1
Introduction

1.1 The issue

The issue we plan to address in this book, that of the average density of matter in the universe, has been a central question in cosmology since the development of the first mathematical cosmological models. As cosmology has developed into a quantitative science, the importance of this issue has not dimininished and it is still one of the central questions in modern cosmology.

Why is this so? As our discussion unfolds, the reason for this importance should become clear, but we can outline three essential reasons right at the beginning. First, the density of matter in the universe determines the *geometry* of space, through Einstein's equations of general relativity. More specifically, it determines the curvature of the spatial sections†: flat, elliptic or hyperbolic. The geometrical properties of space sections are a fundamental aspect of the structure of the universe, but also have profound implications for the space-time curvature and hence for the interpretation of observations of distant astronomical objects. Second, the amount of matter in the universe determines the rate at which the expansion of the universe is decelerated by the gravitational attraction of its contents, and thus its future state: whether it will expand forever or collapse into a future hot big crunch. Both the present rate of expansion and the effect of deceleration also need to be taken into account when estimating the *age* of the universe. The importance of stellar and galactic ages as potential tests of cosmological theories therefore hinges on reliable estimates on the

† One can think of the spatial sections as being like frames of a movie, their continuing succession building up a space-time (Ellis & Williams 1988). The difference is that these frames are intrinsically three-dimensional (and may be curved!). Cosmology provides us with a natural choice for the time coordinate: cosmological proper time, defined below.

mean matter density. Indeed the very viability of our models is threatened if we cannot attain consistency here. Finally, we would also like to know precisely *what the universe is made of*. There is compelling astrophysical and cosmological evidence for the existence of non-baryonic dark matter, i.e. matter which is not in the form of the material with which we are most familiar: protons and neutrons. Indeed, as we shall see, even a conservative interpretation of the data suggests that the bulk of the matter we can infer from observations may not be made of the same stuff that we are. As well as its importance for cosmology, this issue also has more human aspects, if we accept the implication that the material of which we are made is merely a tiny smattering of contaminant in a sea of quite different dark matter.

Historical attempts to determine this fundamental physical parameter of our universe have been fraught with difficulties. Relevant observations are in any case difficult to make, but the issue is also clouded by theoretical and 'aesthetic' preconceptions. In the early days of modern cosmology, Einstein himself was led astray by the conviction that the universe had to be static (i.e. non-expanding). This led him to infer that the universe must be prevented from expansion or contraction by the presence of a cosmological constant term in the equations of general relativity (Einstein 1917; later he called this a great mistake). The first modern attempt to compute the mean density of the matter component was by de Sitter (1917) whose result was several orders of magnitude higher than present estimates, chiefly because distances to the *nebulae* (as galaxies where then known) were so poorly determined. These measurements were improved upon by Hubble (1926), Oort (1932) and Shapley (1934), who estimated the mean density of matter in the universe to be on the order of $\rho \simeq 10^{-30}$ g cm^{-3}. It is interesting to note that as early as 1932, Oort realised that there was evidence of significant quantities of dark matter, whose contribution to the total density would be very difficult for astronomers to estimate. This point was stressed by Einstein (1955):

> ...one can always give a lower bound for ρ but not an upper bound. This is the case because we can hardly form an opinion on how large a fraction a fraction of ρ is given by astronomically unobservable

(not radiating) masses.

In a classic paper, Gott *et al.* (1974) summed up the empirical evidence available at the time, including that for the dark matter, and concluded that the universe had a sufficiently low density for it to be *open*, i.e. with negatively curved spatial sections and with a prognosis of eternal expansion into the future.

The most notable developments in recent years have been (i) on the observational side, the dynamical detection of dark matter from the behaviour of galactic rotation curves, and (ii) on the theoretical side, the introduction by Guth (1981) of the concept of *cosmic inflation* into mathematical models of the universe. We shall discuss inflation further in Chapter 2, but for now it suffices to remark that it essentially involves a period of extremely rapid expansion in the early universe which results in a 'stretching' of the spatial sections to such an extent that the curvature is effectively zero and the universe is correspondingly flat. If the cosmological constant is zero, this implies the existence of a critical density† of matter: $\rho \simeq 10^{-29}$ g cm^{-3} today on average (see §1.2.4).

But is there sufficient matter in the universe for the flat model to be appropriate? There is a considerable controversy raging on this point and – it has to be said – the responsibility for this lies mainly with theorists rather than observers. Without inflation theory, it is probable that most cosmologists would agree to a value of the mean cosmological density of around 10 to 30 per cent of the critical value, but would have an open mind to the existence of more unseen matter than this figure represents, in line with the comments made by Einstein (1955). It is our intention in this book to explore the evidence on this question in as dispassionate a way as possible. Our starting point is that this issue is one that *must* be settled empirically, so while theoretical arguments may give us important insights as to what to expect, they must not totally dominate the argument. We have to separate theoretical prejudice from observational evidence, and ultimately decide the answer on the basis of data about the nature of the real universe. This is not as simple a task as it may seem, because every relevant observa-

† Although this density of matter is sufficient to close the universe, it is, by our standards, a very good vacuum. It corresponds to on the order of one hydrogen atom per cubic metre of space.

tion depends considerably on theoretical interpretation, and there are many subtleties in this. What we are aiming for, therefore, is not a dogmatic statement of absolute truth but an objective discussion of the evidence, its uncertainties and, hopefully, a reasonable interpretation of where the balance of the probabilities lies at present, together with an indication of the most hopeful lines of investigation for arriving at a firm conclusion in the future.

1.2 Cosmological models

We now have to introduce some of the basics of cosmological theory. This book is not intended to be a textbook on basic cosmology, so our discussion here is brief and focusses only on the most directly relevant components of the theory. For more complete introductions see, for example, Weinberg (1972), Kolb & Turner (1990), Narlikar (1993), Peebles (1993) and Coles & Lucchin (1995).

1.2.1 The nature of cosmological models

In this section we shall introduce some of the concepts underpinning the standard big-bang cosmological models†. Before going into the details, it is worth sketching out what the basic ideas are that underlie these cosmological models. Most importantly, there is the realisation that the force of Nature which is most prominent on the large scales relevant to cosmology is gravity‡. The best theory of gravity that we have is Einstein's general theory of relativity. This theory relates three components:

(i) a description of the space-time geometry;
(ii) equations describing the action of gravity;
(iii) a description of the bulk properties of matter.

We shall discuss these further in the following subsections.

† For alternative models, see Ellis (1984a) and Wainwright & Ellis (1996).
‡ It would be electromagnetism if there were a net charge separation on galactic scales.

1.2.2 Geometry: the cosmological principle

The fundamental principle upon which most cosmological models are based is the *cosmological principle*, which is that the universe, at least on large scales, is homogeneous and isotropic. This assumption makes the description of the geometry of cosmological models much simpler than many other situations in which general relativity is employed. One should admit at the outset, however, that one cannot make a watertight case for the global homogeneity and isotropy of the universe (Ellis *et al.* 1996). We know that the universe is quite inhomogeneous on scales up to at least 100 Mpc. The near-homogeneity of the X-ray background, the counts of distant radio sources, and the cosmic microwave background anisotropy limits (Chapter 7) offer at least some circumstantial evidence that the distribution of material on large scales may be roughly homogeneous on scales much larger than this (e.g. Peebles 1993; Stoeger *et al.* 1995a). More recently, large-scale surveys of galaxies have begun to probe depths where one can see, qualitatively at least, the emergence of a homogeneous pattern (Shectman *et al.* 1996). One hopes that this observed behaviour is consistent with the usual treatment of large-scale structure in terms of small perturbations to a globally homogeneous and isotropic background model. We shall discuss this further in Chapter 6. We will, however, take the uncertainty surrounding the detailed application of the cosmological principle on board and, in Chapter 8, explore some issues pertaining to the role of inhomogeneity in cosmology.

Assuming the cosmological principle holds, we first wish to describe the geometrical properties of space-times compatible with it. It turns out that all homogeneous and isotropic space-times can be described in terms of the Friedman–Robertson–Walker (FRW) line element

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a^2(t)\left(\frac{\mathrm{d}r^2}{1-\kappa r^2} + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\mathrm{d}\phi^2\right), \quad (1.1)$$

where κ is the spatial curvature, scaled so as to take the values 0 or ±1, and $u^\alpha = \delta^\alpha_0$ is the average matter four-velocity (defining the world lines of *fundamental observers*). The case $\kappa = 0$ represents flat space sections, and the other two cases are space sections of constant positive or negative curvature, respectively. The time coordinate t is called *cosmological proper time* and it is singled out as a preferred time coordinate by the property of spatial homogeneity: observers can set their clocks according to the local density of matter, which is constant on space-like surfaces orthogonal to the matter four-velocity. The quantity $a(t)$, the *cosmic scale factor*, describes the overall expansion of the universe as a function of time. An important consequence of the expansion is that light from distant sources suffers a cosmological redshift as it travels along a null geodesic in the space-time: $ds = 0$ in equation (1.1). If light emitted at time t_e is received by an observer at t_0 then the redshift z of the source is given by

$$1 + z = \frac{a(t_0)}{a(t_e)}. \tag{1.2}$$

1.2.3 The Friedman equations

The dynamics of an FRW universe are determined by the Einstein gravitational field equations, which can be written, in tensor notation, in the form

$$G^{\nu}_{\mu} = 8\pi G T^{\nu}_{\mu}, \tag{1.3}$$

where T^{ν}_{μ} is the energy-momentum tensor describing the contents of the universe. With this geometry the matter stress-tensor necessarily has the form of a perfect fluid, with $\rho = T^0_0$, $\mathrm{p} = -\frac{1}{3}T^{\alpha}_{\alpha}$, because the matter four-velocity is $u^{\alpha} = \delta^{\alpha}_0$. The Einstein equations then simplify to

$$3\left(\frac{\dot{a}}{a}\right)^2 = 8\pi G\rho - \frac{3\kappa c^2}{a^2} + \Lambda, \tag{1.4}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + 3\frac{p}{c^2}\right) + \frac{\Lambda}{3}, \tag{1.5}$$

$$\dot{\rho} = -3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right). \tag{1.6}$$

These equations determine the time evolution of the cosmic scale factor $a(t)$ (the dots denote derivatives with respect to cosmological proper time t) and therefore describe the global expansion or contraction of the universe. The first equation (the Friedman equation) is a first integral of the other two. In the early phases of the big-bang, the universe is dominated by radiation or relativistic particles for which $p = \rho c^2/3$, while for late times (including now) it is matter-dominated, so that $p \simeq 0$. The crossover between these two regimes occurs when the scale factor a is between 10^{-3} and 10^{-5} of its present value, depending on the density of matter.

In inflationary models of the early universe, there exists a short period of time in which the dynamics of the universe are determined by the action of a scalar field which has an effective equation of state of the form $p = -\rho c^2$ if it is in the slow-rolling regime. We shall discuss this option further in Chapter 2. The cosmological constant Λ, which many cosmologists believe to be exactly zero, changes the acceleration of the universe compared to models containing only matter and/or radiation. A scalar field with $p = -\rho c^2$ behaves in essentially the same way as a cosmological constant term.

It is a property of the homogeneous and isotropic expansion of the universe around every point that these models can easily reproduce Hubble's law for the recession of galaxies:

$$v = H_0 r, \tag{1.7}$$

where r is the proper distance of a galaxy and v is its apparent recession velocity, inferred from the redshifting of spectral lines. The parameter H_0 is called the Hubble constant. In terms of the scale factor, it is straightforward to see that

$$H_0 = (\dot{a}/a)_{t=t_0}, \tag{1.8}$$

with the suffix referring to the present time t_0; in general the expansion rate $H(t) = \dot{a}/a$. The actual value of H_0 is not known with any great accuracy, but is probably in the range 40 km s^{-1} Mpc^{-1} $< H_0 <$ 90 km s^{-1} Mpc^{-1}. This uncertainty is usually parametrised by writing $H_0 = 100h$ km s^{-1} Mpc^{-1}, where h is dimensionless. We shall discuss the observational evidence pertaining to H_0 in Chapter 3: recent estimates suggest $h \simeq 0.7$ (e.g. Freedman *et al.* 1994; Pierce *et al.* 1994) but, to be safe, conservative limits on h are $0.4 \leq h \leq 0.9$.

1.2.4 Open, closed and flat cosmologies

We are now in a position to introduce the relationship between the density of matter and the curvature of the spatial sections we discussed in §1.1. The important parameter for determining the long-term evolution of an FRW universe is the *density parameter*, Ω, which is defined to be

$$\Omega = \frac{8\pi G\rho}{3H^2} = \frac{\rho}{\rho_{cr}}, \tag{1.9}$$

in other words, the ratio of the actual density of the universe to a critical value ρ_{cr}. The present value of this critical density depends on H_0:

$$\rho_{cr} = \frac{3H_0^2}{8\pi G} \simeq 1.9 \times 10^{-29} h^2 \,\mathrm{gm\,cm^{-3}}. \tag{1.10}$$

The value (1.10) is the yardstick against which we shall measure the various determinations of ρ_0 and we shall henceforth give most of our estimates in terms of Ω-values.

If $\Omega > 1$ then $\kappa = +1$ – elliptic spatial sections – and the universe will recollapse to a second singularity (a 'big crunch'); if $\Omega < 1$ then $\kappa = -1$ – hyperbolic spatial sections – and it will expand forever with an ever-decreasing density. In between, $\Omega = 1$ corresponds to the flat $\kappa = 0$ universe favoured by some inflationary models for the early universe (e.g. Guth 1981; Linde 1982). The relationship between κ and Ω becomes more complicated if we allow the cosmological constant to be non-zero.

1.2.5 The equation of state

The Friedman equation (1.4) becomes determinate when we select an equation of state of the form $p = p(\rho)$ for the matter. Particularly relevant examples are $p = 0$ for pressureless matter (or 'dust') and $p = \frac{1}{3}\rho c^2$ for relativistic particles or radiation. One can then use the conservation equation (1.6) to determine $\rho(a)$. The result for pressureless matter is $\rho = M/a^3$, while for radiation it is $\rho = M/a^4$, where M is constant.

Alternatively, one can represent the matter in terms of a scalar field whose evolution is governed by a potential $V(\phi)$ – this is the case in inflationary models in particular (see §2.3 below). In such a case the effective density and pressure are given by

$$\begin{aligned} \rho c^2 &= \frac{1}{2}\dot{\phi}^2 + V(\phi), \\ p &= \frac{1}{2}\dot{\phi}^2 - V(\phi). \end{aligned} \tag{1.11}$$

We discuss some aspects of scalar field cosmologies in Chapter 2.

If $\Omega = 1$ the scale factor evolves according to: (i) $a(t) \propto t^{2/3}$ for pressureless matter, and (ii) $a(t) \propto t^{1/2}$ for radiation. It is plausible that the universe is radiation-dominated from the time of nucleosynthesis to about decoupling, and matter-dominated from then on, and so early on behaves like a radiation universe and later behaves like a matter universe. In a universe with $\Omega < 1$, the universe expands roughly like this until it reaches a fraction Ω of its present size and then goes into free expansion with $a(t) \propto t$. If $\Omega > 1$, it expands more slowly than the critical density case, and then recollapses. This behaviour has implications for structure formation, as we shall see in Chapter 6.

1.2.6 Useful formulae

It is often convenient to express the dynamical equations in terms of H and Ω rather than t and ρ. From the Friedman equation, referred to the reference time t_0, one can write

$$\left(\frac{\dot{a}}{a_0}\right)^2 - \frac{8\pi G\rho}{3}\left(\frac{a}{a_0}\right)^2 = H_0^2(1-\Omega_0) = -\frac{\kappa c^2}{a_0^2} = \kappa_0. \qquad (1.12)$$

One can also write

$$\left(\Omega^{-1} - 1\right)\rho(t)a(t)^2 = -\kappa c^2 = \left(\Omega_0^{-1} - 1\right)\rho_0(t_0)a(t_0)^2, \qquad (1.13)$$

which we shall find useful later, in Chapter 2. The appropriate relation between curvature and Ω_0 in the presence of the Λ-term is

$$\frac{\kappa c^2}{a_0^2} - \frac{\Lambda c^2}{3} = H_0^2(\Omega_0 - 1) \qquad (1.14)$$

(this is just the Friedman equation again, evaluated at $t = t_0$).

1.2.7 The big-bang model

Most cosmologists accept the big-bang model as basically correct: the evidence in favour of it is circumstantial but extremely convincing (Peebles *et al.* 1991). In particular, we can quote the agreement of predicted and observed abundances of atomic light nuclei (Chapter 4) and the existence of the microwave background radiation (Chapter 7), a relic of the primordial fireball. It is important to remember, however, that the big-bang model is not a complete theory. For example, it does not specify exactly what the matter contents of the universe are, nor does it make a prediction for the expansion rate H_0. It also breaks down at the singularity at $t = 0$, where the equations become indeterminate: 'initial conditions' therefore have to be added at some fiducial early time where the equations are valid. Theorists therefore have considerable freedom to play around with parameters of the model in order to fit observations. But the big-bang model does at least provide a conceptual framework within which data can be interpreted and analysed, and allows meaningful scientific investigation of the thermal history of the universe from times as early as the Planck epoch $t = t_P$, where

$$t_P = \left(\frac{\hbar G}{c^5}\right)^{1/2} \simeq 10^{-43}\,\mathrm{s} \tag{1.15}$$

(e.g. Kolb & Turner 1990; Linde 1990).

1.3 Cosmological criteria

It will be apparent to anyone who has glanced at the literature, including the newspapers, that there is a great deal of controversy surrounding the issue of Ω_0, sometimes reinforced by a considerable level of dogmatism in opposing camps. In understanding why this is the case, it is perhaps helpful to note that much of the problem stems from philosophical disagreements about which are the appropriate criteria for choosing an acceptable theory of cosmology. Different approaches to cosmology develop theories aimed at satisfying different criteria, and preferences for the different approaches to a large extent reflect these different initial goals. It would help to clarify this situation if one could make explicit the issues relating to choices of this kind, and separate them from the more 'physical' issues that concern the interpretation of data. Pursuing this line of thought, we now embark on a brief philosophical diversion which we hope will initiate a debate within the cosmological community†.

† Some cosmologists in effect claim that there is no philosophical content in their work and that philosophy is an irrelevant and unnecessary distraction