The homogeneous and isotropic universe

9

- The framework required to study the whole universe as a physical system is general relativity.
- The universe, when observed on distance scales $\gtrsim$100 Mpc, is homogeneous and isotropic.
- Hubble's discovery that the universe is expanding suggests strongly that it had a beginning when all objects were concentrated in a tiny region of extremely high density. The estimate of the age of the universe by astrophysics from observational data is 12 Gyr $\lesssim t_0 \lesssim$ 15 Gyr.
- The mass density of (nonrelativistic) matter in the universe has around a quarter of the "critical density" $\Omega_{\rm M} \simeq 0.25$. There is strong evidence showing that most of the mass in the universe does not shine: while the luminous mass ratio $\Omega_{\rm lum}$ is only half a percent, the nonluminous matter consists of ordinary (baryonic) matter $\Omega_{\rm B} \simeq 0.04$ (mostly as the intergalactic medium) and nonrelativistic exotic dark matter $\Omega_{\rm DM} \simeq 0.21$.
- The spacetime satisfying the cosmological principle (the universe is homogeneous and isotropic at each epoch) is described by the Robertson–Walker metric in comoving coordinates (the cosmic rest frame).
- In an expanding universe with a space that may be curved, any treatment of distance and time must be carried out with care. We study the relations between cosmic redshift and proper, as well as luminosity, distances.

Cosmology is the study of the whole universe as a physical system: What is its matter–energy content? How is this content organized? What is its history? How will it evolve in the future? We are interested in a "smeared" description with the galaxies being the constituent elements of the system. On the cosmic scale the only relevant interaction among galaxies is gravitation; all galaxies are accelerating under their mutual gravity. Thus the study of cosmology depends crucially on our understanding of the gravitational interaction. Consequently, the proper framework for cosmology is general relativity. The solution of Einstein's equation describes the whole universe because it describes the whole of spacetime.

From Chapter 7 we learnt that, for a given gravitational system (M and R being the respectively characteristic mass and length dimensions), one could use the dimensionless parameter

$$\frac{G_{\rm N}M}{c^2R} \equiv \varepsilon \tag{9.1}$$

to decide whether Einstein's theory was required, or a Newtonian description would be adequate. In the context of the spatially isotropic solution, it is just the relative size of the Schwarzschild radius to the distance scale R. Recall $\varepsilon_\odot = O(10^{-6})$ for the sun, cf. Eq. (7.22). Typically the GR effects are small at the level of an ordinary stellar system. On the other hand, we have also considered the case of stellar objects that were so compact that they became black holes when the distance scale is comparable to the Schwarzschild radius, $\varepsilon_{\rm bh} = O(1)$. For the case of cosmology, the mass density is very low. Nevertheless, the distance involved is so large that the total mass M, which increases faster than R, is even larger. This also results in a sizable ε (Problem 9.1). Thus, to describe events on cosmic scales, we must use GR concepts.[1]

[1]Given that the theory had been tested only within the solar system, applying GR to cosmology would involve an extraordinary (something like 15 orders of magnitude) extrapolation. This is a bold assumption indeed.

Soon after the completion of his papers on the foundation of GR, Einstein proceeded to apply his new theory to cosmology. In 1917 he published his paper, "Cosmological considerations on the general theory of relativity". Since then almost all cosmological studies have been carried out in the framework of GR.

9.1 The cosmos observed

We begin with the observational features of the universe: the organization of its matter content, the large-scale motion of its components, its age and mass density.

9.1.1 Matter distribution on the cosmic distance scale

The distance unit traditionally used in astronomy is the parsec (pc). This is defined, see Fig. 9.1(a), as the distance to a star having a parallax of one arcsecond[2] for a base-line equal to the (mean) distance between the earth and the sun (called an AU, the **astronomical unit**). Thus 1pc $= (1''$ in radian$)^{-1} \times$ AU $= 3.1 \times 10^{16}$ m $= 3.26$ light-years. Here we first introduce the organization of stars on the cosmic scales of kpc, Mpc, and even hundreds of Mpc.

[2]One arcsec equals 4.85×10^{-6} rad.

The distance from the solar system to the nearest star is 1.2 pc. Our own galaxy, the Milky Way, is a typical spiral galaxy. It comprises $O\,(10^{11})$ stars in a disk with a diameter of 30 kpc and a disk thickness of about 2 kpc, see Fig. 9.1(b). Galaxies in turn organize themselves into bodies of increasingly large sizes—into a series of hierarchical clusters. Our galaxy is part of a small cluster, called the Local Group, comprising about 30 galaxies in a volume measuring 1 Mpc across; the distance, for example, to the Andromeda Galaxy (M31) is 0.7 Mpc. This cluster is part of the Local, or Virgo, Supercluster over a volume measuring 50 Mpc across, with the Virgo Cluster comprising 2000 galaxies over a distance scale of 5 Mpc as its physical center. (The Virgo

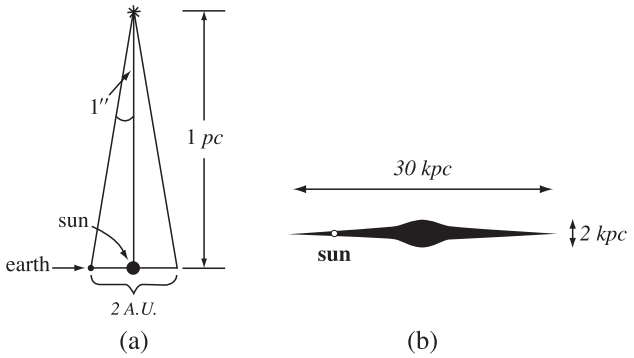

Fig. 9.1 (a) The astronomical distance unit **parsec** (parallax second) defined, see text. (b) Side view of Milky Way as a typical spiral galaxy.

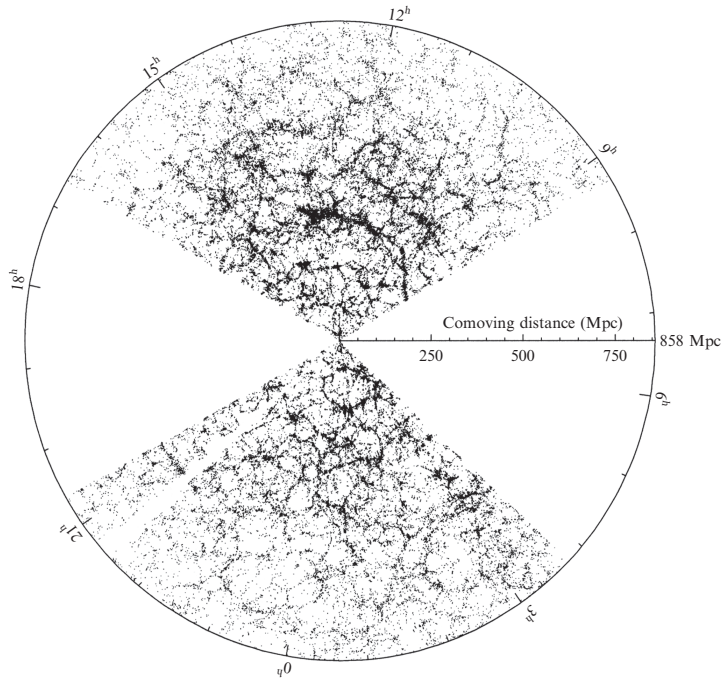

Fig. 9.2 Galaxy distribution out to 858 Mpc, compiled by Gott *et al.* (2005) based on data collected by the SDSS and 2dF surveys.

Cluster is about 15 Mpc from us.) This and other clusters of galaxies, such as the Hydra–Centaurus Supercluster, appear to reside on the edge of great voids. In short, the distribution of galaxies about us is not random, but rather clustered together in coherent patterns that can stretch out up to 100 Mpc. The distribution is characterized by large voids and a network of filamentary structures (see Fig. 9.2). However, beyond this distance scale the universe does appear to be fairly uniform. In fact the largest observable item in the universe is the cosmic microwave background (CMB) radiation which appears to be remarkably uniform.

9.1.2 Cosmological redshift: Hubble's law

Olbers' paradox: Darkness of the night sky Up until about 100 years ago, the commonly held view was that we lived in a static universe (comprising essentially our Milky Way galaxy) that was infinite in age and infinite in size. However, such a cosmic picture is contradicted by the observation that the night sky is dark. If the average luminosity (emitted energy per unit time) of a star is $\mathfrak{L}$, then the brightness (i.e. flux) seen at a distance r would be $f(r) = \mathfrak{L}/4\pi r^2$. The resultant flux from integrating over all the stars in the infinite universe would be unbounded:

$$B = \int n\, f(r) dV = n\mathfrak{L} \int_{r_{\min}}^{\infty} dr = \infty, \tag{9.2}$$

where n, the number density of stars, has been assumed to be a constant. This result of infinite brightness is an over-estimate because stars have finite angular sizes, and the above calculation assumes no obstruction by foreground stars. The correct conclusion is that the night sky in such a universe would have the brightness as if the whole sky were covered by shining suns. Because every line-of-sight has to end at a shining star, although the flux received from a distant star is reduced by a factor of r^{-2}, for a fixed solid angle, the number of unobstructed stars increases with r^2. Thus, there would be an equal amount of flux from every direction. It is difficult to find any physical mechanism that will allow us to evade this result of a night sky ablaze. For example, one might suggest that interstellar dust would diminish the intensity for light having traveled a long distance. But this does not help, because over time, the dust particles would be heated and radiate as much as they absorb. Maybe our universe is not an infinite and static system?[3]

[3] As we shall see, according to modern cosmology, our universe has a finite age and all distant stars in an expanding universe are receding away from us with their emitted light progressively shifted to lower frequencies. Cf. Problem 9.5.

Hubble's discovery

Astronomers have devised a whole series of techniques that can be used to estimate the distances ever farther into space. Each new one, although less reliable, can be used to reach out further into the universe. During the period 1910–1930, the "cosmic distance ladder" reached out beyond 100 kpc. The great discovery was made that our universe was composed of a vast collection of galaxies, each resembling our own Milky Way. One naturally tried to study the motions of these newly discovered "island universes" by using the Doppler effect. When a galaxy is observed at visible wavelengths, its spectrum typically has absorption lines because of the relatively cool upper stellar atmosphere. For a particular absorption line measured in the laboratory as having a wavelength $\lambda_{\rm em}$, the received wavelength by the observer may, however, be different. Such a wavelength shift

$$z \equiv \frac{\lambda_{\rm rec} - \lambda_{\rm em}}{\lambda_{\rm em}} \tag{9.3}$$

is related to the emitter motion by the Doppler effect (cf. Box 3.3), which, for nonrelativistic motion, can be stated as

$$z = \frac{\Delta\lambda}{\lambda} \simeq \frac{v}{c}, \tag{9.4}$$

where v is the recession velocity of the emitter (away from the receiver).

A priori, for different galaxies one expects a random distribution of wavelength shifts: some positive (redshift) and some negative (blueshift). This is more or less true for the Local Group. But beyond the few nearby galaxies, the measurements by Vesto Slipher of some 40 galaxies, over a 10 year period at Arizona's Lowell Observatory, showed that all, except a few in the Local Group, were redshifted. Edwin Hubble (Mt. Wilson Observatory, California) then attempted to correlate these redshift results to the more difficult measurements of the distances to these galaxies. He found that the redshift was proportional to the distance d to the light-emitting galaxy. In 1929, Hubble announced his result:

$$z = \frac{H_0}{c} d \tag{9.5}$$

or, substituting in the Doppler interpretation[4] of (9.4),

$$v = H_0 d, \tag{9.6}$$

with a positive H_0. Namely, we live in an expanding universe. On distance scales greater than 10 Mpc, all galaxies obey Hubble's law: they are receding from us with speed linearly proportional to the distance. The proportional constant H_0, the **Hubble constant**, gives the recession speed per unit separation (between the receiving and emitting galaxies). It is the expansion rate. To obtain an accurate account of H_0 has been a great challenge as it requires one to ascertain great cosmic distances. Only recently has it become possible to yield consistent results among several independent methods. We have the convergent value[5]

$$H_0 = (72 \pm 5 \text{ km/s}) \text{ Mpc}^{-1}, \tag{9.7}$$

where the subscript 0 stands for the present epoch $H_0 \equiv H(t_0)$. An inspection of Hubble's law (9.6) shows that H_0 has the dimension of inverse time, the **Hubble time** $t_{\rm H} \equiv H_0^{-1}$. Similarly, we can also define a **Hubble length** $l_{\rm H} = ct_{\rm H}$.

[4] A Doppler redshift comes about because of the increase in the distance between the emitter and the receiver of a light signal. In the familiar situation, this is due to the relative motion of the emitter and the receiver. This language is being used here in our initial discussion of Hubble's law. However, as we shall show in Section 9.3, especially Eq. (9.43), the proper description of this enlargement of the cosmic distance is reflecting the expansion of the space itself, rather than the motion of the emitter in a static space.

[5] Throughout Chapters 9–11, we shall quote the cosmological parameters as presented by Tegmark *et al.* (2006), cf. Table 11.1 in Section 11.5.

Hubble's law and the Copernican principle

That all galaxies are receding away from us may lead one to suggest erroneously that our location is the center of the universe. The correct interpretation is in fact just the opposite. The Hubble relation in fact follows naturally from a straightforward extension of the **Copernican principle**: our galaxy is not at a privileged position in the universe. The key observation is that this is a **linear relation** between distance and velocity at each cosmic epoch. As a result, it is compatible with the same law holding for all observers at every galaxy. Namely, observers on every galaxy would see all the other galaxies receding away from them according to Hubble's law.

Let us write Hubble's law in vector form:

$$\vec{v} = H_0 \vec{r}. \tag{9.8}$$

That is, a galaxy G, located at position $\vec{r}$, will be seen by us (at the origin O) to recede at velocity $\vec{v}$ proportional to $\vec{r}$. Now consider an observer on another galaxy O′ located at $\vec{r}'$ from us as in Fig. 9.3. Then, according to Hubble's law,

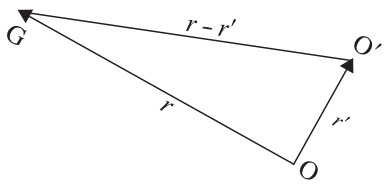

Fig. 9.3 Relative positions of a galaxy G with respect to two observers located at two other galaxies: O and O'.

it must be receding from us according to

$$\vec{v}' = H_0 \vec{r}' \tag{9.9}$$

with the **same** Hubble constant H_0 which is independent of distance and velocity. The difference of these two equations yields

$$(\vec{v} - \vec{v}') = H_0(\vec{r} - \vec{r}'). \tag{9.10}$$

But $(\vec{r} - \vec{r}')$ and $(\vec{v} - \vec{v}')$ are the respective location and velocity of G as viewed from O′. Since $\vec{v}$ and $\vec{v}'$ are in the same direction as $\vec{r}$ and $\vec{r}'$, the vectors $(\vec{v} - \vec{v}')$ and $(\vec{r} - \vec{r}')$ must also be parallel. That is, the relation (9.10) is just Hubble's law valid for the observer on galaxy O′. Clearly such a deduction would fail if the velocity and distance relation, at a given cosmic time, were nonlinear (i.e. if H_0 depends either on position and/or on velocity).

Distance measurement by redshift

We can turn the Hubble relation around and use it as a means to find the distance to a galaxy by its observed redshift. In fact, the development of new techniques of multi-fiber and multi-slip spectrographs allowed astronomers to measure redshifts for hundreds of galaxies simultaneously. This made large surveys of galaxies possible. In the 1980s there was the Harvard–Smithsonian Center for Astrophysics (CfA) galaxy survey, containing more than 15 000 galaxies. Later, the Las Campanas mapping eventually covered a significantly larger volume and found the "greatness limit" (i.e. cosmic structures have a maximum size and on any larger scale the universe would appear to be homogeneous). But this was still not definitive. The modern surveys culminated in two recent parallel surveys: the Anglo-Australian Two-Degree Field Galaxy Redshift Survey (2dF) and the Sloan Digital Sky Survey (SDSS) collaborations have measured some quarter of a million galaxies over a significant portion of the sky, confirming the basic cosmological assumption that the universe of a large distance $\gtrsim 100$ Mpc is homogeneous and isotropic. (For further discussion see Sections 9.3 and 9.4.) In fact, an important tool for modern cosmology is just such large-structure study. Detailed analysis of survey data can help us to answer questions such as whether the cosmic structure observed today came about in a top–down (i.e. the largest structure formed first, then the smaller ones by fragmentation) or in a bottom–up process. (The second route is favored by observational data.) In fact many of the cosmological parameters, such as Hubble's constant and the energy density of the universe, etc. can also be extracted from such analysis.

9.1.3 Age of the universe

If all galaxies are rushing away from each other now presumably they must have been closer in the past. Unless there was some new physics involved, extrapolating back in time there would be a moment, "**the big bang**", when all objects were concentrated at one point of infinite density[6]. This is taken to be the origin of the universe. How much time has evolved since this fiery beginning? What is then the age of our universe?

[6]See Problem 9.10 for a brief description of the alternative cosmology called the **steady-state theory** which avoids the big bang beginning by having a constant mass density, maintained through continuous spontaneous matter creation as the universe expands.

It is useful to note that the inverse of the Hubble's constant at the present epoch, the **Hubble time**, has the value of

$$t_{\rm H} \equiv H_0^{-1} = 13.6 \text{ Gyr}. \tag{9.11}$$

By Hubble "constant," we mean that, at a given cosmic time, H is independent of the separation distance and the recessional velocity—the Hubble relation is a linear relation. The proportional coefficient between distance and recessional speed is not expected to be a constant with respect to time: there is matter and energy in the universe, and their mutual gravitational attraction will slow down the expansion, leading to a monotonically decreasing expansion rate $H(t)$—a **decelerating universe**. Only in an "empty universe" do we expect the expansion rate to be a constant throughout its history, $H(t) = H_0$. In that case, the age t_0 of the empty universe is given by the Hubble time

$$t_{0\text{empty}} = \frac{d}{v} = \frac{1}{H_0} = t_{\rm H}. \tag{9.12}$$

For a decelerating universe full of matter and energy, the expansion rate must be larger in the past: $H(t) > H_0$ for $t < t_0$. Because the universe was expanding faster than the present rate, this would imply that the age of the decelerating universe must be shorter than the empty universe age: $t_0 < t_{\rm H}$. Nevertheless, we shall often use the Hubble time as a rough benchmark value for the age of the universe, which has a current horizon[7] of $ct_{\rm H} = l_{\rm H} \simeq 4300$ Mpc.

[7] Two objects, separated by a distance of $ct_{\rm H}$, would recede from each other, according to the Hubble relation of (9.6), at the speed of light c.

Phenomenologically, we can estimate the age of the universe from observational data. For example, from astrophysical calculation, we know the relative abundance of nuclear elements when they are produced in a star. Since they have different decay rates, their present relative abundance will be different from the initial value. The difference is a function of time. Thus, from the decay rates, the initial and observed relative abundance, we can estimate the time that has elapsed since their formation. Typically, such a calculation gives the ages of stars to be around 13.5 Gyr. This only gives an estimate of time when stars were first formed, thus only a lower bound for the age of the universe. However, our current understanding informs us that the formation of stars started a hundred million years or so after the big bang, thus such a lower limit still serves as a useful estimate of t_0.

An important approach to the study of the universe's age has been the research work on systems of 10^5 or so old stars known as **globular clusters**. These stars are located in the halo, rather than the disk, of our Galaxy. It is known that a halo lacks the interstellar gas for star formation. These stars must be created in the early epochs after the big bang (as confirmed by their lack of elements heavier than lithium, cf. Section 10.4). Stars spend most of their lifetime undergoing nuclear burning. From the observed brightness (flux) and the distance to the stars, one can deduce their intrinsic luminosity (energy output per unit time). From such properties, astrophysical calculations based on established models of stellar evolution, allowed one to deduce their ages (Krauss and Chaboyer, 2003):

$$12 \text{ Gyr} \lesssim t_{0\text{gc}} \lesssim 15 \text{ Gyr}. \tag{9.13}$$

For reference, we note that the age of our earth is estimated to be around 4.6 Gyr.

9.2 Mass density of the universe

It is useful to express the mass density in terms of a benchmark value for a universe with expansion rate given by the Hubble constant H. One can check that the ratio, with H^2 divided by Newton's constant G_N, has the units of mass density. With an appropriate choice[8] of the coefficient, we have the expression of the **critical density**

$$\rho_c \equiv \frac{3H^2}{8\pi G_N}. \tag{9.14}$$

[8]We can remember ρ_c as the density of a universe with its radius $R = ct_H = c/H$ just equal to the Schwarzschild radius: $R = 2G_N M/c^2$ where $M = (4\pi R^3/3)\rho_c$.

The significance of this quantity will be discussed in Chapter 10 when the Einstein equation for cosmology will be presented. In the meantime, we introduce the notation for the **density ratio**

$$\Omega \equiv \frac{\rho}{\rho_c}. \tag{9.15}$$

Since the Hubble constant is a function of cosmic time, the critical density also evolves with time. We denote the values for the present epoch with the subscript 0. For example, $\rho(t_0) \equiv \rho_0$, $\rho_c(t_0) \equiv \rho_{c,0}$, and $\Omega(t_0) \equiv \Omega_0$, etc. For the present Hubble constant H_0 as given in (9.7), the critical density has the value

$$\rho_{c,0} = (0.97 \pm 0.08) \times 10^{-29}\ \text{g/cm}^3 \tag{9.16}$$

or, equivalently, a **critical energy density**[9] of

$$\rho_{c,0}c^2 \simeq 0.88 \times 10^{-10}\ \text{J/m}^3 \simeq 5500\ \text{eV/cm}^3. \tag{9.17}$$

[9]In the natural unit system of quantum field theory, this critical density is approximately $\rho_c c^2 \approx (2.5 \times 10^{-3} eV)^4/(\hbar c)^3$, where $\hbar$ is Planck's constant (over 2π) with $\hbar c \approx 1.9 \times 10^{-5}$eV · cm. Also, $\rho_{c,0}c^2 \simeq 5.5\,\text{GeV/m}^3$ is equivalent to the rest energy of $\simeq 6$ protons per cubic meter.

In the following we shall discuss the measurement of the universe's various mass densities (averaged over volumes on the order of 100 Mpc^3) for both luminous and nonluminous matter. In recent years, these parameters have been deduced rather accurately by somewhat indirect method—including detailed statistical analysis of the temperature fluctuation in the cosmic microwave background (CMB) radiation and from large-structure studies by the 2dF and SDSS galaxy surveys mentioned above. The large-structure study involves advanced theoretical tools that are beyond the scope of this introductory presentation. In the following we choose to present a few methods that involve rather simple physical principles, even though they may be somewhat "dated" in view of recent cosmological advances. Our discussion will in fact be only semiquantitative. Subtle details of derivation as well as qualification of the stated results will be omitted. The purpose is to provide some general idea as to how cosmological parameters can in principle be deduced phenomenologically.

9.2.1 Luminous matter and the baryonic density

Luminous matter

The basic idea of measuring the mass density for luminous matter is through its relation to the luminosity $\mathfrak{L}$, which is the energy emitted per unit time,

$$\rho_{\rm lum} = \begin{pmatrix} \text{luminosity} \\ \text{density} \end{pmatrix} \times \left(\frac{M}{\mathfrak{L}}\right). \tag{9.18}$$

(Here we omit the subscript 0 for the present epoch.) That is, one finds it convenient to decompose the mass density into two separate factors: the luminosity density and the mass-to-luminosity ratio. The luminosity density can be obtained by a count of galaxies per unit volume, multiplied by the average galactic luminosity. Several surveys have resulted in a fairly consistent conclusion of 200 million solar luminosity $\mathfrak{L}_\odot$ per Mpc^3 volume,

$$\begin{pmatrix} \text{luminosity} \\ \text{density} \end{pmatrix} \approx 2 \times 10^8 \frac{\mathfrak{L}_\odot}{(\text{Mpc})^3}. \tag{9.19}$$

The ratio $(M/\mathfrak{L})$ is the amount of mass associated, on the average, with a given amount of light. This is the more difficult quantity to ascertain. Depending on the selection criteria one gets a range of values for the mass-to-luminosity ratio. The average of these results came out to be $(M/\mathfrak{L}) \approx 4M_\odot/\mathfrak{L}_\odot$. Plugging this and (9.19) into (9.18) we obtain an estimate of the density for luminous matter $\rho_{\rm lum} \approx 8 \times 10^8 M_\odot/\text{Mpc}^3 \approx 5 \times 10^{-32}\ \text{g/cm}^3$, or in terms of the density ratio defined in (9.15)

$$\Omega_{\rm lum} \approx 0.005. \tag{9.20}$$

Total amount of baryonic matter and the intergalactic medium

We designate the type of matter, for which we cannot directly detect its presence through its electromagnetic emissions, as nonluminous matter. This includes such matter as neutrinos which have no electromagnetic interaction, as well as matter such as intergalactic hydrogen molecules, which, although they do not "shine," can be detected through their absorption of electromagnetic radiation.

Ordinary matter made of baryons (protons and neutrons) and electrons is referred to in cosmology as **baryonic matter**.[10] Baryons is the particle physics name for strongly interacting particles, composed of quark triplets, that carry nontrivial **baryon numbers**—as are the cases of protons and neutrons. For our purpose here, the baryon number is just the proton plus neutron numbers. Other types of particles, such as photons, electrons, and neutrinos, carry zero baryon number. Baryon matter (protons, neutrons and electrons) can clump to form atoms and molecules, leading to large astronomical bodies. Luminous matter (shining stars) is baryonic matter; but some of the baryonic matter, such as interstellar or intergalactic gas, may not shine—they are nonluminous baryonic matter.[11] That is, baryonic matter can be luminous stars or optically nonluminous gas[12] of ordinary atoms:

$$\Omega_{\rm B} = \Omega_{\rm lum} + \Omega_{\rm gas}. \tag{9.21}$$

[10] Such a name neglects electrons (one species of leptons), which constitute less than 0.1% of the baryonic matter masses.

[11] Besides the interstellar gas around galaxies, nonluminous baryonic matter can be planets or stellar remnants such as black holes, white dwarfs, and brown dwarfs (the last category being stars of the size of Jupiter, with not enough mass to trigger the thermonuclear reaction to make it shine).

[12] This includes X-ray emitting hot gas.

As it turns out, we have methods that can deduce the total baryonic abundance Ω_B regardless of whether they are luminous or nonluminous. The light nuclear elements (helium, deuterium, etc.) were produced predominantly in the early universe at the cosmic time $O(10^2\ s)$, cf. Section 10.4. Their abundance (in particular deuterium) is sensitive to the baryonic abundance. From such considerations one deduces the result (Burles *et al.* 2001; Tegmark *et al.* 2006)

$$\Omega_B = 0.042 \pm 0.002, \tag{9.22}$$

which is confirmed by the latest CMB anisotropy measurements (see Chapter 11), as well as gravitational microlensing (see Box 7.2).

From (9.20), we see that $\Omega_B \gg \Omega_{lum}$. This means that most of the "ordinary matter" is not visible to us. Theoretical studies, backed up by detailed simulation calculations, indicate that a major portion of it is in the form of unseen neutral gas in galaxies as well as in the space between galaxies. Such an intergalactic medium (IGM) is in the form of wispy filaments that connect galaxy clusters. Their presence has been verified by careful examination of quasar spectra. Quasars are among the most powerful light sources in the universe. Their light reaches us after passing through successive layers of IGM at various distances resulting in a quasar spectrum imprinted with neutral hydrogen absorption lines. From the line depths one can infer the amount and distribution of the absorbing gas. Such studies were able to account[13] for the difference between Ω_B and Ω_{lum}; namely, an optically nonluminous $\Omega_{gas} \approx 0.038$.

9.2.2 Dark matter and the total mass density

One of the great discoveries of modern cosmology has been the finding that there is more mass in the universe than just baryonic matter. That is, the bulk of the nonluminous matter is not baryonic. We call such nonluminous and nonbaryonic matter, **dark matter**.

Dark matter vs. baryonic matter

Dark matter is supposedly made up of exotic particles that have neither electromagnetic emission nor absorption. Namely, they have no electromagnetic interaction at all (i.e. they do not have strong or electromagnetic charges). Neutrinos are cases in point. They only feel the weak nuclear force. With their masses being extremely small, neutrinos are expected to be in relativistic motion. They are examples of "**hot dark matter**." Also, there may exist "**cold dark matter**" composed of nonrelativistic heavy particles. Hot and cold dark matter have distinctly different effects on the formation of galaxies and clusters of galaxies from the initial density inhomogeneity in the universe. Research in the past decade favors the possibility of cold dark matter.[14] The prime examples of CDM are the "weakly interacting massive particles" (WIMPs) predicted by the various extensions of the Standard Model of particle interactions.[15] WIMPs are expected to be much more massive than nucleons (in the 50–1000 GeV/c^2 range) but interact weakly—a particle with such a mass and interaction rate can produce just the correct CDM abundance in the big bang cosmology, to be discussed in Chapter 10 (see in particular the related subject of primordial

[13] Until very recently, such an IGM has been detected in the early universe; finding such nonluminous atoms in the nearby universe had not been successful. However, theoretical studies (e.g. Cen and Ostriker, 1999) suggest such IGM baryons should have been shock-heated by the large-scale collapsing and squeezing that formed the foamy cosmic structure. The corresponding absorption lines of the heated atoms move up to the far-ultraviolet and X-ray region. Measurements of such absorption lines with the expected intensity has recently been reported (Danforth and Shull, 2008).

[14] If the dark matter had been fast moving (hot) particles, they would be able to stream away from high density regions, thus smooth out small density perturbations. This would have left only the large-scale perturbations, leading to the formation of the largest structure (superclusters) first, with the smaller structures (galaxies) being produced from fragmentation. This top-down scenario is inconsistent with observation.

[15] It has been suggested that the Standard Model of particle physics be extended by the inclusion of supersymmetry (cf. discussion in Section 11.7.3). Every known elementary particle must then have a supersymmetric partner, with a spin differing by half a unit. The lightest of such hypothesized supersymmetry particles are expected to be **neutralino** fermions (partners to the neutral Higgs scalar and weak gauge bosons) and should be stable against spontaneous decay. They can in principle make up the bulk of the required dark matter WIMPs.

neutrinos in Section 10.5.3). For a recent review,[16] see for example Bertone, Hooper and Silk (2005).

The total mass of the universe can thus be divided into two categories: baryonic, which may be luminous or nonluminous, and dark matter,[17] which has only weak interaction:

$$\Omega_{\rm M} = \Omega_{\rm B} + \Omega_{\rm DM}. \tag{9.23}$$

Although the dark matter does not emit electromagnetic radiation, it still feels gravitational effects. In the following we first list several methods of detecting the total amount of masses, whether due to luminous or nonluminous matter, through their gravitational interaction.

Galactic rotation curves

The most direct evidence of dark matter's existence comes from measured "rotation curves" in galaxies. Consider the gravitational force that a spherical (or ellipsoidal) mass distribution exerts on a mass m located at a distance r from the center of a galaxy, see Fig. 9.4(a). Since the contribution outside the Gaussian sphere (radius r) cancels out, only the interior mass $M(r)$ enters into the Newtonian formula for gravitational attraction. The object is held by this gravity in circular motion with centripetal acceleration v^2/r. Hence

$$v(r) = \sqrt{\frac{G_{\rm N} M(r)}{r}}. \tag{9.24}$$

In this way, the tangential velocity inside a galaxy is expected to rise linearly with the distance from the center ($v \sim r$) if the mass density is approximately constant. For a light source located outside the galactic mass distribution (radius R), the velocity is expected to decrease as $v \sim 1/\sqrt{r}$, see Fig. 9.4(b).

The velocity of particles located at different distances (the rotation curves) can be measured through the 21-cm lines of the hydrogen atoms. The surprising discovery was that, beyond the visible portion of the galaxies ($r > R$), instead of this fall-off, they are observed to stay at the constant peak value (as far as the measurement can be made). See, for example, Cram *et al.* (1980). This indicates that the observed object is gravitationally pulled by other than the luminous matter; hence it constitute direct evidence for the existence of dark matter. Many subsequent studies confirm this discovery. The general picture of a galaxy that has emerged is that of a disk of stars and gas embedded in a large halo of dark matter, see Fig. 9.5. According to (9.24), the flatness

[16] Among other examples of speculated CDM particles are **axions** and **Kaluza–Klein particles**. Axions are associated with our effort to explain how the strong interaction theory of QCD avoids having a large violation of the combined symmetry of charge conjugation and parity. The KK particles are associated with the speculated existence of compactified extra spatial dimensions.

[17] Hot dark matter such as neutrinos contribute a negligible amount.

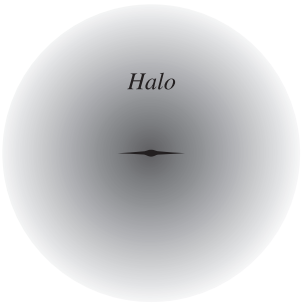

Fig. 9.5 The dark matter halo surrounding the luminous portion of the galaxy. In our simple presentation, we take the halo to be spherical. In reality the dark matter halo may not be spherical and its distribution may not be smooth.

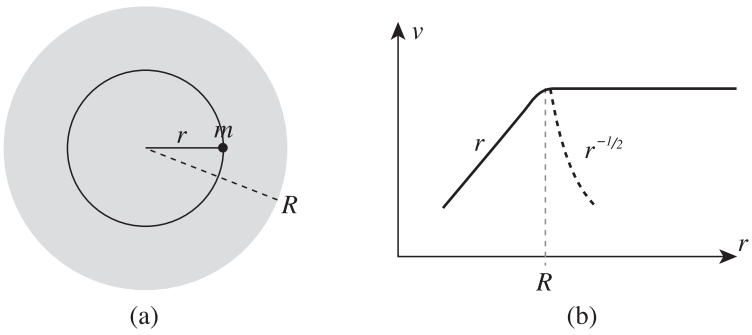

Fig. 9.4 (a) Gravitational attraction on a mass m due to a spherical mass distribution (shaded disk). The circle passing through m represents the Gaussian spherical surface. (b) The solid line is the observed rotation velocity curve $v(r)$. It does not fall as $r^{-1/2}$ beyond R, the edge of the visible portion of a galaxy.

of the rotation curve means that $M \propto r$. We can think of the halo as a sphere with mass density decreasing as r^{-2}. Measurements of the rotational curve for spiral galaxies have shown that halo radii are at least ten times larger than the visible radii of the galaxies. This leads to a lower bound on the dark matter density of $\Omega_{\rm DM} \gtrsim 0.1$.

Use of the virial theorem to infer gravitational mass

Because the rotation curves cannot be measured far enough out to determine the extent of the dark matter halo, we have to use some other approach to fix the mass density of the dark matter in the universe. Here we discuss one method which allows us to measure the total (luminous and dark) mass in a system of galaxies (binaries, small groups, and large clusters of galaxies), that are bound together by their mutual gravitational attraction. This involves measurements of the mean-square of the galactic velocities $\langle v^2 \rangle$ and the average galactic inverse separation $\langle s^{-1} \rangle$ of the luminous components of the system. These two quantities, according to the **virial theorem** of statistical mechanics, $-2\langle T \rangle = \langle V \rangle$, relating the average kinetic and potential energy, are proportional to each other—with the proportional constant given by the total gravitational mass M (luminous and dark) of the system,

$$\langle v^2 \rangle = G_{\rm N} M \left\langle \frac{1}{s} \right\rangle . \tag{9.25}$$

The proof of this theorem is left as an exercise (Problem 9.6). Here we shall merely illustrate it with a simple example. Consider a two-body system (M, m), with $M \gg m$, separated by distance s. The Newtonian equation of motion $G_{\rm N} Mm/s^2 = mv^2/s$ immediately yields the result in (9.25). From such considerations,[18] one obtains a total mass density that is something like 50 times larger than the luminous matter. Thus the luminous matter, being what we can see when looking out into space, represents only a tiny fraction of the mass content of the universe.

There are now several independent means to determine the mass density at the present era $\Omega_{\rm M,0}$: one approach is through gravitational lensing by galaxies, and clusters of galaxies[19] (see Section 7.2); another is by comparing the number of galaxy clusters in galaxy superclusters throughout the cosmic age; yet another is from measured CMB temperature fluctuations. A value for the total mass density that is generally consistent with the above discussed results has been obtained (Tegmark *et al.* 2006):

$$\Omega_{\rm M,0} = 0.245 \pm 0.025. \tag{9.26}$$

We shall show in the next chapter that the whole universe is permeated with radiation. However, its energy density is considerably smaller so that $\Omega_{\rm R,0} \ll \Omega_{\rm M,0}$.

A historical note That there might be a significant amount of dark matter in the universe was first pointed out by Fritz Zwicky in the 1930s. The basis of this proposal is just the method we have outlined here. Zwicky noted that, given the observed radial velocities of the galaxies, the combined mass of the visible stars and gases in the Coma Cluster was simply not enough to hold

[18] While the argument involving the virial theorem may appear to be somewhat abstract, its result can be understood crudely as saying that the constituents of a system held gravitationally cannot move too fast so as to exceed the escape velocity $v_{\rm esc}^2 = 2G_{\rm N}M/r$. For a review of the simple concept of escape velocity, see Eq. (10.15) in Section 10.1.2.

[19] After subtracting out the peaks corresponding to the stars and galaxies from the mass distribution as deduced by gravitational lensing (e.g. Fig. 7.5), one is still left with a huge smooth bulged piece that can only be accounted for by the existence of dark matter and optically nonluminous gas.

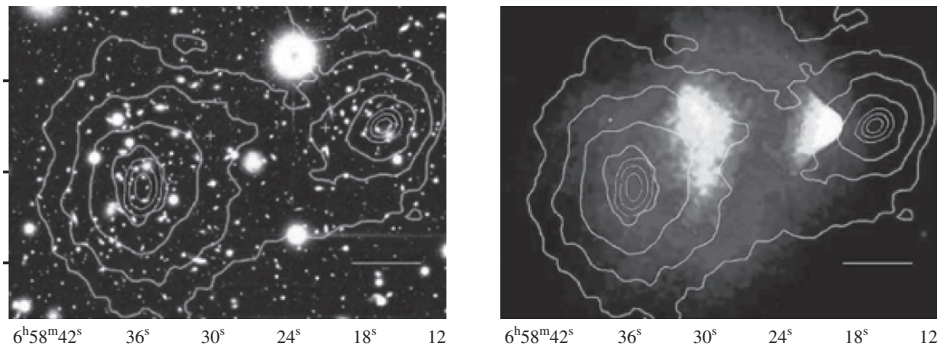

Fig. 9.6 Images of the "Bullet Cluster" 1E0657−558 from Clowe, *et al*. (2006) showing it as having been produced by the collision of two galactic clusters, resulting in the separation of hot gas and dark matter (with their embedded stars and galaxies). (a) Optical image from the Hubble Space Telescope; (b) X-ray image from the Chandra telescope. Mass density contours from gravitational lensing reconstruction, showing two mass peaks separated from the hot gaseous regions.

them together gravitationally. That is, what is holding together a galaxy or a cluster of galaxies must be some form of dark matter. The modern era began in 1970 when Vera Rubin and W. Kent Ford, using more sensitive techniques, were able to extend the rotation curve measurements far beyond the visible edge of gravitating systems of galaxies and clusters of galaxies.

Bullet Cluster offers direct empirical evidence of dark matter

In all the above discussions, the presence of dark matter was deduced through its gravitational effects (finding total $\Omega_{\rm M} > \Omega_{\rm B}$). On might wonder whether it is possible to explain the observation, instead of postulating the existence of a new form of matter, by modifying the law of gravity. Here we present a piece of evidence for dark matter that simply cannot be evaded by this alternative explanation. This is the observational result shown in Fig. 9.6. Three images of the galaxy cluster 1E0657-558, the "Bullet Cluster," are displayed here. The picture on the left shows galaxies that make up a few percent of the cluster mass; the picture on the right is the X-ray image from the Chandra telescope showing where the bulk of the hot gas is located. Superimposed on top of these two pictures is the mass contours as derived from gravitational lensing. These counters have two mass peaks which, while they more or less track the locations of observed galaxies, are situated at very different positions from the atomic gas. Such an observation cannot be explained by any modified law of gravity but is consistent with the interpretation that this Bullet Cluster came about because of a collision of two clusters of galaxies. The dark matter and baryonic gas are separated because the dark matter (having small interaction cross-section) passes through[20] "like a bullet" while the baryonic gases are left behind.

[20]Most of the galxies track the deep dark matter gravitational potentials.

Matter densities in the universe: A summary Dark matter is mostly nonrelativistic particles having only gravitational and weak interactions. It does not emit or absorb electromagnetic radiation. Its presence has been deduced from the velocity distribution of a gravitationally bound system. The most direct empirical evidence is the mass distribution in the Bullet Cluster. On an even larger scale the abundance of dark matter can be

quantified from the study of large cosmic structure and CMB. All this leads to a total mass density equal approximately to a quarter of the critical density:

$$\Omega_{\mathrm{M}} = \Omega_{\mathrm{B}} + \Omega_{\mathrm{DM}} \simeq 0.25. \tag{9.27}$$

The total baryonic (atomic) density can be deduced from the observed amount of light nuclear elements and the big bang nucleosynthesis theory or from the observed CMB temperature anisotropy:

$$\Omega_{\mathrm{B}} \simeq 0.04, \tag{9.28}$$

The bulk of which is in the intergalactic medium and has been detected through its electromagnetic absorption lines. What we can see optically, stars and galaxies, is only a small part of this baryonic density:

$$\Omega_{\mathrm{B}} = \Omega_{\mathrm{gas}} + \Omega_{\mathrm{lum}} \quad \text{with} \quad \Omega_{\mathrm{gas}} \gg \Omega_{\mathrm{lum}} \simeq 0.005. \tag{9.29}$$

Thus the luminous matter associated with the stars we see in galaxies represents about 2% of the total mass content. Most of the matter is dark ($\Omega_{\mathrm{DM}} \simeq 0.21$) composed mostly of exotic nonrelativistic particles such as WIMPs. The exact nature of these exotic nonbaryonic CDM particles remains one of the unsolved problems in physics.

9.3 The cosmological principle

That the universe is homogeneous and isotropic on the largest scale of hundreds of Mpc has been confirmed by direct observation only recently (cf. the discussion at the end of Section 9.1.2). Other evidence for its homogeneity and isotropy came in the form of the extremely uniform CMB radiation. This is the relic thermal radiation left over from an early epoch when the universe was only 10^5 years old. The nonuniformity of CMB is on the order of 10^{-5} (cf. Sections 10.5 and 11.3.1). This shows that the "baby universe" can be described as being highly homogeneous and isotropic.

But long before obtaining such direct observational evidence, Einstein had adopted the **strategy** of starting the study of cosmology with a basic working hypothesis called **the cosmological principle** (CP): at each epoch (i.e. each fixed value of cosmological time t) the universe is homogeneous and isotropic. It presents the **same** aspects (except for local irregularities) from each point: the universe has no center and no edge.

- This statement that there is no privileged location in the universe (hence homogeneous and isotropic) is sometimes referred to as the **Copernican cosmological principle**. It is in essence the ultimate generalization of the Copernican principle.
- This is a priori the most reasonable assumption, as it is difficult to think of any other alternative. Also, in practice, it is also the most "useful," as it involves the least number of parameters. There is some chance for the theory to be predictive. Its correctness can then be checked by observation. Thus CP was invoked in the study of cosmology long before there was any direct evidence for a homogeneous and isotropic universe, but it is now fully supported by observation.

- The observed irregularities, i.e. the structure, in the universe (stars, galaxies, clusters of galaxies, superclusters, voids, etc.) are assumed to arise because of gravitational clumping around some initial density unevenness. Various mechanisms for seeding such density perturbation have been explored. Most of the efforts have been concentrated around the idea that, in the earliest moments, the universe passed through a phase of extraordinarily rapid expansion, the "**cosmic inflationary epoch**." The small quantum fluctuations were inflated to astrophysical size and they seeded the cosmological density perturbation (cf. Sections 11.2.3 and 11.3.1).

The cosmological principle gives rise to a picture of the universe as a physical system of a "cosmic fluid." The fundamental particles of this fluid are galaxies, and a fluid element has a volume that contains many galaxies, yet is small compared to the whole system of the universe. Thus, the motion of a cosmic fluid element is the smeared-out motion of the constituent galaxies. It is determined by the gravitational interaction of the entire system—the self-gravity of the universe. This means that each element is in free-fall; all elements follow geodesic worldlines. (In reality, the random motions of the galaxies are small, on the order of 10^{-3}.)

Such a picture of the universe allows us to pick a privileged coordinate frame, the **comoving coordinate system**, where

$t \equiv$ the proper time of each fluid element

$x^i \equiv$ the spatial coordinates carried by each fluid element.

A comoving observer flows with a cosmic fluid element. The comoving coordinate time can be synchronized over the whole system. For example, t is inversely proportional to the temperature of the cosmic background radiation (see Section 10.3) which decreases monotonically. Thus, we can in principle determine the cosmic time by a measurement of the background radiation temperature. This property allows us to define space-like slices, each with a fixed value of the coordinate time, and each is homogenous and isotropic.

Because each fluid element carries its own position label the comoving coordinate is also the cosmic rest frame—as each fluid element's position coordinates are unchanged with time. But we must remember that in GR the coordinates do not measure distance, which is a combination of the coordinates and the metric. As we shall detail below, viewed in this comoving coordinate, the expanding universe, with all galaxies rushing away from each other, is described not by changing position coordinates, but by an ever-increasing metric. This emphasizes the physics underlying an expanding universe not as something exploding in the space, but as the expansion of space itself.

9.4 The Robertson–Walker spacetime

9.4.1 The metric in the comoving coordinate system

The cosmological principle says that, at a fixed cosmic time, each space-like slice of the spacetime is homogeneous and isotropic. In Section 7.1

spherical symmetry has been found to restrict the metric to the form of $g_{\mu\nu} = \text{diag}(g_{00},\ g_{rr},\ r^2,\ r^2\sin^2\theta)$ with only two scalar functions (g_{00} and g_{rr}). In this section we discuss the geometry resulting from the cosmological principle: when expressed in comoving coordinates, it has a Robertson–Walker metric.

The time components

Because the coordinate time is the proper time of fluid elements, we must have $g_{00} = -1$. The fact that the spacelike slices for fixed t can be defined means that the spatial axes are orthogonal to the time axes:[21]

$$g_{00} = -1 \quad \text{and} \quad g_{0i} = g_{i0} = 0. \tag{9.30}$$

The self-consistency of this choice of coordinates can be checked as follows. A particle at rest in the comoving frame is a particle in free fall under the mutual gravity of the system; it should follow a geodesic worldline obeying Eq. (6.9):

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0. \tag{9.31}$$

Being at rest, $dx^i = 0$ with $i = 1, 2, 3$, we only need to calculate the Christoffel symbol Γ^μ_{00}. But the metric properties of (9.30) imply that $\Gamma^\mu_{00} = 0$ for all μ. Thus these fluid elements at rest with respect to the comoving frame ($dx^i/d\tau = d^2x^i/d\tau^2 = 0$) do satisfy (trivially) the geodesic equation.

The metric for a 3D space with constant curvature

Let g_{ij} be the spatial part of the metric; the relations in (9.30) imply that the 4D metric that satisfies the cosmological principle is block-diagonal:

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 \\ 0 & g_{ij} \end{pmatrix}. \tag{9.32}$$

The invariant interval expressed in terms of the comoving coordinates is

$$\begin{aligned} ds^2 &= -c^2dt^2 + g_{ij}dx^i dx^j \\ &\equiv -c^2dt^2 + dl^2. \end{aligned} \tag{9.33}$$

Because of the CP requirement (i.e. no preferred direction and position), the time dependence in g_{ij} must be an **overall** length factor $R(t)$, sometimes referred to as the radius of the universe, with no dependence on any of the indices:

$$dl^2 = R^2(t)d\tilde{l}^2 \tag{9.34}$$

where the reduced length element $d\tilde{l}$ is both t-independent and dimensionless. It is also useful to define a dimensionless **scale factor**

$$a(t) \equiv \frac{R(t)}{R_0}, \tag{9.35}$$

[21] Because fixed-time space-like slices of space exist, we can consider an event as separated from two other events in two distinctive ways. The first connects the event to another on a space-like space containing all events with the same cosmic time: $da^\mu = (0, dx^i)$ for a definite spatial index i; the second is an interval connecting the event to another one along the worldline of a comoving observer: $db^\mu = (dt, 0)$. The inner product of these two intervals $g_{\mu\nu}da^\mu db^\nu = g_{i0}dx^i dt$ (the repeated μ and ν indices are summed, but not the i indices) is an invariant, valid in any coordinate system including the local Minkowski frame. This makes it clear that the left-hand side vanishes. The above equality then implies $g_{i0} = 0$.

with denominator on the right-hand side (RHS) $R_0 \equiv R(t_0)$ so that the scale factor is normalized at the present epoch by $a(t_0) = 1$.

One has the picture of the universe as a three-dimensional (3D) map with cosmic fluid elements labeled by the fixed comoving coordinates $\tilde{x}_i$. Time evolution enters entirely through the time dependence of the map scale $R(t) = a(t)R_0$, see Fig. 9.7,

$$x_i(t) = a(t)R_0\tilde{x}_i \tag{9.36}$$

with $\tilde{x}_i$ being the fixed (t-independent) dimensionless map coordinates, while $a(t)$ is the size of the grids and is independent of the map coordinates. As the universe expands, the **relative distance** relations (i.e. the shape of things) are not changed.

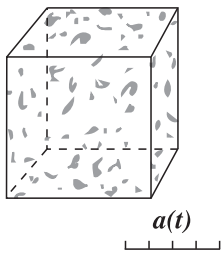

Fig. 9.7 A three-dimensional map of the cosmic fluid with elements labeled by t-independent $\tilde{x}_i$ comoving coordinates. The time dependence of any distance is entirely determined by the t-dependent scale factor.

The Robertson–Walker metric

The Robertson–Walker (RW) metric is for a spacetime which, at a give time, has a 3D homogeneous and isotropic space. One naturally expects this space to have a constant curvature. In Section 5.3.2 we have already written down the metric[22] for the 3D spaces with constant curvature in two spherical coordinate systems, with the dimensionless radial coordinates being $\chi = r/R_0$ and $\xi = \rho/R_0$, respectively, and the differential solid angle $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$:

- Equation (5.53) for the comoving "polar" coordinates (χ, θ, ϕ):

$$dl^2 = R_0^2 a^2(t) d\chi^2 + k^{-1}(\sin^2\sqrt{k}\chi)d\Omega^2. \tag{9.37}$$

- Equation (5.55) for the comoving "cylindrical" coordinates (ξ, θ, ϕ)

$$dl^2 = R_0^2 a^2(t)\left(\frac{d\xi^2}{1-k\xi^2} + \xi^2 d\Omega^2\right). \tag{9.38}$$

The parameter k in g_{ij} can take on the values $\pm 1, 0$ with $k = +1$ for a 3-sphere, $k = -1$ for a 3-pseudosphere, and $k = 0$ for a 3D Euclidean (flat) space. Some properties of such spaces, such as their embedding and their volume evaluation, were also discussed in Problems 5.7 and 5.8. In the context of cosmology, the universe having a $k = +1$ positively curved space is called a "**closed universe**," a $k = -1$ negatively curved space an "**open universe**," and $k = 0$ a "**flat universe**."

In practice, one can use either one of the two coordinates displayed in (9.37) and (9.38); they are equivalent. In the following, for definiteness, we shall work with the (ξ, θ, ϕ) coordinate system of (9.38).

9.4.2 Distances in the RW geometry

In an expanding universe with a space that may be curved, we must be very careful in any treatment of distance. In the following sections we shall deal with several kinds of distance, starting with conceptually the simplest: the proper distance.

[22] While the deduction of the 3D spatial metric given in Section 5.3.2 is only heuristic, in Section 14.4.1 we shall provide an independent derivation of the same result.

The proper distance

The proper distance $d_{\rm p}(\xi, t)$ to a point at the comoving radial distance ξ and cosmic time t can be calculated from the metric (9.38) with $d\Omega = 0$ and $dt = 0$:

$$d_{\rm p}(\xi, t) = a(t) R_0 \int_0^{\xi} \frac{d\xi'}{(1 - k\xi'^2)^{1/2}} \tag{9.39}$$

so that the time dependence (due to expansion of the universe) on the RHS is entirely contained in the scale factor $a(t)$

$$d_{\rm p}(\xi, t) = a(t) d_{\rm p}(\xi, t_0) \tag{9.40}$$

where the fixed (comoving) distance at the present epoch is

$$d_{\rm p}(\xi, t_0) = R_0 \int_0^{\xi} \frac{d\xi'}{(1 - k\xi'^2)^{1/2}} = \left(\frac{R_0}{\sqrt{k}}\right) \sin^{-1}(\sqrt{k}\xi). \tag{9.41}$$

Namely, for a space with positive curvature $k = +1$, we have $d_{\rm p}(\xi, t_0) = R_0 \sin^{-1} \xi$; negative curvature, $R_0 \sinh^{-1} \xi$, and a flat space $R_0 \xi = \rho$.

Hubble's law follows from CP The relation (9.40) implies a proper velocity of

$$v_{\rm p}(t) = \frac{d(d_{\rm p})}{dt} = \frac{\dot{a}(t)}{a(t)} d_{\rm p}(t). \tag{9.42}$$

Evidently the velocity is proportional to the separation. This is just Hubble's law with the Hubble constant expressed in terms of the scale factor:

$$H(t) = \frac{\dot{a}(t)}{a(t)} \quad \text{and} \quad H_0 = \dot{a}(t_0). \tag{9.43}$$

Recall that the appearance of an overall scale factor in the spatial part of the Robertson–Walker metric follows from our imposition of the homogeneity and isotropy condition. The result in (9.42) confirms our expectation that in any geometrical description of a dynamical universe which satisfies the cosmological principle, hence the distance scaling relation (9.40), Hubble's law emerges automatically. We emphasize that, in the GR framework, the expansion of the universe is described as the expansion of space, and "big bang" is not any sort of "explosion of matter in space," but rather it is an "expansion of space itself." Space is a dynamic quantity, which is expanding; that is, the metric function of spacetime is the solution to Einstein equation and its scale factor increases with time.

Relating distance to the scale factor at emission To relate distance to the redshift of a light source located at the comoving distance $\xi_{\rm em}$, we use the fact that the observer and emitter are connected by a light ray along a radial path ($d\Omega = 0$),

$$ds^2 = -c^2 dt^2 + R_0^2 a^2(t) \frac{d\xi^2}{1 - k\xi^2} = 0. \tag{9.44}$$

Moving c^2dt^2 to one side and taking the minus sign for the square-root for incoming light, we have

$$-\int_{t_0}^{t_{\rm em}} \frac{cdt}{a(t)} = R_0 \int_0^{\xi_{\rm em}} \frac{d\xi}{(1-k\xi^2)^{1/2}} = d_{\rm p}(\xi_{\rm em}, t_0) \tag{9.45}$$

where (9.41) has been used to express the second integral in terms of the proper distance at $t = t_0$. The first integral can be put into a more useful form by changing the integration variable to the scale factor,

$$-\int_{t_0}^{t_{\rm em}} \frac{cdt}{a(t)} = -\int_1^{a_{\rm em}} \frac{cda}{a(t)\dot{a}(t)} = -\int_1^{a_{\rm em}} \frac{cda}{a^2(t)H(t)}, \tag{9.46}$$

where we used (9.43) to reach the last expression. In this way (9.45) becomes the relation tbetween the proper distance and scale factor at the emission time

$$d_{\rm p}(\xi_{\rm em}, t_0) = -\int_1^{a_{\rm em}} \frac{cda}{a^2 H(a)}. \tag{9.47}$$

Once again, this is the distance between us and the light emitter located at comoving radial coordinate $\xi_{\rm em}$ with light emitted when the scale factor was $a_{\rm em}$.

Redshift and the scale factor We see that the scale factor $a(t)$ is the key quantity in our description of the time evolution of the universe. In fact, because $a(t)$ is generally a monotonic function, it can serve as a kind of cosmic clock. How can the scale factor be measured? The observable quantity that has the simplest relation to $a(t)$ is the wavelength shift of a light signal.

The spectral shift, according to (9.3), is

$$z = \frac{\Delta\lambda}{\lambda} = \frac{\lambda_{\rm rec}}{\lambda_{\rm em}} - 1. \tag{9.48}$$

We expect that the wavelength (in fact any length) scales as $a(t)$ (see Problem 9.8 for a more detailed justification):

$$\frac{\lambda_{\rm rec}}{\lambda_{\rm em}} = \frac{a(t_{\rm rec})}{a(t_{\rm em})}. \tag{9.49}$$

Since the "received time" is at t_0 with $a(t_0) = 1$, we have the basic relation

$$1 + z = \frac{1}{a(t_{\rm em})}. \tag{9.50}$$

For example, at the redshift of $z = 1$, the universe had a linear size half as large as at the present one. In fact a common practice in cosmology is to refer to "the redshift of an era" instead of its cosmic time. For example, the "photon decoupling time," when the universe became transparent to light (cf. Section 10.5), is said to occur at $z = 1100$, etc.

Distance in terms of redshift Changing the integration variable in (9.47) to the redshift, we have the relation between proper distance and redshift in the Robertson–Walker spacetime:

$$d_{\rm p}(z) = \int_0^z \frac{c\,dz'}{H(z')}. \tag{9.51}$$

The functional dependence of distance on the redshift is, of course, the Hubble relation. Different cosmological models having a Hubble constant with different z dependence would yield a different distance–redshift relation. Thus the Hubble curve can be used to distinguish between different cosmological scenarios. As we shall discuss in the next chapter, our universe has been discovered to be in an accelerating expansion phase. By fitting the Hubble curve we shall deduce that the universe's dominant energy component is some unknown "dark energy," which provides the repulsion in causing the expansion to proceed at an ever faster rate.

Luminosity distance and standard candle

The principal approach in calculating the distance to any stellar object is to estimate its true luminosity and compare that with the observed flux (which is reduced by the squared distance). Thus it is important to have stars with known intrinsic luminosity that can be used to gauge astronomical distances. Stars with luminosity that can be deduced from other properties are called "standard candles." A well-known class of standard candles is the Cepheid variable stars, which have a definite correlation between their intrinsic luminosity and their pulse rates. In fact, Edwin Hubble used Cepheids to deduce the distances of the galaxies collected for his distance vs. redshift plot. Clearly, the reliability of the method depends on one's ability to obtain the correct estimate of the intrinsic luminosity. A famous piece of history is that Hubble underestimated the luminosity of his Cepheids by almost a factor of 50, leading to an underestimation of the distances, hence an overestimate of the Hubble constant H_0 by a factor of seven. This caused a "cosmic age problem" because the resultant Hubble time (which should be comparable to the age of the universe) became much shorter than the estimated ages of many objects in the universe. This was corrected only after many years of further astronomical observation and astrophysical modeling. Here, we assume that the intrinsic luminosity of a standard candle can be reliably obtained.

In this section, we study the distance that can be obtained by measuring the light flux from a remote light source with known luminosity. Because we use observations of light emitted in the distant past of an evolving universe, this requires us to be attentive in dealing with the concept of time.

The measured flux of watts per unit area is related to the intrinsic luminosity $\mathfrak{L}$, which is the total power-radiated by the emitting object, as

$$f \equiv \frac{\mathfrak{L}}{4\pi d_{\rm L}^2}. \tag{9.52}$$

This defines the luminosity distance $d_{\rm L}$. Let us note that in space with any constant curvature the area of a "sphere" is given by $4\pi d_{\rm p}^2$ where $d_{\rm p}$ is the

proper radial distance, cf. (9.41). In a static universe, the luminosity distance equals the proper distance to the source: $d_{\rm p} = d_{\rm L}$.

$$f_{\rm (static)} = \frac{\mathfrak{L}}{4\pi d_{\rm p}^2}. \tag{9.53}$$

In an expanding universe the observed flux, being proportional to the energy transfer per unit time, is reduced by a factor of $(1+z)^2$: one power of $(1+z)$ comes from energy reduction due to wavelength lengthening of the emitted light, and another power due to the increasing time interval. Let us explain: The energy being proportional to frequency ω, the emitted energy, compared to the observed one, is given by the ratio,

$$\frac{\omega_{\rm em}}{\omega_0} = \frac{\lambda_0}{\lambda_{\rm em}} = \frac{1}{a(t_{\rm em})} = 1+z, \tag{9.54}$$

where we have used $a(t_0) = 1$ and (9.49) and (9.50). Just as frequency is reduced by $\omega_0 = \omega_{\rm em}(1+z)^{-1}$, the time interval must be correspondingly increased by $\delta t_0 = \delta t_{\rm em}(1+z)$, leading to a reduction of energy transfer rate by another power of $(1+z)$:

$$\frac{\omega_0}{\delta t_0} = \frac{\omega_{\rm em}}{\delta t_{\rm em}}(1+z)^{-2}. \tag{9.55}$$

Thus the observed flux in an expanding universe, in contrast to the static universe result of (9.53), is given by

$$f = \frac{\mathfrak{L}}{4\pi d_{\rm p}^2 (1+z)^2}. \tag{9.56}$$

Namely, the luminosity distance (9.52) differs from the proper distance by

$$d_{\rm L} = d_{\rm p}(1+z). \tag{9.57}$$

In Chapter 10 the cosmological equations will be solved to obtain the epoch-dependent Hubble's constant in terms of the energy/mass content of the universe. In this way we can find how the proper distance $d_{\rm p}$ (thus also the luminosity distance) depends on the redshift z via (9.51) for the general relation. (Problem 9.11 works out the case of small z.) In Box 9.1 we explain the astronomy practice of plotting the Hubble diagrams of redshift vs. **distance modulus** (instead of luminosity distance), which is effectively the logarithmic luminosity distance.

Box 9.1 Logarithmic luminosity and distance modulus

Ancient Greek astronomers classified the brightness (observed flux) of stars as having "first magnitude" to "sixth magnitude" for the brightest to the faintest stars visible to the naked eye—the brighter a star is, the smaller its magnitude. Since for this magnitude range of $m_{(6)} - m_{(1)} = 5$ the apparent luminosities span roughly a factor of 100, namely, $f_{(1)}/f_{(6)} \simeq 100$,

(*cont.*)

Box 9.1 (*Continued*)

a definition of **apparent magnitude** m is suggested:

$$m \equiv -2.5 \log_{10} \frac{f}{f_0} \tag{9.58}$$

so that $m_{(6)} - m_{(1)} = 2.5 \log_{10} f_{(1)}/f_{(6)} = 5$. The reference flux is taken to be $f_0 \equiv 2.52 \times 10^{-8} W\ m^{-2}$ so that the brightest visible stars correspond to $m = 1$ objects. In this scale, for comparison, the sun has an apparent magnitude $m_\odot = -26.8$.

Similar to (9.58), we can define a logarithmic scale, called **absolute magnitude,** for the intrinsic luminosity of a star:

$$M \equiv -2.5 \log_{10} \frac{\mathfrak{L}}{\mathfrak{L}_0} \tag{9.59}$$

where the reference luminosity $\mathfrak{L}_0$ is defined so that a star with this power output will be seen at a distance 10 pc away to have a flux f_0:

$$f_0 = \frac{\mathfrak{L}_0}{4\pi\ (10\ \text{pc})^2}. \tag{9.60}$$

This works out to be $\mathfrak{L}_0 = 78.7 \mathfrak{L}_\odot$. Using the definition of luminosity distance as given in (9.52), Eq. (9.60) can be translated into an expression for the luminosity ratio

$$\frac{f}{\mathfrak{L}} = \frac{f_0}{\mathfrak{L}_0} \left(\frac{10\ \text{pc}}{d_\text{L}} \right)^2. \tag{9.61}$$

Taking the logarithm of this equation leads to the definition of **distance modulus** $(m - M)$, which can be related to luminosity distance by taking the difference of (9.58) and (9.59) and substituting in (9.61):

$$m - M = 5 \log_{10} \frac{d_\text{L}}{10\ \text{pc}}. \tag{9.62}$$

In the astronomy literature one finds the common practice of plotting the Hubble diagram with one axis being the redshift z and another axis, instead of luminosity distance, its logarithmic function, the distance modulus (e.g. Fig. 11.8 and Fig. 11.11)

Review questions

1. What does it mean that Hubble's law is a linear relation? What is the significance of this linearity? Support your statement with a proof.
2. What is the Hubble time t_H? Under what condition is it equal to the age of the universe t_0? In a universe full of matter and energy, what would be the expected relative

magnitude of these two quantities ($t_{\rm H} > t_0$ or $t_{\rm H} < t_0$)? What is the lower bound for t_0 deduced from the observation data on globular clusters?

3. What are "galaxy rotation curves?" What feature would we expect if the luminous matter were a good representation of the total mass distribution? What observational feature of the rotation curve told us that there were significant amounts of nonluminous matter associated with galaxies and clusters of galaxies?
4. Give a simple example that illustrates the content of the virial theorem for a gravitational system. How can this be used to estimate the total mass of the system?
5. What is baryonic matter? The bulk of baryonic matter resides in the intergalactic medium (IGM) and does not shine. Why don't we count it a part of dark matter?
6. What are the values that we have for the total mass density $\Omega_{\rm M}$, for the luminous matter $\Omega_{\rm lum}$, and for the baryonic matter $\Omega_{\rm B}$? From this deduce an estimate of the dark matter density $\Omega_{\rm DM}$. All values are for the present epoch, and list them only to one significant figure.
7. What is the cosmological principle? What are comoving coordinates?
8. Write out the form of the Robertson–Walker metric for two possible coordinate systems. What is the input (i.e. the assumption) used in the derivation of this metric?
9. What are the physical meanings of the scale factor $a(t)$ and the parameter k in the Robertson–Walker metric? How is the epoch-dependent Hubble constant $H(t)$ related to the scale factor $a(t)$?
10. What is the scaling behavior of wavelength? From this derive the relation between the scale factor $a(t)$ and the redshift z.
11. Derive the integral expression for the proper distance $d_p = c\int H^{-1}dz$ to the light source with redshift z.
12. What is luminosity distance? How is it related to the proper distance?

Problems

9.1 **The universe as a strong gravitational system** One can check that the universe as a whole corresponds to a system of strong gravity that requires a GR description by making a crude estimate of the parameter ε in Eq. (9.1). For this calculation you can assume a static Euclidean universe having a finite spherical volume with radius given by a horizon length cH_0^{-1} and having a mass density comparable to the critical density as given in (9.14).

9.2 **Luminosity distance to the nearest star** The nearest star appears to us to have a brightness $f_* \simeq 10^{-11} f_\odot$ ($f_\odot$ being the observed solar flux). Assuming that it has the same intrinsic luminosity as the sun, estimate the distance d_* to this star, in the distance unit of **parsec**, as well as in the **astronomical unit** AU $\simeq 5 \times 10^{-6}$pc.

9.3 **Gravitational frequency shift contribution to the Hubble redshift** Hubble's linear plot of redshift vs. distance relies on spectral measurement of galaxies beyond the Local Group with redshift $z \gtrsim 0.01$. A photon emitted by a galaxy suffers not only a redshift because of cosmic recession, but also a gravitational redshift. Is the latter a significant factor when compared to the recessional effect? *Suggestion*: compare the gravitational redshifts of light from a galaxy with mass $M_{\rm G} = O(10^{11} M_\odot)$ and linear dimension $R_{\rm G} = O(10^{12} R_\odot)$ to the redshift for light leaving the surface of the sun, $z_\odot = O(10^{-6})$.

9.4 **Energy content due to starlight** By assuming the stars have been shining with the same intensity since the beginning of the universe and always had the luminosity density as given in (9.19), estimate the density ratio $\Omega_* = \rho_*/\rho_c$ for starlight. For this rough calculation you can take the age of universe to be the Hubble time t_H.

9.5 **Night sky as bright as day** Olbers' paradox is solved in our expanding universe because the age of the universe is not infinite $t_0 \simeq t_H$ and, having a horizon length $\simeq ct_H$, it is effectively not infinite in extent. Given the present luminosity density of (9.19), with the same approximation as Problem 9.4, estimate the total flux due to starlight. Compare your result with the solar flux $f_\odot = \mathfrak{L}_\odot/(4\pi\ (\text{AU})^2)$. We can increase the star light flux by increasing the age of the universe t_0. How much older does the universe have to be in order that the night sky is as bright as day?

9.6 **The virial theorem** Given a general bound system of mass points (located at $\mathbf{r}_n$) subject to gravitational forces (central and inverse square) $\mathbf{F}_n = -\nabla V_n$ with $V_n \propto r_n^{-1}$, by considering the time derivative, and average, of the sum of dot-products of momentum and position $G \equiv \Sigma_n \mathbf{p}_n \cdot \mathbf{r}_n$ (called the **virial**), show that the time-averages of the kinetic and potential energy are related by $2\langle T\rangle = -\langle V\rangle$.

9.7 **Proper distance from comoving coordinate χ** In the text we worked out the proper distance from a point with radial coordinate ξ as in (9.40). Now perform the same calculation (and obtain a similar result) for a point labeled by the alternative radial coordinate χ with a metric given by (9.37).

9.8 **Wavelength in an expanding universe** By a careful consideration of the time interval between emission and reception of two successive wavecrests, prove that in an expanding universe with a scale factor $a(t)$, the wavelength scales as expected:

$$\frac{\lambda_{\text{rec}}}{\lambda_{\text{em}}} = \frac{a(t_{\text{rec}})}{a(t_{\text{em}})}.$$

Suggestion: Apply Eq. (9.45) to two successive emissions/receptions of waves.

9.9 **The deceleration parameter and Taylor expansion of the scale factor** Display the Taylor expansion of the scale factor $a(t)$ and $[a(t)]^{-1}$ around $t = t_0$, up to $(t - t_0)^2$, in terms of the Hubble's constant H_0 and the **deceleration parameter** defined by

$$q_0 \equiv \frac{-\ddot{a}(t_0)\, a(t_0)}{\dot{a}^2(t_0)}. \tag{9.63}$$

9.10 **The steady-state universe** The conventional interpretation of an ever increasing scale factor (expanding universe) means that all objects must have been closer in the past, leading to a big bang beginning. We also mentioned in Section 9.4.2 that, because of an initial overestimate of the Hubble constant (by a factor of seven), there was a "cosmic age problem." To avoid this difficulty, an alternative cosmology, called the **steady-state universe (SSU)**, was proposed by Hermann Bondi, Thomas Gold, and Fred Hoyle. It was suggested that, consistent with the Robertson–Walker description of an expanding universe, all cosmological quantities besides the scale factor (the expansion rate, deceleration parameter, spatial curvature, matter density, etc.) are time independent. A constant mass density means that the universe did not have a big hot beginning; hence there cannot be a cosmic age problem. To have a constant mass density in an expanding universe requires the continuous, energy-nonconserving, creation of matter. To SSU's advocates, this spontaneous mass creation is no more peculiar than the creation of all matter at the instant of big bang. In fact, the name "big bang" was invented by Fred Hoyle as a somewhat disparaging description of the competing cosmology.

(a) Supporters of SSU find this model attractive on theoretical grounds—because it is compatible with the "**perfect cosmological principle**." From the above outline of SSU and the cosmological principle in Section 9.3, can you infer what this "perfect CP" must be?

(b) RW geometry, hence (9.43), also holds for SSU, but with a constant expansion rate $H(t) = H_0$. From this, deduce the explicit t-dependence of the scale factor $a(t)$. What is the SSU prediction for the deceleration parameter q_0 defined in (9.63)?

(c) SSU has a 3D space with a curvature $K = k/R^2$ that is not only constant in space but also in time. Does this extra requirement fix its spatial geometry? If so, what is it?

(d) Since the matter density is a constant $\rho_M(t) = \rho_{M,0} \simeq 0.3\rho_{c,0}$ and yet the scale factor increases with time, SSU requires spontaneous matter creation. What must be the rate of this mass creation per unit volume? Express it in terms of the number of hydrogen atoms created per cubic kilometer per year.

9.11 **z^2 correction to the Hubble relation** The Hubble relation (9.5) is valid only in the low velocity limit. Namely, it is the leading term in the power series expansion of the proper distance in terms of the redshift. Use the definition of deceleration parameter introduced in (9.63) to show that, including the next order, the Hubble relation reads as

$$d_p(t_0) = \frac{cz}{H_0}\left(1 - \frac{1 + q_0}{2} z\right). \tag{9.64}$$

(a) One first uses (9.45) to calculate the proper distance up to the quadratic term in the "look-back time" $(t_0 - t_{\text{em}})$.

(b) Use the Taylor series of Problem 9.10 to express the redshift in terms of the look-back time up to $(t_0 - t_{\text{em}})^{2.}$

(c) Deduce the claimed result of (9.64) by using the result obtained in (a) and inverting the relation between the redshift and look-back time obtained in (b).