaître (1931, 1933a,b) found the ultimate simplification: if the pressure can be neglected, the motion of each mass shell is the same as in some homogeneous world model. (Of course when mass shells cross, the motion of each follows an altered cosmological model.) Discussion of how a density irregularity might evolve thus is made simple: each mass shell goes its separate way.[4]

Lemaître pointed out that this result, which might at first sight seem remarkable, is in fact "obvious" at least for small-scale condensations because the general relativity description of a small region is equivalent to the weak-field limit, the Newtonian description.[5] The argument, in a fuller form than Lemaître gave, develops as follows (Callan, Dicke, and Peebles 1965). Suppose the mass $M(< r)$ within the shell of physical radius r centered on the condensation satisfies $GM(< r)/rc^2 \ll 1$, and suppose this mass inside the shell is temporarily removed. Then an earlier discussion by Lemaître (1931) describes the space inside the hollow: according to Birkhoff's (1923, p. 253) theorem, which generalizes Newton's iron sphere theorem, space must be flat, unaffected by the matter outside. The mass $M(<r)$ can be placed in this flat space and treated in the Newtonian approximation, so the iron sphere theorem applied once again indicates the acceleration of the surface of the sphere is the same as if $M(<r)$ were uniformly distributed within r. Thus the motion of the shell must agree with that of some zero pressure homogeneous world model. (A more general discussion of Newtonian gravity physics in relativistic cosmology is given in Sections 6 and 84 below.)

Tolman (1934a) discussed some interesting consequences of Lemaître's solution.[6] Einstein's static world model evidently suffers from an instability more general than that noted by Lemaître and Eddington, for if in the originally static case some matter were carried from one spot to another, the more dense spot would collapse, the less dense spot expand, and the universe would grow strongly irregular. For a generally expanding universe, since different mass shells can evolve independently, Tolman observed that there clearly is no "general kind of gravitational action which would necessarily lead to the disappearance of inhomogeneities in cosmological models" (1934a, p. 175). As Dingle also remarked, there

[4]Though the point is simple, it was by no means self-evident, as is illustrated by the lengthy computations by McVittie (1932), Dingle (1933*b*), and Sen (1934), all of whom assumed spherical symmetry but did not hit on Lemaître's trick.

[5]The agreement between Newtonian and relativistic descriptions of a spherically symmetric condensation was independently noted by McCrea and Milne (1934), but they offered no explanation of why it should be.

[6]Though Tolman may well have hit on Lemaître's result independently, he refers to Lemaître's prior discovery. Thus I find it curious that this often is called the Bondi (1947)-Tolman solution. Another standard reference is Einstein and Straus (1945).

appears to be nothing in Einstein's gravitational field equations that would guarantee that different parts of the universe must expand at the same rate or even that all parts of the universe, as observed by us, must be expanding, not contracting. Tolman noted that this conclusion might be modified by the effects of "more drastic kinds of inhomogeneities" than those spherically symmetric about one point and that nongravitational forces might promote homogeneity. He concluded that, pending the possible discovery of such effects, we should be cautious about extrapolating the observed behavior of "our neighborhood" to great distances in space or to the remote past or the distant future.[7] Similar cautionary remarks are expressed in his book and are contrasted with the views of Milne, "who would regard the homogeneity of the universe as a fundamental principle" (Tolman 1934b, p. 364).

The implications of Tolman's remark are worth emphasizing. For example, it appears to be conceivable that the part of the universe we see in the Northern Hemisphere could have been slightly too dense overall at the time of the big bang, so that it expands for 10^{10} years and then collapses, while the part we see in the Southern Hemisphere had slightly low density overall and so expands indefinitely. According to our present understanding of physical principles, this is a possible universe but one that would look markedly unlike what we observe.

How would a strong initial irregularity behave? A model that is easy to understand goes as follows. Suppose that at some very early time t_i the universe is everywhere homogeneous and isotropic, with uniform density and expansion rate, except within a spherical patch of radius $r_i > ct_i$. Toward the center of this patch the density is high and space is strongly curved, so unless conditions are specially adjusted space soon collapses to a singularity. How does space outside the patch behave? Birkhoff's theorem tells us there is no gravitational signal of what happens inside, and if $r_i > ct_i$, there is no pressure signal, so the exterior is unaffected, evolves as a homogeneous model (§ 87). If at the present epoch this patch came within the horizon, we would see a mass concentration, perhaps surrounded by an empty region (though the hole could be filled by interactions with neighboring irregularities). There is no observational problem with low mass black holes, but we can only account for the absence of black holes of extreme mass, like the absence of large density fluctuations on very large scales, by presuming that they were excluded by the initial conditions.

An important aspect of the puzzle is that in the model a light cone

[7]A tendency now is to distinguish extrapolations backward and forward in time (Peebles 1967*a*, 1972); in the spherical model it certainly can be arranged that the universe starts out highly irregular and grows homogeneous by adjusting the starting times for the expansion of each mass shell, but that requires very particular adjustment and so seems contrived.

traced back to the singularity encompasses only a limited part of the matter in the universe. The visible part increases with time, reaching zero as $t \to 0$.[8] A distant galaxy at high redshift is near our horizon and in the past would not have been "visible" from our position (unless there is some way to trace light rays back through the singularity). If causal connection is reckoned from the time of the big bang, galaxies at high redshift have not previously been in contact with us, and they have not been in contact with galaxies in other parts of the sky. How then do we account for the familiar appearance of the galaxies at high redshift? How do we account for the remarkable uniformity of the microwave radiation coming from parts of the universe that have not been in communication since the time of the big bang?

De Sitter (1917) noted the horizon effect in his cosmological solution. Tolman (1934b) derived the effect for Friedmann-Lemaître models but only briefly pointed to the conceptual problems it raises. Milne (1935, §§463–474) discussed it at length, mainly as an argument against relativistic cosmology. There was a revival of interest in connection with steady state cosmology and a new analysis by Rindler (1956). Misner (1967, 1968) and McCrea (1968) emphasized the importance of the causality puzzle, and Misner proposed an ingenious solution: the horizon would be broken if the early universe were not at all like the model, but chaotic. Perhaps, as Tolman had noted, "more drastic" irregularities could promote homogeneity. This idea, and the homogeneous anisotropic mixmaster model by which Misner illustrated it, has been an important stimulus, but one that must be applied with caution because the lesson of the spherical model certainly is that gravity promotes inhomogeneity, not homogeneity. If the early universe were chaotic how did it avoid becoming a tangle of black holes?

Tolman's 1934 discussion now seems reasonable, and the reaction to Milne's ideas stimulated some questions that seem interesting. Must the universe be homogeneous? Is it always that way? As it happened these questions attracted little notice; attention concentrated on the homogeneous models, their relative merits, and possible tests. The cosmological principle (and perfect cosmological principle) served a useful function in keeping the discussion focused on some well-defined and useful research problems. On the other hand, in elevating homogeneity to a principle

[8]The depth of the visible universe is $\sim ct \sim cH^{-1}$, and the number of baryons in the visible universe is $N \sim n(t)\ (ct)^3$. Since $n \propto a(t)^{-3}$ and $a \propto t^{2/3}$ in an Einstein-de Sitter model, $N \propto t$. The horizon, of course, also limits the propagation of a pressure wave. That is why in the preceeding discussion it was possible to ignore the radiation pressure gradient in the early universe.

people did tend to lose sight of the important observational and theoretical problems behind it.

The present situation is curious. Einstein did predict the large-scale homogeneity of the universe, and the observational developments mainly have agreed with the prediction. Einstein's idea was codified in the cosmological principle and has played a central role in cosmology. But it seems that with present dynamic theory we cannot account for this homogeneity; we cannot say whether Einstein's argument was a lucky guess or a deep insight into the way the universe must be. For now, it appears, we must accept it as an initial condition or, with Milne, as something to be added to the principles of physics.

4. How did galaxies and clusters of galaxies form?

A. The role of gravity

Lemaître (1933a,b, 1934) pointed out that if the evolving homogeneous and isotropic world model is a reasonable first approximation, then the next step is to account for the departures from homogeneity in structures like galaxies and clusters of galaxies. Like Jeans (1928) he supposed that in the remote past matter was uniformly spread through the universe and that gravitational instability caused the distribution to fragment into separate nebulae. Jeans had proposed that the size of each fragment would be comparable to the critical Jeans length (the minimum length at which the self gravitation of a developing irregularity exceeds the opposing pressure gradient). He had assumed as a matter of course that the density of the universe is independent of time, and he cited the opinion expressed by Newton on the point:

> It seems to me, that if the matter of our sun and planets, and all the matter of the universe, were evenly scattered throughout all the heavens, and every particle had an innate gravity towards all the rest, and the whole space throughout which this matter was scattered, was finite, the matter on the outside of this space would by its gravity tend towards all the matter on the inside, and by consequence fall down into the middle of the whole space, and there compose one great spherical mass. But if the matter were evenly disposed throughout an infinite space, it could never convene into one mass; but some of it would convene into one mass and some into another, so as to make an infinite number of great masses, scattered great distances from one to another throughout all that infinite space. And thus might the sun and fixed stars be formed, supposing the matter were of a lucid nature.

Einstein (1917) with many others felt that Jeans' static uniform Newtonian background model is not self-consistent. Lemaître had a more definite theoretical basis in the relativistic world models. He originally supposed that the expansion of the universe can be traced back to the static Einstein model in the distant past but soon turned to the Lemaître model (1933a) where the universe is assumed to expand from a dense initial state, decelerate until gravity and cosmic repulsion nearly balance, remain in this quasi-static phase for some length of time, and then resume expansion with Λ dominating ρ. Apparently one reason he liked the model is that it gives a preferred epoch to the formation of structures.[9] He supposed that in the dense early stage there were small irregularities in the matter distribution. In a patch where the density (evaluated when the local expansion rate has some chosen value) is slightly higher than average the matter may dwell in the quasi-static phase for a longer time and where the initial density is high enough the patch may collapse rather than resume the general expansion. Such a collapsing patch would end up as a galaxy. In larger volumes containing many protogalaxies, the initial density contrast must be smaller, and there are spots where the contrast is just such that the patch stays in the quasi-equilibrium phase for a very long time. He identified these patches with clusters of galaxies. The equilibrium between gravitational attraction and cosmic repulsion gives $\rho \sim (4\pi G)^{-1}\,\Lambda$ in such patches (eq. 3.1). This is the predicted density within clusters, the minimum density in a stable system. Lemaître (1934) argued that with current estimates of H and Λ the predicted density gave a reasonable fit to the typical density within a cluster.

Lemaître's approach was phenomenological; he asked whether small initial fluctuations could develop into irregularities that match in some detail what is observed, and he left for some deeper theory the origin of the initial fluctuations. The problem of accounting for the origins of galaxies and clusters of galaxies certainly is a worthy one, and the general approach Lemaître formulated now seems fairly useful: it is the subject of this book. But it is curious to note how little his ideas were discussed during the 1930s and how little they influenced the developments in the next several decades. One reason was his tendency to stick with the Lemaître universe while others were considering other models and many were arguing that Λ ought to be dropped. Another certainly was the excitement of the gathering storm over homogeneous models.

The next important development was Lifshitz's (1946) general analysis

[9]Another reason (Lemaître 1933b) was that the time since zero radius could be made larger than H^{-1}, thus relieving the time-scale problem resulting, as we now know, from an overestimate of Hubble's constant.

of linear perturbations in a Friedmann-Lemaître model. Unfortunately because he did not examine the details of joining the limiting behavior at high redshift where he assumed the relativistic equation of state $p = \rho c^2/3$ to the solution for $p \ll \rho c^2$ at low redshift, he decided that "we can apparently conclude that gravitational instability is not the source of condensation of matter into separate nebulae" (1946, p. 116). Novikov (1964a) was the first to point out that this is not quite right.

One can see the origin and resolution of the problem by the following heuristic argument. Consider an expanding model $\Lambda = 0$ cosmologically flat or close to it. The characteristic time for collapse or expansion is then

$$t \sim (G\rho)^{-1/2}. \tag{4.1}$$

If the velocity of sound in the matter (that we shall imagine behaves like a perfect fluid) is c_s, the critical Jeans length is (§ 16)

$$\lambda_J \sim c_s t. \tag{4.2}$$

If the density fluctuation occupies a patch smaller than λ_J, the acoustic response time r/c_s is shorter than the collapse time t, so the fluctuation oscillates like an acoustic wave. If $r > \lambda_J$, gravity dominates and the fluctuation can grow more prominent. Note in particular that if the universe is radiation dominated, $p = \rho c^2/3$, the velocity of sound is $c/3^{1/2}$, and the Jeans length is comparable to the horizon,

$$\lambda_J \sim ct, \qquad p \sim \rho c^2. \tag{4.3}$$

Consider now a patch with contrast $\delta\rho/\rho = \delta(t)$ and physical size $r(t)$. If $p = \rho c^2/3$ and $r \gg ct$, then in linear perturbation theory one finds that the contrast grows as $\delta \propto t$ (§ 86) and r closely follows the general expansion, $r(t \propto a(t) \propto t^{1/2}$. If $p = 0, \delta \propto t^{2/3}$ with $r \propto a(t) \propto t^{2/3}$ (§ 11). In either case the potential energy per unit mass associated with the fluctuation is

$$\phi c^2 \sim G\delta M/r \sim G\rho\, \delta\, r^2 \sim \delta(r/ct)^2 c^2. \tag{4.4}$$

Using the results quoted above, one sees ϕ is independent of time. The perturbation to the geometry due to the density fluctuation is on the order of the dimensionless potential ϕ so, if linear perturbation theory is to be valid, ϕ must be much less than unity. But then equation (4.4) indicates that, when $r = ct$, $\delta \ll 1$. That is, when the fluctuation appears on the horizon, the contrast δ must be small. After this epoch, if $p = \rho c^2/3$, $r < \lambda_J \sim ct$, and so δ is forced to oscillate like an acoustic wave: it cannot

grow large. However, Novikov pointed out that if $p \to 0$ while r still is larger than ct, then δ can continue to grow after it appears on the horizon and can finally develop into a stable system ($\delta \gtrsim 1$). When this happens, the object has energy $E \sim -\phi c^2$. Since we require $\phi \ll 1$, the object is nonrelativistic, which, of course, is what is wanted.

Lemaître's spherical solution gives another useful way to think of the behavior of the perturbation. Suppose pressure gradients may be neglected. Then Lemaître (1933a,b) showed that the perturbed patch behaves like a section of a homogeneous world model. If $\Lambda = 0$, the cosmological equation is

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8}{3}\pi G\rho - \left(\frac{c}{aR}\right)^2, \tag{4.5}$$

where the curvature of space in proper units is

$$R_p = Ra(t), \tag{4.6}$$

with R a constant. Suppose R^{-2} is positive, so the patch has negative energy. The expansion parameter can be chosen to agree with the proper radius of the perturbed patch,

$$r(t) = a(t). \tag{4.7}$$

The ratio of the size of the patch to the space curvature in the patch is then

$$r(t)/R_p(t) = R^{-1}. \tag{4.8}$$

If this number is small, the curvature within the patch can be likened to a wrinkle in the background geometry; if $R^{-1} \sim 1$, it can be likened to a knob (fig. 87.1). One notices that R^{-1} is independent of time so the perturbation to the geometry does not change: a wrinkle stays a wrinkle (as long as this simple spherical model applies). This corresponds to the result ϕ = constant (eq. 4.4) in linear perturbation theory.

When the patch stops expanding, $da/dt = 0$, equations (4.5) and (4.7) indicate the radius is

$$r_m \sim c(G\rho)^{-1/2}R^{-1} \sim ct_m R^{-1}. \tag{4.9}$$

This is a relation between the time t_m when the protoobject breaks away from the general expansion (eq. 4.1), the radius r_m of the object when this

happens, and the parameter R^{-1} that measures the perturbation to the geometry. For galaxies and clusters of galaxies $r_m \ll ct_m$, so $R^{-1} \ll 1$, and these objects would form out of wrinkles in the geometry (eq. 4.8). This corresponds to the condition $\phi \ll 1$.

The conclusion from this discussion is that, as Lemaître showed, one can think of small density fluctuations in the early universe growing into prominent irregularities like galaxies. However, the consequence that was only fully recognized later, is that in this picture one must accept the idea that the universe had primeval wrinkles (Novikov 1964a, Peebles 1967a).

Another aspect of the linear perturbation theory result was noted by Lifshitz (1946) and by Bonnor (1956, 1957, 1967). The density contrast $\delta\rho/\rho$ in an Einstein-de Sitter model grows as $t^{2/3}$, much less strongly than the exponential growth one usually associates with an instability. Bonnor noted as an example that if one starts the calculation at $t_i = 1$ sec, then one finds $\delta\rho/\rho$ grows by the factor $(t_0/t_i)^{2/3} \sim 10^{12}$. If at t_i the matter were hydrogen atoms distributed uniformly at random, the density fluctuations on the scale of a large galaxy (10^{11} $M_\odot \sim 10^{68}$ atoms) would be $\delta\rho/\rho_i \sim N^{-1/2} \sim 10^{-34}$. The growth factor thus is inadequate by many orders of magnitude.

Though all the steps in Bonnor's calculation are valid, one can revise the conclusion. The choice $t_i = 1$ sec is an impressively small value, but we must nevertheless suppose that the universe did not begin then, that it and the density irregularities had a still earlier history. If $\delta\rho/\rho$ at the chosen hypersurface $t_i = 1$ sec happened to agree with the thermal fluctuation value $N^{-1/2}$, then the fluctuations traced back to $t_i \ll 1$ sec would have to have been much smaller or much larger than this, depending on the initial velocities. Either case would be puzzling. The other side of this is that, if $(\delta\rho/\rho)_i$ were given, one could always choose t_i small enough to secure the wanted amplification to fit present fluctuations (Zel'dovich 1965a, Peebles 1968, Nariai and Tomita 1971). But finally there is no known reason to assume $(\delta\rho/\rho)_i \sim N^{-1/2}$ at any chosen t_i. Though the relaxation time may be very short, the maximum distance over which particles or energy can be shared is limited by the horizon, which at $t_i = 1$ sec contains only about the number of baryons in the sun. In sum, because we do not know how initial conditions were set up across the horizon at the time of the big bang, we do not know the growth factor in the gravitational instability picture; we cannot say whether in this picture galaxies could have formed.

In the 1950s and early 1960s cosmologists generally tended to accept the conclusions of Lifshitz and Bonnor. Perhaps most important was the effect on Gamow's thoughts. He had earlier adopted the instability picture and, with Teller (1939), had given a heuristic analysis of the effect. Gamow was very excited to learn of Lifshitz's work (according to the recollection of

J. A. Wheeler) and quickly accepted it. Apparently this was reenforced by the results of his own calculation with S. Ulam and N. Metropolis (reported by Gamow 1952 but not published). He then turned to primeval turbulence (§ 4D below).

The instability picture certainly was not abandoned during the 1950s. For example, Hoyle (1949b) used it as an argument against the big bang model: one might have thought we ought to have seen dense patches left over from very early epochs because, as he argued, the expanding universe is unstable. Raychaudhuri (1952) used the spherical model to argue that one can find a middle ground between the conclusions of Hoyle and Lifshitz.

A good illustration of the rather confused state of affairs is the discussion at the 1958 Solvay Conference on the Structure and Evolution of the Universe. In a report on the theoretical situation in cosmology, Adams, Mjolsness, and Wheeler (1958) accepted Lifshitz's conclusion and proposed that condensations like galaxies form "during the stage of contraction towards the end of the previous oscillation" of the universe. Lemaître (1958) described his ideas on cluster formation (which now included the thought that there is an ongoing exchange of galaxies entering and leaving clusters, a concept that since has not seemed promising). He mentioned that Bonnor had worked on this subject, but made no comments on the objections he and Lifshitz had raised. Hoyle (1958, p. 61) suggested the instability picture is not very promising:

> The formation of galaxies presents a curious problem, for the universe combines both expansion and condensation. This apparent contradiction is overcome in Lemaître's cosmology by arranging for the formation of galaxies to have occurred at an epoch when the universe was quasi-stationary. No such provision is made in other forms of relativistic cosmology, the origin of the galaxies being by-passed with the rather vague hypothesis that islands of higher density were present within the expanding cosmological material. At a certain stage these islands are supposed to have resisted the general expansion and to have condensed into stars. How and why this condensation took place is left in an equally vague condition.

He suggested that in the steady state model galaxies could form by thermal instability: where the density happens to be high the cooling time is low, so the pressure drops and the pressure gradient tends to push more material in to enhance the irregularity. Oort (1958) was largely unaware of all the debate on gravitational instability. In his report he considered reasonable processes for the formation of a system like a spiral or elliptical galaxy or a cluster of galaxies in an expanding universe. He did not use the

jargon but he arrived at the conclusion that systems like the Virgo cluster might form by the gravitational instability process, while spiral galaxies require in addition something like primeval turbulence to account for their angular momenta.

Oort's remarks on cluster formation were taken up by van Albada (1960) who considered the evolution of the single-galaxy distribution function $F(\mathbf{r}, \mathbf{v}, t)$ in the self-consistent spherically symmetric potential well $\phi(r, t)$. He was able to find numerical solutions that commence as nearly uniform, expanding with the general expansion, and end up roughly reproducing the density run in a cluster as well as the observed tendency of the line of sight velocity dispersion to decrease with increasing projected radius. It is interesting to see, in the proceedings of the 1961 Conference on *Problems of Extragalactic Research,* the rather vigorous objections to van Albada's approach because of the slow rate of growth of irregularities in an expanding universe (van Albada 1962, p. 427). His response[10] was that in the solutions the galaxy distribution nevertheless does vary from nearly uniform at the initial time to strongly clustered at the final time and that the final state does bear some resemblance to a real cluster. This recalls Lemaître's original project to discover whether there is a self-consistent scenario of evolution that matches what is observed. As has been described here, people have objected in effect that if the gravitational instability picture were valid, it ought to be capable of giving an *ab initio* theory of galaxies, and that is not so within present fundamental theory. But we are left with the phenomenological approach.

B. Clustering without preferred quantities

Gravity physics with $\Lambda = 0$ does not involve any fundamental quantities of length, time, or mass, and the coupling constant G affords only variants of the one dimensionless relation as GM/rc^2 or $G\rho t^2$. The Einstein-de Sitter model ($\Lambda = p = R^{-2} = 0$) does not offer any fixed quantities either. It is not suprising therefore that in this model the dimensionless density contrast $\delta\rho/\rho$ varies as a power of time while preserving whatever initial spatial shape was given (in a pure mode: § 11), for this is the only possible functional form: the relation $\delta\rho/\rho \propto \exp t/\tau$, which often has been cited as what is wanted, is not possible because it requires the quantity τ that does not exist in the theory. It follows that in this model we cannot hope to predict the masses of systems that break away from the general expansion or when this happens. Though this point has not often been explicitly

[10]In his 1960 paper van Albada argued that the conclusion of Lifshitz and of Adams, Mjolsness, and Wheeler was mistaken. It is not clear, however, whether he considered the important role radiation pressure played in these earlier analyses.

discussed, it must have been apparent to many, to judge from the many schemes that have been proposed to introduce fixed characteristic quantities. And it must be counted as one of the reasons people have considered the instability picture unsatisfactory, as one sees, for example, in Hoyle's comments quoted above and in the detailed discussion by Harrison (1967a,b).

There are two ways to proceed. First one can consider how gravitation might be augmented by other effects, like fluid pressure, that in combination with gravity yield characteristic quantities like the Jeans length. That is reviewed in part (C) below. Second one can argue, as a virtue out of necessity, that the search for characteristic quantities is only a part of the problem and perhaps not even central to it. If we had derived the length $\sim$10 kpc from the fundamental theory to account for the nominal sizes of large galaxies, we would still have to account for tight groups of galaxies at perhaps 100 h^{-1} kpc diameter, for the dense parts of rich clusters at $\sim 1\ h^{-1}$ Mpc, and for the pattern of clustering that extends beyond that to at least 40 h^{-1} Mpc. If the theory had predicted an exponential growth of $\delta\rho/\rho$ with time, then we would have had one characteristic time, but again the problem seems richer than that. Large galaxies generally are old: though there may be some young galaxies, the era of galaxy formation seems pretty well over. On the other hand, the density contrast in a supercluster of Abell clusters is not very large; so if the universe really is expanding and evolving, these systems could only have broken away from the general expansion quite recently.

Characteristic quantities certainly are important: galaxies appear as definite objects with definite properties to account for. However, it may be that continuity of phenomena is the more fundamental clue. Though a galaxy is very different from a supercluster of galaxies, the two can be considered extremes of a continuous range of objects. Just as one can trace a continuous progression from gas to liquid phase, one can find examples of galaxies with double or multiple nuclei, compact pairs of galaxies, compact groups, looser and richer galaxy associations, and so on through a more or less continuous spectrum.

Carpenter (1938) noted that in the scatter plot of radii and mean densities within clusters of galaxies there is a rather well-defined upper envelope representing the densest clusters found for each size of the form

$$n(r) \propto r^{-\gamma}. \tag{4.10}$$

This led Carpenter to speculate "that there is no basic and essential distinction between the large, rich clusters and the small, loose groups. Rather, the objects commonly recognized as physical clusterings are

merely extremes of a nonuniform though not random distribution which is limited by density as well as by population. From this point of view, the term 'supergalaxy' is of questionable propriety, since it implies a distinctive and coherent organic structure inherently of a higher order than individual galaxies themselves" (1938, p. 355). De Vaucouleurs (1960, 1970, 1971) reexamined Carpenter's relation, adjusting the power law index to

$$\gamma = 1.7. \tag{4.11}$$

He remarked that the typical radii and densities of galaxies fit onto this relation, and he left as an open question whether there is a natural division or gap in the spectrum of clustering between a galaxy and a compact group or between a compact group and a rich cluster and so on. We are presented with a series of physically significant lengths if there is and with the continuity of the clustering phenomena if there is not.

Kiang (1967) arrived at the concept of continuous clustering from the attempts to model the distributions of galaxies and of rich compact Abell clusters of galaxies. Kiang estimated the autocorrelation among counts of Abell clusters counted in a mesh of cells across the sky, and he compared the results to a model of Neyman and Scott in which the clusters are in Gaussian-shaped superclusters, the superclusters being distributed uniformly at random. Kiang found that the best value of the supercluster radius (width of the Gaussian) varies with the lag angle θ of the correlation function at which the model is fitted to the data: the Gaussian supercluster model does not reproduce the shape of the cluster autocorrelation function. Earlier Neyman, Scott, and Shane (1956) had found the same problem in fitting this Gaussian model of galaxy clustering to the galaxy autocorrelation function in the Lick sample. They suggested that one may have to account for the clustering of clusters of galaxies, and they noted also that if clusters in the model overlap appreciably, the concept of an individual cluster may be only a convenient but oversimplified construct. Kiang was more direct: if there is no best value for the standard deviation σ in the Gaussian clustering model, then perhaps one should consider "the hypothesis, that clustering of galaxies occurs on all scales," with "no preferred sizes" (1967, p. 17).

The same point was made by Totsuji and Kihara (1969) who noted that the galaxy correlation function $\xi(r)$ (eq. 2.2) found by Neyman, Scott, and Shane for the Lick data approximates a power law at $10' \lesssim \theta \lesssim 3°$. They checked these results with their own estimates of the correlation function at 1° to 3.° They remarked that if the angular correlation function is close to a power law, then it is not very convenient to use Gaussian or

exponential functions that have characteristic lengths to model the cluster shapes or to fit to the spatial autocorrelation function, as earlier workers had done (Neyman, Scott, and Shane 1956, Limber 1954, Rubin 1954). Totsuji and Kihara's fit to the power law model is

$$\xi \propto r^{-\gamma}, \qquad \gamma \approx 1.8. \tag{4.12}$$

This expression was independently discovered (Peebles 1974a,b) in the Zwicky catalog of galaxies (Zwicky et al., 1961–68). Like Carpenter's law (eq. 4.10), it certainly agrees with the idea that there is no preferred scale over a substantial range in the clustering.

The autocorrelation function is a useful measure of the nature of the galaxy distribution, but of course it contains only very limited information, so there is not a unique interpretation of a given $\xi(r)$. One systematic way to add more detailed information is to examine progressively higher order correlation functions. As will be described in Chapters III and IV, this approach proves convenient both for the reduction of the data and the theoretical analysis of clustering dynamics. The galaxy three-point function is known in some detail, and we have schematic estimates of the four-point function. The results (§ 61) are in good agreement with Fournier d'Albe's (1907) picture of a scale-invariant clustering hierarchy (§ 3a): when the distribution is viewed with resolution r, the mass appears in patches of size $\sim r$, typical density

$$n(r) \propto r^{-\gamma}, \qquad \gamma = 1.8. \tag{4.13}$$

This applies on scales as large as $\sim 10\ h^{-1}$ Mpc and down to and perhaps including that of an individual galaxy. At $r \gtrsim 10\ h^{-1}$ Mpc the indication is that the clustering pattern is starting to wash out into a uniform background (§ 2).

Carpenter's power law expression in equation (4.10) agrees with equation (4.13), but we must consider that this agreement is at least in part fortuitous because Carpenter had in mind separate and distinct clusters, not a clustering hierarchy. Carpenter's relation as adapted by de Vaucouleurs does describe a clustering hierarchy, and it is notable that the values of the index γ found by de Vaucouleurs (equation 4.11) and established from the correlation functions agree very well.

If the continuous clustering hierarchy picture is a valid first approximation, attempts to find theories of origin of specific objects may have been addressing the wrong question. Partly because of the continuity of the galaxy clustering, more importantly because of the scale invariance of the theory, there have been a number of discussions of possible theoretical