© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

1

Introduction

A **stellar system** is a gravitationally bound assembly of stars or other point masses. Stellar systems vary over more than fourteen orders of magnitude in size and mass, from binary stars, to star clusters containing 10^2 to 10^6 stars, through galaxies containing 10^5 to 10^{12} stars, to vast clusters containing thousands of galaxies.

The behavior of these systems is determined by Newton's laws of motion and Newton's law of gravity,[1] and the study of this behavior is the branch of theoretical physics called **stellar dynamics**. Stellar dynamics is directly related to at least three other areas of theoretical physics. Superficially, it is closest to celestial mechanics, the theory of planetary motions—both involve the study of orbits in a gravitational field—however, much of the formalism of celestial mechanics is of little use in stellar dynamics, since it is based on perturbation expansions that do not converge when applied to most stellar systems. The most fundamental connections of stellar dynamics are with classical statistical mechanics, since the number of stars in a star cluster or galaxy is often so large that a statistical treatment of the dynamics is necessary. Finally, many of the mathematical tools that have been developed to study stellar systems are borrowed from plasma physics, which also involves the study of large numbers of particles interacting via long-range forces.

[1] As yet, there is no direct evidence for stellar systems in which relativistic effects are important, although such systems are likely to be present at the centers of galaxies.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For an initial orientation, it is useful to summarize a few orders of magnitude for a typical stellar system, the one to which we belong. Our Sun is located in a stellar system called the **Milky Way** or simply **the Galaxy**. The Galaxy contains four principal constituents:

(i) There are about 10^{11} stars, having a total mass $\simeq 5\times10^{10}$ solar masses (written $5\times10^{10}\,\mathcal{M}_\odot$; $1\,\mathcal{M}_\odot = 1.99\times10^{30}\,\mathrm{kg}$).[2] Most of the stars in the Galaxy travel on nearly circular orbits in a thin disk whose radius is roughly 10^4 parsecs (1 parsec $\equiv 1\,\mathrm{pc} \equiv 3.086\times10^{16}\,\mathrm{m}$), or 10 kiloparsecs (kpc). The thickness of the disk is roughly 0.5 kpc and the Sun is located near its midplane, about 8 kpc from the center.

(ii) The disk also contains gas, mostly atomic and molecular hydrogen, concentrated into clouds with a wide range of masses and sizes, as well as small solid particles ("dust"), which render interstellar gas opaque at visible wavelengths over distances of several kpc. Most of the atomic hydrogen is neutral rather than ionized, and so is denoted **HI**. Together, the gas and dust are called the **interstellar medium** (ISM). The total ISM mass is only about 10% of the mass in stars, so the ISM has little direct influence on the dynamics of the Galaxy. However, it plays a central role in the chemistry of galaxies, since dense gas clouds are the sites of star formation, while dying stars eject chemically enriched material back into the interstellar gas. The nuclei of the atoms in our bodies were assembled in stars that were widely distributed through the Galaxy.

(iii) At the center of the disk is a black hole, of mass $\simeq 4\times10^6\,\mathcal{M}_\odot$. The black hole is sometimes called Sagittarius A* or Sgr A*, after the radio source that is believed to mark its position, which in turn is named after the constellation in which it is found.

(iv) By far the largest component, both in size and mass, is the **dark halo**, which has a radius of about 200 kpc and a mass of about $10^{12}\,\mathcal{M}_\odot$ (both these values are quite uncertain). The dark halo is probably composed of some weakly interacting elementary particle that has yet to be detected in the laboratory. For most purposes, the halo interacts with the other components of the Galaxy only through the gravitational force that it exerts, and hence stellar dynamics is one of the few tools we have to study this mysterious yet crucial constituent of the universe.

The typical speed of a star on a circular orbit in the disk is about $200\,\mathrm{km\,s^{-1}}$. It is worth remembering that $1\,\mathrm{km\,s^{-1}}$ is almost exactly 1 pc (actually 1.023) in 1 megayear (1 megayear $\equiv 1\,\mathrm{Myr} = 10^6$ years). Thus the time required to complete one orbit at the solar radius of 8 kpc is 250 Myr. Since the age of the Galaxy is about 10 gigayears (1 gigayear $\equiv 1\,\mathrm{Gyr} = 10^9\,\mathrm{yr}$), most disk stars have completed over forty revolutions, and it is reasonable to assume that the Galaxy is now in an approximately steady state. The steady-state

[2] See Appendix A for a tabulation of physical and astronomical constants, and Tables 1.1, 1.2 and 2.3 for more precise descriptions of the properties of the Galaxy.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

approximation allows us to decouple the questions of the present-day *equilibrium* and *structure* of the Galaxy, to which most of this book is devoted, from the thornier issue of the *formation* of the Galaxy, which we discuss only in the last chapter of this book.

Since the orbital period of stars near the Sun is several million times longer than the history of accurate astronomical observations, we are forced to base our investigation of Galactic structure on what amounts to an instantaneous snapshot of the system. To a limited extent, the snapshot can be supplemented by measurements of the angular velocities (or **proper motions**) of stars that are so close that their position on the sky has changed noticeably over the last few years; and by **line-of-sight velocities** of stars, measured from Doppler shifts in their spectra. Thus the *positions* and *velocities* of some stars can be determined, but their *accelerations* are almost always undetectable with current observational techniques.

Using the rough values for the dimensions of the Galaxy given above, we can estimate the mean free path of a star between collisions with another star. For an assembly of particles moving on straight-line orbits, the mean free path is $\lambda = 1/(n\sigma)$, where n is the number density and σ is the cross-section. Let us make the crude assumption that all stars are like the Sun so the cross-section for collision is $\sigma = \pi(2\,R_\odot)^2$, where $R_\odot = 6.96 \times 10^8\,\mathrm{m} = 2.26 \times 10^{-8}\,\mathrm{pc}$ is the solar radius.[3] If we spread 10^{11} stars uniformly over a disk of radius 10 kpc and thickness 0.5 kpc, then the number density of stars in the disk is $0.6\,\mathrm{pc}^{-3}$ and the mean free path is $\lambda \simeq 2\times10^{14}\,\mathrm{pc}$. The interval between collisions is approximately λ/v, where v is the random velocity of stars at a given location. Near the Sun, the random velocities of stars are typically about $50\,\mathrm{km\,s}^{-1}$. With this velocity, the collision interval is about $5\times10^{18}\,\mathrm{yr}$, over 10^8 times longer than the age of the Galaxy. Evidently, near the Sun collisions between stars are so rare that they are irrelevant—which is fortunate, since the passage of a star within even 10^3 solar radii would have disastrous consequences for life on Earth. For similar reasons, hydrodynamic interactions between the stars and the interstellar gas have a negligible effect on stellar orbits.

Thus, each star's motion is determined solely by the gravitational attraction of the mass in the galaxy—other stars, gas, and dark matter. Since the motions of weakly interacting dark-matter particles are also determined by gravitational forces alone, the tools that we develop in this book are equally applicable to both stars and dark matter, despite the difference of 70 or more orders of magnitude in mass.

We show in §1.2 that a useful first approximation for the gravitational field in a galaxy is obtained by imagining that the mass is continuously distributed, rather than concentrated into discrete mass points (the stars

[3] This calculation neglects the enhancement in collision cross-section due to the mutual gravitational attraction of the passing stars, but this increases the collision rate by a factor of less than 100, and hence does not affect our conclusion. See equation (7.195).

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

and dark-matter particles) and clouds (the gas). Thus we begin Chapter 2 with a description of Newtonian potential theory, developing methods to describe the smoothed gravitational fields of stellar systems having a variety of shapes. In Chapter 3 we develop both quantitative and qualitative tools to describe the behavior of particle orbits in gravitational fields. In Chapter 4 we study the statistical mechanics of large numbers of orbiting particles to find equilibrium distributions of stars in phase space that match the observed properties of galaxies, and learn how to use observations of galaxies to infer the properties of the underlying gravitational field.

The models constructed in Chapter 4 are **stationary**, that is, the density at each point is constant in time because the rates of arrival and departure of stars in every volume element balance exactly. Stationary models are appropriate to describe a galaxy that is many revolutions old and hence presumably in a steady state. However, some stationary systems are unstable, in that the smallest perturbation causes the system to evolve to some quite different configuration. Such systems cannot be found in nature. Chapter 5 studies the stability of stellar systems.

In Chapter 6 we describe some of the complex phenomena that are peculiar to galactic disks. These include the beautiful spiral patterns that are usually seen in disk galaxies; the prominent bar-like structures seen at the centers of about half of all disks; and the warps that are present in many spiral galaxies, including our own.

Even though stellar collisions are extremely rare, the gravitational fields of passing stars exert a series of small tugs that slowly randomize the orbits of stars. Gravitational encounters of this kind in a stellar system are analogous to collisions of molecules in a gas or Brownian motion of small particles in a fluid—all these processes drive the system towards energy equipartition and a thermally relaxed state. Relaxation by gravitational encounters operates so slowly that it can generally be neglected in galaxies, except very close to their centers (see §1.2); however, this process plays a central role in determining the evolution and present form of many star clusters. Chapter 7 describes the kinetic theory of stellar systems, that is, the study of the evolution of stellar systems towards thermodynamic equilibrium as a result of gravitational encounters. The results can be directly applied to observations of star clusters in our Galaxy, and also have implications for the evolution of clusters of galaxies and the centers of galaxies.

Chapter 8 is devoted to the interplay between stellar systems. We describe the physics of collisions and mergers of galaxies, and the influence of the surrounding galaxy on the evolution of smaller stellar systems orbiting within it, through such processes as dynamical friction, tidal stripping, and shock heating. We also study the effect of irregularities in the galactic gravitational field—generated, for example, by gas clouds or spiral arms—on the orbits of disk stars.

Throughout much of the twentieth century, galaxies were regarded as "island universes"—distinct stellar systems occupying secluded positions in

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

space. Explicitly or implicitly, they were seen as isolated, permanent structures, each a dynamical and chemical *unit* that was formed in the distant past and did not interact with its neighbors. A major conceptual revolution in **extragalactic astronomy**—the study of the universe beyond the edges of our own Galaxy—was the recognition in the 1970s that this view is incorrect. We now believe in a model of **hierarchical galaxy formation**, the main features of which are that: (i) encounters and mergers of galaxies play a central role in their evolution, and in fact galaxies are formed by the mergers of smaller galaxies; (ii) even apparently isolated galaxies are surrounded by much larger dark halos whose outermost tendrils are linked to the halos of neighboring galaxies; (iii) gas, stars, and dark matter are being accreted onto galaxies up to the present time. A summary of the modern view of galaxy formation and its cosmological context is in Chapter 9.

1.1 An overview of the observations

1.1.1 Stars

The luminosity of the Sun is $L_\odot = 3.84 \times 10^{26}\,\mathrm{W}$. More precisely, this is the **bolometric luminosity**, the total rate of energy output integrated over all wavelengths. The bolometric luminosity is difficult to determine accurately, in part because the Earth's atmosphere is opaque at most wavelengths. Hence astronomical luminosities are usually measured in one or more specified wavelength bands, such as the **blue** or B band centered on $\lambda = 450\,\mathrm{nm}$; the **visual** or V band at $\lambda = 550\,\mathrm{nm}$; the R band at $\lambda = 660\,\mathrm{nm}$; the near-infrared I band at $\lambda = 810\,\mathrm{nm}$; and the infrared K band centered on the relatively transparent atmospheric window at $\lambda = 2200\,\mathrm{nm} = 2.2\,\mu\mathrm{m}$, all with width $\Delta\lambda/\lambda \simeq 0.2$ (see Binney & Merrifield 1998, §2.3; hereafter this book is abbreviated as BM). For example, the brightest star in the sky, Sirius, has luminosities

$$L_V = 22\,L_{\odot V} \quad ; \quad L_R = 15\,L_{\odot R}, \tag{1.1}$$

while the nearest star, Proxima Centauri, has luminosities

$$L_V = 5.2 \times 10^{-5}\,L_{\odot V} \quad ; \quad L_R = 1.7 \times 10^{-4}\,L_{\odot R}. \tag{1.2}$$

This notation is usually simplified by dropping the subscript from $L_\odot$, when the band to which it refers is clear from the context.

Luminosities are often expressed in a logarithmic scale, by defining the **absolute magnitude**

$$M \equiv -2.5 \log_{10} L + \text{constant}. \tag{1.3}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

The constant is chosen separately and arbitrarily for each wavelength band. The solar absolute magnitude is

$$M_{\odot B} = 5.48 \quad ; \quad M_{\odot V} = 4.83 \quad ; \quad M_{\odot R} = 4.42. \tag{1.4}$$

Sirius has absolute magnitude $M_V = 1.46$, $M_R = 1.47$, and Proxima Centauri has $M_V = 15.5$, $M_R = 13.9$. The **flux** from a star of luminosity L at distance d is $f = L/(4\pi d^2)$, and a logarithmic measure of the flux is provided by the **apparent magnitude**

$$m \equiv M + 5\log_{10}(d/10\,\mathrm{pc}) = -2.5\log_{10}\left[L(10\,\mathrm{pc}/d)^2\right] + \mathrm{constant}; \tag{1.5}$$

thus, the absolute magnitude is the apparent magnitude that the star would have if it were at a distance of 10 parsecs. Note that *faint* stars have *large* magnitudes. Sirius is at a distance[4] of $(2.64 \pm 0.01)\,\mathrm{pc}$ and has apparent magnitude $m_V = -1.43$, while Proxima Centauri is at $(1.295 \pm 0.004)\,\mathrm{pc}$ and has apparent magnitude $m_V = 11.1$. The faintest stars visible to the naked eye have $m_V \simeq 6$, and the limiting magnitude of the deepest astronomical images at this time is $m_V \simeq 29$. The apparent magnitudes m_V and m_R are often abbreviated simply as V and R.

The **distance modulus** $m - M = 5\log_{10}(d/10\,\mathrm{pc})$ is often used as a measure of distance.

The **color** of a star is measured by the ratio of the luminosity in two wavelength bands, for example by L_R/L_V or equivalently by $M_V - M_R = m_V - m_R = V - R$. Sirius has color $V - R = -0.01$ and Proxima Centauri has $V - R = 1.67$. Stellar spectra are approximately black-body and hence the color is a measure of the temperature at the surface of the star.

A more precise measure of the surface temperature is the **effective temperature** $T_{\rm eff}$, defined as the temperature of the black body with the same radius and bolometric luminosity as the star in question. If the stellar radius is R, then the Stefan–Boltzmann law implies that the bolometric luminosity is

$$L = 4\pi R^2 \sigma T_{\rm eff}^4, \tag{1.6}$$

where $\sigma = 5.670 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$. The relation between color and effective temperature is tabulated in BM §3.4 and shown in Figure 1.1.

A third measure of the surface temperature of a star is its **spectral class**, which is assigned on the basis of the prominence of various absorption lines in the stellar spectrum. In order of decreasing temperature, the spectral classes are labeled O, B, A, F, G, K, M, L, and T, and each class is divided into ten subclasses by the numbers $0, 1, \ldots, 9$. Thus a B0 star is slightly

[4] Throughout this book, quoted errors are all 1 standard deviation, i.e., there is a probability of 0.68 that the actual error is less than the quoted error. The reader is warned that astronomical errors are often dominated by unknown systematic effects and thus tend to be underestimated.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

cooler than an O9 star. Using this scheme, experienced observers can determine the effective temperature of a star to within about 10% from a quick examination of its spectrum. For example, Sirius has spectral class A1 and effective temperature 9500 K, while Proxima Centauri has spectral class M5 and $T_{\rm eff} = 3000\,{\rm K}$. The Sun is a quite ordinary G2 star, with $T_{\rm eff} = 5780\,{\rm K}$.

The ultraviolet emission from the hottest stars ionizes nearby interstellar gas, forming a sphere of ionized hydrogen called an **HII region** (BM §8.1.3). The brightest star-like objects in other galaxies are often HII regions shining in emission lines, rather than normal stars shining from thermal emission (see, for example, Plate 1).

The **color-magnitude** diagram is a plot of absolute magnitude against color; since color is related to effective temperature, each point on this diagram corresponds to a unique luminosity, effective temperature, and stellar radius (through eq. 1.6). In older work spectral type sometimes replaces color, since the two quantities are closely related, and in this case the plot is called a **Hertzsprung–Russell** or **HR** diagram (BM §3.5). This simple diagram has proved to be the primary point of contact between observations and the theory of stellar structure and evolution.

The distribution of stars in the color-magnitude diagram depends on the age and chemical composition of the sample of stars plotted. Astronomers refer to all elements beyond helium in the periodic table as "metals"; with the exception of lithium, such elements are believed to be formed in stars rather than at the birth of the universe (see §1.3.5). To a first approximation the abundances of groups of elements vary in lockstep since they are formed in the same reaction chain and injected into interstellar space by the same type of star. At an even cruder level the chemical composition of a star can be approximately specified by a single number Z, the **metallicity**, which is the fraction by mass of all elements heavier than helium. Similarly, the fractions by mass of hydrogen and helium are denoted X and Y ($X + Y + Z = 1$). The Sun's initial composition was $X_\odot = 0.71$, $Y_\odot = 0.27$, $Z_\odot = 0.019$.

Figure 1.1 shows the color-magnitude diagram for about 10^4 nearby stars. The most prominent feature is the well-defined band stretching from $(B-V, M_V) \simeq (0, 0)$ to $(B-V, M_V) \simeq (1.5, 11)$. This band, known as the **main sequence**, contains stars that are burning hydrogen in their cores. In this stage of a star's life, the mass—and to a lesser extent, chemical composition—uniquely determine both the effective temperature and the luminosity, so stars remain in a fixed position on the color-magnitude diagram. Main-sequence stars are sometimes called **dwarf** stars, to distinguish them from the larger giant stars that we discuss below. The main sequence is a mass sequence, with more massive stars at the upper left (high luminosity, high temperature, blue color) and less massive stars at the lower right (low luminosity, low temperature, red color). The most and least luminous main-sequence stars in this figure, with absolute magnitudes $M \simeq -2$ and $+12$, have masses of about $10\,\mathcal{M}_\odot$ and $0.2\,\mathcal{M}_\odot$, respectively (BM Table 3.13). Objects smaller than about $0.08\,\mathcal{M}_\odot$ never ignite hydrogen in their cores, and

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

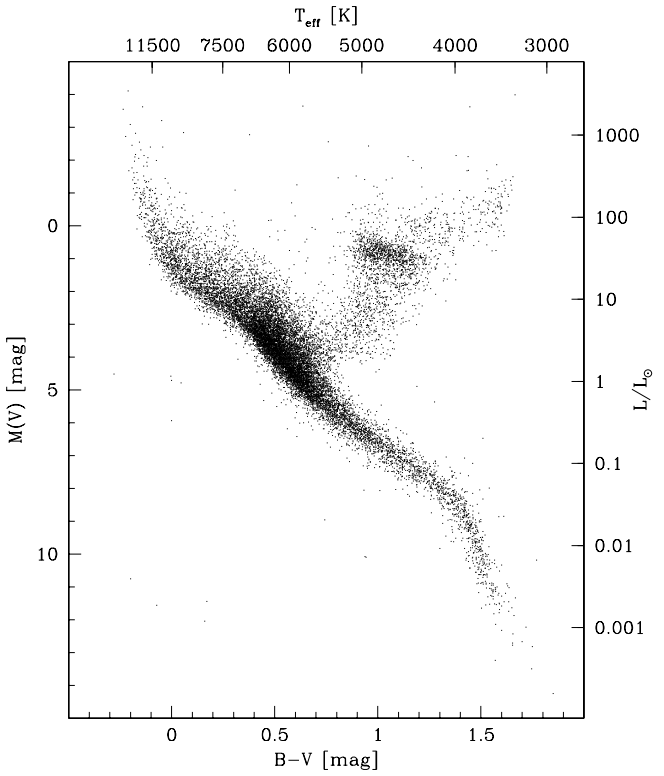

Figure 1.1 The color-magnitude diagram for over 10^4 nearby stars. Close binary stars have been excluded, since the companion contaminates the color and magnitude measurements. Most of the stars fall on the main sequence, which runs from upper left to lower right. The red-giant branch runs upward and to the right from the main sequence. The red-clump stars form a prominent concentration in the middle of the red-giant branch. The giants are chosen from a larger volume than the other stars to enhance their numbers and thereby show the structure of the giant branch more clearly. The absolute magnitudes are in the V band and the colors are based on $B-V$, which gives the ratio of fluxes between $\lambda = 450\,\mathrm{nm}$ and $550\,\mathrm{nm}$ (BM Table 2.1). The right axis shows the V-band luminosity in solar units, and the top axis shows approximate values of the effective temperature, which follows from the color and luminosity because the stars are approximate black bodies. From Perryman et al. (1995).

are visible mainly from the radiation they emit as they contract and cool. The luminosity of these objects, known as **brown dwarfs**, therefore depends on both mass and age and they do not form a one-parameter sequence like their more massive siblings.

The massive, high-luminosity stars at the upper end of the main sequence exhaust their fuel rapidly and hence are short-lived ($\lesssim 100\,\mathrm{Myr}$ for stars with $M_V = -2$), while stars at the lower end of the main sequence

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

burn steadily for much longer than the current age of the universe. The Sun has a lifetime of 10 Gyr on the main sequence.

From equation (1.6), stars that are luminous and cool (upper right of the color-magnitude diagram) must be large, while stars that are dim and hot (lower left of the diagram) must be small. Since the main sequence crosses from upper left to lower right, this argument suggests that radius is not a strong function of luminosity along the main sequence. The mass-radius relation in Table 3.13 of BM bears this out: between $M \simeq 0$ and $M \simeq 10$, a factor of 10^4 in luminosity, the radius varies by only a factor of six, from $3\,R_\odot$ to $0.5\,R_\odot$.

The color-magnitude diagram contains a handful of dim blue stars around $(B-V, M) \simeq (0, 12)$. These are **white dwarfs**, stars that have exhausted their nuclear fuel and are gradually cooling to invisibility. As their location in the diagram suggests, white dwarfs are very small, with radii of order $10^{-2}\,R_\odot$. White dwarfs are so dense that the electron gas in the interior of the star is degenerate; in other words, gravitational contraction is resisted, not by thermal pressure as in main-sequence stars, but rather by the Fermi energy of the star's cold, degenerate electron gas.

Figure 1.1 also contains a prominent branch slanting up and to the right from the main sequence, from $(B-V, M) \simeq (0.3, 4)$ to $(B-V, M) \simeq (1.5, -1)$. These are **red giants**, stars that have exhausted hydrogen in their cores and are now burning hydrogen in a shell surrounding an inert helium core. As their location in the color-magnitude diagram suggests—red therefore cool, yet very luminous—red giants are much larger than main-sequence stars; the stars at the tip of the giant branch have radii $\gtrsim 100\,R_\odot$. In contrast to the main sequence, on which stars remain in a fixed position determined by their mass, the red-giant branch is an evolutionary sequence: stars climb the giant branch from the main sequence to its tip, over an interval of about 1 Gyr for stars like the Sun.

The prominent concentration in the middle of the red-giant branch, near $(B-V, M) \simeq (1, 1)$, is called the **red clump**. This feature arises from a later evolutionary stage, which happens to coincide with the red-giant branch for stars of solar metallicity. Red-clump stars have already ascended the giant branch to its tip and returned, settling at the red clump when they begin burning helium in their cores (see below).

Red giants are rare compared to main-sequence stars because the red-giant phase in a star's life is much shorter than its main-sequence phase. Nevertheless, red giants are so luminous that they dominate the total luminosity of many stellar systems. Another consequence of their high luminosity is that a far larger fraction of red giants is found in flux-limited samples than in volume-limited samples. For example, over half of the 100 brightest stars are giants, but none of the 100 nearest stars is a giant.

Figure 1.2 illustrates the color-magnitude diagram of a typical globular star cluster (§1.1.4). The advantage of studying a star cluster is that all of its members lie at almost the same distance, and have the same age and

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

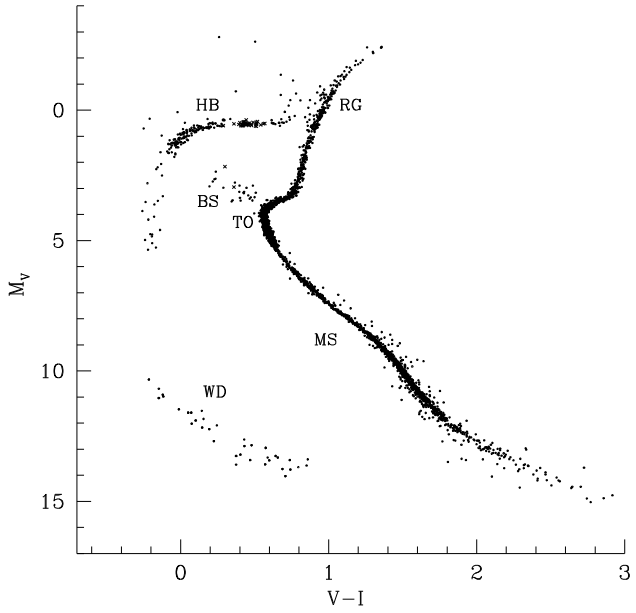

Figure 1.2 The color-magnitude diagram for metal-poor globular clusters. The horizontal axis is the color $V-I$. Labels denote the main sequence (MS), the main-sequence turnoff (TO), red giants (RG), horizontal branch (HB), blue stragglers (BS), and white dwarfs (WD). This is a composite diagram in which data from five globular clusters have been combined selectively to emphasize the principal sequences; thus the relative numbers of different types of stars are not realistic. From data supplied by W. E. Harris; see also Harris (2003).

chemical composition. Thus age and composition differences and distance errors do not blur the diagram, so the sequences are much sharper than in Figure 1.1 (BM §6.1.2). In this diagram the main sequence stretches from $(V-I, M_V) \simeq (0.6, 4)$ to $(V-I, M_V) \simeq (2.4, 14)$. In contrast to Figure 1.1, the main sequence terminates sharply at $M_V \simeq 4$ (the **turnoff**); the more luminous, bluer part of the main sequence that is seen in the sample of nearby stars is absent in the cluster, because such stars have lifetimes shorter than the cluster age. Figure 1.2 shows a few stars situated along the extrapolation of the main sequence past the turnoff point; these "blue stragglers" may arise from collisions and mergers of stars in the dense core of the cluster or mass transfer between the components of a binary star (page 628). The white dwarfs are visible near $(V-I, M_V) \simeq (0.4, 13)$, and the tip of the red-giant branch lies at $(V-I, M_V) \simeq (1.3, -2)$.

As a star evolves, it climbs the giant branch until, at the tip of the giant branch, helium starts to burn in its core. The stars then evolve rapidly

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

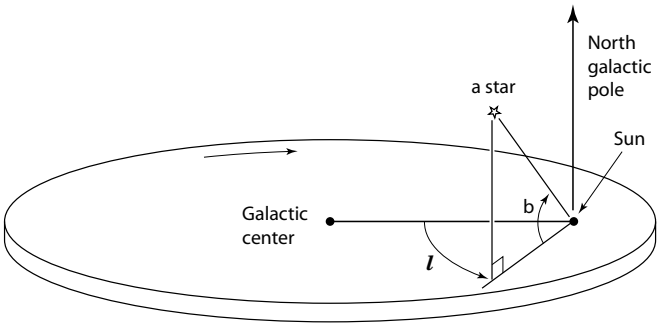

Figure 1.3 A schematic picture of the Sun's location in the Galaxy, illustrating the Galactic coordinate system. An arrow points in the direction of Galactic rotation, which is clockwise as viewed from the north Galactic pole.

to the **horizontal branch** (the sequence of stars near $M_V \simeq 0$, stretching from $V - I \simeq 0$ to 0.8), where they remain until the helium in the core is exhausted. The form of the horizontal branch depends on the metallicity: as the metallicity increases from very low values ($Z \simeq 0.01 Z_\odot$) the horizontal branch shortens and moves to the right, until at near-solar metallicity it is truncated to the red clump seen in Figure 1.1.

Stars in metal-poor globular clusters are among the oldest objects in the Galaxy. Fits of theoretical models of stellar evolution to the color-magnitude diagrams of metal-poor globular clusters yield ages of (12.5 ± 1.5) Gyr (Krauss & Chaboyer 2003). This result is consistent with the age of the universe determined from measurements of the cosmic background radiation, $t_0 = (13.7 \pm 0.2)$ Gyr (eq. 1.77), if the globular clusters formed when the universe was about 1 Gyr old.

1.1.2 The Galaxy

Most of the stars in the Galaxy lie in a flattened, roughly axisymmetric structure known as the **Galactic disk**. On clear, dark nights the cumulative light from the myriad of faint disk stars is visible as a luminous band stretching across the sky, which is the source of the name "Milky Way" for our Galaxy. The midplane of this disk is called the **Galactic plane** and serves as the equator of **Galactic coordinates** (ℓ, b), where ℓ is the **Galactic longitude** and b is the **Galactic latitude** (BM §2.1.2). The Galactic coordinate system is a heliocentric system in which $\ell = 0, b = 0$ points to the Galactic center and $b = \pm 90^\circ$ points to the **Galactic poles**, normal to the disk plane (see Figure 1.3).

The Sun is located at a distance R_0 from the center of the Galaxy; the best current estimate $R_0 = (8.0 \pm 0.5)$ kpc comes from the orbits of stars near the black hole that is believed to mark the center (see §1.1.6 and Eisenhauer

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

et al. 2003). One measure of the distribution of stars in the Galactic disk is the surface brightness, the total stellar luminosity emitted per unit area of the disk (see Box 2.1 for a more precise definition). Observations of other disk galaxies suggest that the surface brightness is approximately an exponential function of radius,

$$I(R) = I_{\rm d} \exp(-R/R_{\rm d}). \tag{1.7}$$

The **disk scale length** $R_{\rm d}$ is difficult to measure in our Galaxy because of our position within the disk. Current estimates place $R_{\rm d}$ between about 2 and 3 kpc. Thus the Sun lies farther from the Galactic center than about 75–90% of the disk stars. The resulting concentration of luminosity towards the Galactic center is not apparent to the naked eye, since interstellar dust absorbs the light from distant disk stars (the optical depth in the V band along a line of sight in the Galactic midplane is unity at a distance of only about 0.7 kpc); however, in the infrared, where dust extinction is unimportant, our position near the edge of the disk is immediately apparent from the strong concentration of light in the direction of the constellation Sagittarius (at the center of the image in Plate 2). By contrast, the Galaxy is nearly transparent in the direction of the Galactic poles, which greatly facilitates studying the extragalactic universe.

The stars of the disk travel in nearly circular orbits around the Galactic center. The speed of a star in a circular orbit of radius R in the Galactic equator is denoted $v_{\rm c}(R)$, and a plot of $v_{\rm c}(R)$ versus R is called the **circular-speed curve**. The circular speed at the solar radius R_0 is

$$v_0 \equiv v_{\rm c}(R_0) = (220 \pm 20)\,{\rm km\,s^{-1}}. \tag{1.8}$$

A strong additional constraint on v_0 and R_0 comes from the angular motion of the radio source Sgr A* relative to extragalactic sources: if Sgr A* coincides with the black hole at the Galactic center, and if this black hole is at rest in the Galaxy—both very plausible assumptions, but not certainties—then the angular speed of the Sun is $v_0/R_0 = (236 \pm 1)\,{\rm km\,s^{-1}}/(8\,{\rm kpc})$ (Reid & Brunthaler 2004).

The **Local Standard of Rest** (LSR) is an inertial reference frame centered on the Sun and traveling at speed v_0 in the direction of Galactic rotation. Since most nearby disk stars are on nearly circular orbits, their velocities relative to the Local Standard of Rest are much smaller than v_0. For example, the Sun's velocity relative to the LSR (the **solar motion**) is (BM §10.3.1)

$$13.4\,{\rm km\,s^{-1}} \text{ in the direction } \ell = 28^\circ,\ b = 32^\circ. \tag{1.9}$$

The root-mean-square (RMS) velocity of old disk stars relative to the Local Standard of Rest is $50\,{\rm km\,s^{-1}}$, larger than the Sun's velocity but still small compared to the circular speed v_0.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

In the direction perpendicular to the Galactic plane (usually called the "vertical" direction), the density of stars falls off exponentially,

$$\rho(R,z) = \rho(R,0)\mathrm{e}^{-|z|/z_{\mathrm{d}}(R)}, \tag{1.10}$$

where z is the distance from the midplane and $z_{\mathrm{d}}(R)$ is the **scale height** at radius R.[5] The thickness z_{d} of the Galactic disk depends on the age of the stars that are being examined. Older stellar populations have larger scale heights, probably because stochastic gravitational fields due to spiral arms and molecular clouds gradually pump up the random velocities of stars (see §8.4). In the solar neighborhood, the scale height ranges from $\lesssim 100\,\mathrm{pc}$ for the young O and B stars to $\simeq 300\,\mathrm{pc}$ for the stars with ages of order 10 Gyr that constitute the bulk of the disk mass.

A more accurate representation of the vertical structure of the disk is obtained by superimposing two populations with densities described by equation (1.10): the **thin disk** with $z_{\mathrm{d}} \simeq 300\,\mathrm{pc}$, and the **thick disk** with $z_{\mathrm{d}} \simeq 1\,\mathrm{kpc}$ (BM Figure 10.25). The stars of the thick disk are older and have a different chemical composition from those of the thin disk—thick-disk stars have lower metallicities, and at a given metallicity they have higher abundances of the α nuclides (^{16}O, ^{20}Ne, ^{24}Mg, ^{28}Si, etc.; see BM §5.2.1 and Figure 10.17) relative to ^{56}Fe. The surface density of the thick disk is about 7% of that of the thin disk, so in the midplane, thin-disk stars outnumber thick-disk stars by about 50:1. The thick disk was probably created when the infant thin disk was shaken and thickened by an encounter with a smaller galaxy early in its history.

The enhanced α nuclides found in the thick disk are the signature of stars formed early in the history of the disk, for the following reason. The interstellar gas is polluted with heavy elements by two main processes: (i) "core-collapse" supernovae, arising from the catastrophic gravitational collapse of massive stars, which lag star formation by no more than $\sim 40\,\mathrm{Myr}$, and produce ejecta that are rich in α nuclides; (ii) "thermonuclear" or Type Ia supernovae, which are caused by runaway nuclear burning on the surface of white-dwarf stars in binary systems, lag star formation by of order 0.5–10 Gyr, and produce mostly nuclei near ^{56}Fe. Thus the chemical composition of thick-disk stars suggests that the thick disk formed in less than about 1 Gyr. In contrast, it appears that stars in the thin disk have formed at a steady rate throughout the lifetime of the Galaxy.

Throughout this book, we shall distinguish the **kinematics** of a stellar system—the observational description of the positions and motions of the stars in the system—from its dynamics—the interpretation of these motions in terms of physical laws (forces, masses, etc.). Thus, the description of the

[5] This formula has a discontinuous slope at $z = 0$, which reflects the gravitational attraction of the much thinner gas layer on the stars. The vertical distribution of stars in a thin disk is explored theoretically in Problem 4.22.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Galaxy in this subsection has so far been kinematic. The simplest approximate dynamical description of the Galaxy is obtained by assuming that its mass distribution is spherical. Let the mass interior to radius r be $M(r)$. From Newton's theorems (§2.2.1) the gravitational acceleration at radius r is equal to that of a point whose mass is the same as the total mass interior to r; thus the inward acceleration is $GM(r)/r^2$, where the **gravitational constant** $G = 6.674 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$. The central or **centripetal** acceleration required to hold a body in a circular orbit with speed v_0 is v_0^2/r. Thus the mass interior to the solar radius R_0 in this crude model is

$$M(R_0) = \frac{v_0^2 R_0}{G} = 9.0 \times 10^{10}\,\mathcal{M}_\odot \left(\frac{v_0}{220\,\mathrm{km\,s^{-1}}}\right)^2 \left(\frac{R_0}{8\,\mathrm{kpc}}\right). \qquad (1.11)$$

The approximation that the mass distribution is spherical is reasonable for the dark halo, but not for the flat stellar disk. Better models suggest that this estimate is probably high by about 30%, since a disk requires less mass to produce a given centripetal acceleration (see Figure 2.17).

Most of our understanding of stellar astrophysics comes from observations of stars within a few hundred parsecs of the Sun. This distance is much smaller than the disk scale length, and hence it is reasonable to assume that the distribution of properties of these stars (chemical compositions, ages, masses, kinematics, fraction of binary stars, etc.) is constant within this region, even though there may be large-scale gradients in these properties across the Galactic disk. To formalize this assumption, we define the **solar neighborhood** to be a volume centered on the Sun that is much smaller than the overall size of the Galaxy but large enough to contain a statistically useful sample of stars. The concept is somewhat imprecise but nevertheless extremely useful. The appropriate size of the volume depends on which stars we wish to investigate: for white dwarfs, which are both common and dim, the "solar neighborhood" may consist of a sphere of radius only 30 pc centered on the Sun, while for the luminous but rare O and B stars, the solar neighborhood may be considered to extend as far as 1–2 kpc from the Sun.

Our best estimate of the inventory of the solar neighborhood is summarized in Table 1.1. The category "visible stars" includes all main-sequence and giant stars. The category "stellar remnants" includes white dwarfs and neutron stars, while "ISM" (interstellar medium) includes atomic and molecular hydrogen, ionized gas, and a small contribution from interstellar dust. The volume density and luminosity density are quoted in the Galactic midplane and the surface density and surface brightness are integrated over a column perpendicular to the Galactic plane, extending to ± 1.1 kpc from the midplane. "Dynamical" denotes determinations of the total volume or surface density from the dynamics of disk stars (see §4.9.3). The dynamically determined volume density in the midplane is consistent with the observed density in stars and gas to within about 10%, so there is no evidence for a significant component of dark matter in the disk—in other words, the inventory in Table 1.1 appears to be complete. The dynamically determined

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Table 1.1 Inventory of the solar neighborhood

component	volume density ($\mathcal{M}_\odot\,\mathrm{pc}^{-3}$)	surface density ($\mathcal{M}_\odot\,\mathrm{pc}^{-2}$)	luminosity density ($L_\odot\,\mathrm{pc}^{-3}$)	surface brightness ($L_\odot\,\mathrm{pc}^{-2}$)
visible stars	0.033	29	0.05	29
stellar remnants	0.006	5	0	0
brown dwarfs	0.002	2	0	0
ISM	0.050	13	0	0
total	0.09 ± 0.01	49 ± 6	0.05	29
dynamical	0.10 ± 0.01	74 ± 6	–	–

NOTES: Volume and luminosity densities are measured in the Galactic midplane and surface density is the total within ± 1.1 kpc of the plane. Luminosity density and surface brightness are given in the R band. Dynamical estimates are from §4.9.3. Most other entries are taken from Flynn et al. (2006).

surface density appears to be higher than the surface density in stars and gas, by $(25 \pm 9)\,\mathcal{M}_\odot\,\mathrm{pc}^{-2}$; if significant, this excess probably represents the contribution of the dark halo. The dark halo also contributes to the volume density in the midplane, but this contribution is undetectably small.

A stellar system is often characterized by its **mass-to-light ratio**, which we denote by Υ and write in units of the solar ratio, $\Upsilon_\odot = \mathcal{M}_\odot / L_\odot$. According to Table 1.1, the mass-to-light ratio of the solar neighborhood in the R band is $\Upsilon_R \simeq 2\,\Upsilon_\odot$ in the midplane and $\simeq 2.5\,\Upsilon_\odot$ after integrating to ± 1.1 kpc from the plane. The second value is higher because the scale height z_d of luminous young stars is smaller than that of older, dimmer stars.

In addition to the disk, the Galaxy contains a **bulge**, a small, amorphous, centrally located stellar system that is thicker than the disk and comprises $\sim 15\%$ of the total luminosity. The Galactic bulge is clearly visible at the center of the disk in infrared images of the Galaxy (Plate 2). The evolutionary history, kinematics, and chemical composition of bulge stars are quite different from those of disk stars near the Sun. The bulge stars are believed to date from near the time of formation of the Galaxy, whereas the disk stars have a wide range of ages, since star formation in the disk is an ongoing process. While disk stars in the solar neighborhood are found in nearly circular orbits with speeds $v_\mathrm{c}(R) \simeq 220\,\mathrm{km\,s}^{-1}$ and RMS velocity relative to this speed of only $50\,\mathrm{km\,s}^{-1}$, the velocity vectors of bulge stars are randomly oriented, with RMS velocity $\simeq 150\,\mathrm{km\,s}^{-1}$. The bulge stars exhibit a wide range of metallicities, spread around a median metallicity of about $0.4 Z_\odot$ (Zoccali et al. 2003), substantially smaller than the metallicity of young stars in the solar neighborhood—presumably because the interstellar gas from which the local disk stars form has steadily become more and more polluted by the metal-rich debris of exploding supernovae.

By analogy to the statistical-mechanical concept that temperature is proportional to mean-square velocity, a stellar population like the disk in

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

which the random velocities are much smaller than the ordered or mean velocity is said to be **cool**, while the bulge population, in which the random velocities are larger than the mean velocity, is said to be **hot**. A hypothetical disk in which the stars move on precisely circular orbits would be **cold**.

Although the distribution of bulge stars is symmetric about the Galactic midplane, the bulge is somewhat brighter and thicker on one side of the Galactic center (longitude $\ell > 0$) than on the other. This asymmetry arises because the bulge is triaxial: the lengths of the two principal axes that lie in the Galactic plane are in the ratio 3:1, and the triaxial structure extends to about 3 kpc from the center. The long axis is oriented about 20° from the line between the Galactic center and the Sun (§2.7e). Thus the bulge is brighter and thicker at positive longitudes simply because that side is closer to the Sun. Because the bulge is triaxial it is also sometimes called a "bar" and the Milky Way is said to be a barred galaxy (see §1.1.3).

About 1% of the stellar mass in the Galaxy is contained in the **stellar halo**, which contains old stars of low metallicity (median about $0.02 Z_\odot$). The stellar halo has little or no mean rotation, and a density distribution that is approximately spherical and a power-law function of radius, $\rho \propto r^{-3}$, out to at least 50 kpc. The metal-poor globular clusters that we describe below (§1.1.4) are members of the stellar halo. The low metallicity of this population suggests that it was among the first components of the Galaxy to form. Much of the halo comprises the debris of disrupted stellar systems, such as globular clusters and small satellite galaxies.

The dark halo is the least well understood of the Galaxy's components. We have only weak constraints on its composition, shape, size, mass, and local density. A wide variety of candidates for the dark matter have been suggested, most falling into one of two broad classes: (i) some unknown elementary particle—the preferred candidates are WIMPs, an acronym for weakly interacting massive particles, but there are also more exotic possibilities such as axions; (ii) non-luminous macroscopic objects, such as neutron stars or black holes, which are usually called MACHOs, for massive compact halo objects. Measurements of the optical depth to gravitational lensing through the halo exclude MACHOs in the range 10^{-7}–$30\,\mathcal{M}_\odot$ as the dominant component of the dark halo (Alcock et al. 2001; Tisserand et al. 2007), and indirect dynamical arguments (§8.2.2e) suggest that more massive compact objects are also excluded. On the other hand, hypothetical massive, neutral, weakly interacting particles could be formed naturally in the early universe in approximately the numbers required to make a substantial contribution to the overall density. Thus most physicists and astronomers believe that the dark halo is probably composed of WIMPs. Ordinary matter—stars, dust, interstellar gas, MACHOs, etc., whether luminous or dark—derives almost all of its mass from baryons and hence is usually referred to as **baryonic matter** to distinguish it from **non-baryonic matter** such as WIMPs.[6]

[6] A baryon is a strongly interacting fermion. The word derives from barus, the Greek

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

The formation of flat astrophysical systems such as the solar system or a galaxy disk requires dissipation, which removes energy but conserves angular momentum and therefore leads naturally to a rapidly rotating thin disk. Since WIMPs cannot dissipate energy, the dark halo is expected to be approximately spherical. Numerical simulations of the formation of dark halos suggest that they are triaxial rather than precisely spherical, with minor-to-major axis ratios of 0.4–0.6, but there is little direct observational evidence on halo shapes (§9.3.3).

The total size and mass of the Galaxy's halo can be constrained by the kinematics of distant globular clusters and nearby galaxies. Using this method Wilkinson & Evans (1999) find a best-fit mass of $2\times10^{12}\,\mathcal{M}_\odot$, with a **median** or **half-mass radius** (the radius containing half the total mass) of 100 kpc; however, these values are very uncertain and masses as small as $2\times10^{11}\,\mathcal{M}_\odot$ or as large as $5\times10^{12}\,\mathcal{M}_\odot$ are allowed. A reasonable guess of the total mass of the Galaxy inside 100 kpc radius is

$$M(r<100\,\mathrm{kpc}) = 5\text{–}10\times10^{11}\,\mathcal{M}_\odot. \tag{1.12}$$

The mass distribution in the dark halo is equally uncertain at smaller radii: the halo contribution to the radial gravitational force at the solar radius, which determines the circular speed, could lie anywhere from less than 10% to almost 50% of the total force without violating the observational constraints (§6.3.3). The uncertain halo mass distribution inside R_0 implies an uncertain halo density at R_0, which is a significant concern to experimentalists hoping to detect the dark matter in laboratory experiments (Gaitskell 2004).

It is useful to parametrize the relative amounts of dark and luminous matter in a stellar system by the mass-to-light ratio. Stellar systems composed entirely of stars usually have mass-to-light ratios Υ_R in the range $1\text{–}10\Upsilon_\odot$, depending on the age and chemical composition of the stars, while systems composed entirely of dark matter would have $\Upsilon\to\infty$. The largest mass-to-light ratios known, $\Upsilon_R\sim500\Upsilon_\odot$, occur in dwarf spheroidal galaxies (page 24). In the R band, the luminosity of the Galaxy is $3\times10^{10}\,L_\odot$, so its mass-to-light ratio is $\sim60\Upsilon_\odot$, with large uncertainties ($7\text{–}170\Upsilon_\odot$) because of the uncertain mass of the dark halo.

A summary of properties of the Galaxy is provided in Table 1.2; for more details see §2.7.

for heavy. The use of the term "baryonic matter" for ordinary matter is conventional, but less than ideal for several reasons: (i) ordinary matter includes electrons, which are leptons, not baryons; (ii) the unknown dark-matter particle is likely to be even heavier than any baryonic particle; (iii) it is not clear whether to count neutrinos as "ordinary" matter, because they have been known for decades, or as dark matter, because they have mass and interact only weakly so they are WIMPs. Usually the latter choice is made.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Table 1.2 Properties of the Galaxy

Global properties:	
disk scale length R_d	$(2.5 \pm 0.5)\,\mathrm{kpc}$
disk luminosity	$(2.5 \pm 1) \times 10^{10}\,L_\odot$
bulge luminosity	$(5 \pm 2) \times 10^{9}\,L_\odot$
total luminosity	$(3.0 \pm 1) \times 10^{10}\,L_\odot$
disk mass	$(4.5 \pm 0.5) \times 10^{10}\,\mathcal{M}_\odot$
bulge mass	$(4.5 \pm 1.5) \times 10^{9}\,\mathcal{M}_\odot$
dark halo mass	$(2\,^{+3}_{-1.8}) \times 10^{12}\,\mathcal{M}_\odot$
dark halo half-mass radius	$(100\,^{+100}_{-80})\,\mathrm{kpc}$
disk mass-to-light ratio Υ_R	$(1.8 \pm 0.7)\Upsilon_\odot$
total mass-to-light ratio Υ_R	$(70\,^{+100}_{-63})\Upsilon_\odot$
black-hole mass	$(3.9 \pm 0.3) \times 10^{6}\,\mathcal{M}_\odot$
Hubble type	Sbc
Solar neighborhood properties:	
solar radius R_0	$(8.0 \pm 0.5)\,\mathrm{kpc}$
circular speed v_0	$(220 \pm 20)\,\mathrm{km\,s^{-1}}$
angular speed: from v_0/R_0	$(27.5 \pm 3)\,\mathrm{km\,s^{-1}\,kpc^{-1}}$
from Sgr A*	$(29.5 \pm 0.2)\,\mathrm{km\,s^{-1}\,kpc^{-1}}$
disk density ρ_0	$(0.09 \pm 0.01)\,\mathcal{M}_\odot\,\mathrm{pc^{-3}}$
disk surface density Σ_0	$(49 \pm 6)\,\mathcal{M}_\odot\,\mathrm{pc^{-2}}$
disk thickness Σ_0/ρ_0	$500\,\mathrm{pc}$
scale height z_d (old stars)	$300\,\mathrm{pc}$
rotation period $2\pi/\Omega_0$	$(220 \pm 30)\,\mathrm{Myr}$
vertical frequency $\nu_0 = \sqrt{4\pi G\rho_0}$	$(2.3 \pm 0.1) \times 10^{-15}\,\mathrm{Hz}$
	$= (70 \pm 4)\,\mathrm{km\,s^{-1}\,kpc^{-1}}$
vertical period $2\pi/\nu_0$	$87\,\mathrm{Myr}$
Oort's A constant	$(14.8 \pm 0.8)\,\mathrm{km\,s^{-1}\,kpc^{-1}}$
Oort's B constant	$-(12.4 \pm 0.6)\,\mathrm{km\,s^{-1}\,kpc^{-1}}$
epicycle frequency $\kappa_0 = \sqrt{-4B(A-B)}$	$(37 \pm 3)\,\mathrm{km\,s^{-1}\,kpc^{-1}}$
radial dispersion of old stars	$(38 \pm 2)\,\mathrm{km\,s^{-1}}$
vertical dispersion of old stars	$(19 \pm 2)\,\mathrm{km\,s^{-1}}$
RMS velocity of old stars	$(50 \pm 3)\,\mathrm{km\,s^{-1}}$
escape speed $v_\mathrm{e}(R_0)$	$(550 \pm 50)\,\mathrm{km\,s^{-1}}$

NOTES: See §2.7 and §§10.1 and 10.3 of BM for more detail. Luminosities are in the R band at $\lambda = 660\,\mathrm{nm}$. The halo mass and half-mass radius are taken from Wilkinson & Evans (1999). The angular speed of the central black hole (Sgr A*) relative to an extragalactic frame is from Reid & Brunthaler (2004). The density in the midplane of the disk, ρ_0, and the surface density Σ_0 are taken from Table 1.1. The scale height z_d is defined by equation (1.10). The RMS velocity of old stars is the square root of the sum of the squared dispersions along the three principal axes of the velocity-dispersion tensor (BM Table 10.2). Escape speed is from Smith et al. (2007).

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

1.1.3 Other galaxies

The nearest known galaxy to our own is the **Sagittarius dwarf galaxy**, (see Table 4.3 of BM or van den Bergh 2000 for a list of nearby galaxies). The Sagittarius galaxy has a total luminosity $L \simeq 2 \times 10^7\, L_\odot$ and is located on the opposite side of the Galaxy from us, about 24 kpc from the Sun and 16 kpc from the Galactic center. The line of sight to the Sagittarius dwarf passes only 15° from the Galactic center, so the galaxy is masked by the dense star fields of the Galactic bulge, and thus was discovered only in 1994 (from anomalies in the kinematics of what were thought to be bulge stars). The orbit of Sagittarius carries it so close to the center of the Galaxy that it is being disrupted by the Galactic tidal field, and a tidal tail or streamer—a trail of stars torn away from the main body of the galaxy—can be traced across most of the sky (Figure 8.10).

Our next nearest neighbor is the **Large Magellanic Cloud** or LMC. Although some 50 times as luminous than Sagittarius, with luminosity $L_R \simeq 1 \times 10^9\, L_\odot$, the LMC is still a relatively modest galaxy. The LMC is 45–50 kpc from the Sun and is visible to the naked eye in the southern hemisphere as a faint patch of light (see Plates 2 and 11). Because of its proximity and its location at relatively high Galactic latitude, where foreground contamination and dust obscuration are small ($b = -30°$), the LMC provides a unique laboratory for studies of interstellar gas and dust, stellar properties, and the cosmological distance scale. Also visible to the naked eye is the **Small Magellanic Cloud**, located 20° from the LMC on the sky, 20% further away, and with 20% of its luminosity. It is likely that the two Clouds are a former binary system that has been disrupted by tidal forces from the Galaxy.

The nearest large disk galaxy similar to our own is called the **Andromeda galaxy**, **M31**, or **NGC 224** (see Plate 3 and Hodge 1992). M31 is more than ten times as far away as the LMC ($d \simeq 740$ kpc) and more than ten times as luminous ($L \simeq 4 \times 10^{10}\, L_\odot$). Only the central parts of M31 are visible to the naked eye, but deep telescopic images show that its stellar disk extends across more than six degrees on the sky.

Our Galaxy is just one member of a vast sea of some 10^9 galaxies stretching to a distance of several thousand megaparsecs (1 megaparsec $\equiv$ 1 Mpc $= 10^6$ pc $= 3.086 \times 10^{22}$ m). The determination of distances to these galaxies is one of the most important tasks in extragalactic astronomy, since many of the properties derived for a galaxy depend on the assumed distance. Methods for measuring galaxy distances are described in detail in Chapter 7 of BM. For our purposes it is sufficient to note that in a universe that is homogeneous and isotropic, the relative velocity v between two galaxies that are separated by a large distance r is given by the **Hubble law**,

$$v = H_0 r, \tag{1.13}$$

where H_0 is the **Hubble constant** (see §1.3.1). Our universe is nearly homogeneous and isotropic on large scales, so the velocity field or **Hubble flow**

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

implied by the Hubble law is approximately correct: the only significant error comes from random velocities of a few hundred $\mathrm{km\,s^{-1}}$ that are generated by the gravitational acceleration from small-scale irregularities in the cosmic matter density. Thus, for example, the distance of a galaxy with a velocity of $7000\,\mathrm{km\,s^{-1}}$ is known to within about 5% once the Hubble constant is known.

By comparing the flux from Cepheid variable stars in the Large Magellanic Cloud and more distant galaxies, Freedman et al. (2001) deduce that

$$H_0 = (72 \pm 8)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}; \tag{1.14}$$

while measurements of small fluctuations in the cosmic background radiation (§1.3.5 and Spergel et al. 2007) give

$$H_0 = (73.5 \pm 3.2)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}. \tag{1.15}$$

When precision is required, we shall write the Hubble constant as

$$\begin{aligned} H_0 &\equiv 70\,h_7\,\mathrm{km\,s^{-1}\,Mpc^{-1}} \\ &= 2.268\,h_7 \times 10^{-18}\,\mathrm{Hz}, \\ H_0^{-1} &= 13.97\,h_7^{-1}\,\mathrm{Gyr}, \end{aligned} \tag{1.16}$$

where the dimensionless parameter h_7 is probably within 10% of unity. Any uncertainty in the Hubble constant affects the whole distance scale of the universe and hence is reflected in many of the average properties of galaxies; for example, the mean density of galaxies scales as h_7^3 and the mean luminosity of a galaxy of a given type scales as h_7^{-2}.

If galaxies suffered no acceleration due to external gravitational forces, the distance between any two galaxies would be a linear function of time. Combined with the Hubble law (1.13), this assumption implies that the distance between any two galaxies was zero at a time H_0^{-1} (the **Hubble time**) before the present. The Hubble time provides a rough estimate of the age of the universe. The actual age is somewhat different because the relative velocities of galaxies are decelerated by the gravitational attraction of baryonic and dark matter and accelerated by vacuum energy (see §1.3.3); these effects nearly cancel at present, so the best estimate of the age, $t_0 = 13.7\,\mathrm{Gyr}$ (eq. 1.77), is accidentally very close to the Hubble time.

Galaxies can usefully be divided into four main types according to the **Hubble classification system**—see BM §4.1.1 or Sandage & Bedke (1994) for a more complete description.

(a) Elliptical galaxies These are smooth, featureless stellar systems containing little or no cool interstellar gas or dust and little or no stellar disk. The galaxy M87 shown in Plate 4 is a classic example of this type. The stars

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

in most elliptical galaxies are old, having ages comparable to the age of the universe, consistent with the absence of gas from which new stars can form.

The fraction of luminous galaxies that are elliptical depends on the local density of galaxies, ranging from about 10% in low-density regions to over 40% in the centers of dense clusters of galaxies (BM §4.1.2).

As the name suggests, the contours of constant surface brightness, or **isophotes**, of elliptical galaxies are approximately concentric ellipses, with axis ratio b/a ranging from 1 to about 0.3. The **ellipticity** is $\epsilon \equiv 1 - b/a$. In the Hubble classification system, elliptical galaxies are denoted by the symbols E0, E1, etc., where a galaxy of type En has axis ratio $b/a = 1 - n/10$. The most elongated elliptical galaxies are type E7. Since we see only the projected brightness distribution, it is impossible to determine directly whether elliptical galaxies are axisymmetric or triaxial; however, indirect evidence strongly suggests that both shapes are present (BM §§4.2 and 4.3).

The surface brightness of an elliptical galaxy falls off smoothly with radius, until the outermost parts are undetectable against the background sky brightness. Because galaxies do not have sharp outer edges, their sizes must be defined with care. One useful measure of size is the **effective radius** $R_{\rm e}$, the radius of the isophote containing half of the total luminosity[7] (or the geometric mean of the major and minor axes of this isophote, if the galaxy is elliptical). The effective radius is correlated with the luminosity of the elliptical galaxy, ranging from 20 kpc for a giant galaxy such as M87 (Plate 4) to 0.2 kpc for a dwarf such as M32 (Plate 3).

The Hubble classification is based on the ellipticity of the isophotes near the effective radius. In many galaxies the isophotes become more elliptical at large radii; thus, for example, M87 is classified as E0 but the isophotal axis ratio is only 0.5 in its outermost parts.

Several empirical formulae have been used to fit the surface-brightness profiles of ellipticals. One of the most successful is the **Sérsic** law

$$I_m(R) = I(0)\exp(-kR^{1/m}) = I_{\rm e}\exp\left\{-b_m[(R/R_{\rm e})^{1/m} - 1]\right\}; \qquad (1.17)$$

here $I(R)$ is the surface brightness at radius R and $I_{\rm e}$ is the surface brightness at the effective radius $R_{\rm e}$. The parameter m is the **Sérsic index**, which is correlated with the luminosity of the elliptical galaxy, luminous ellipticals having $m \simeq 6$ and dim ones having $m \simeq 2$. The middle of this range is $m = 4$, which defines the **de Vaucouleurs** or $R^{1/4}$ law (de Vaucouleurs 1948). The function b_m must be determined numerically from the condition $\int_0^{R_{\rm e}} {\rm d}R\, RI_m(R) = \frac{1}{2}\int_0^\infty {\rm d}R\, RI_m(R)$; but the fitting formula $b_m = 2m - 0.324$ has fractional error $\lesssim 0.001$ over the range $1 < m < 10$ (see Ciotti & Bertin 1999 for properties of Sérsic laws). For $m = 1$ the Sérsic law reduces to the

[7] The effective radius is measured on the plane of the sky, and is not to be confused with the half-light or median radius (page 17), the radius of a sphere containing half the luminosity.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

exponential profile (1.7) that describes the surface-brightness distribution of disk galaxies.

The total luminosity of a galaxy is difficult to define precisely because the outer parts are too faint to measure. One approach is to define a **model luminosity** by fitting the surface-brightness profile to a Sérsic or de Vaucouleurs profile and then estimating the luminosity as $L = \int \mathrm{d}^2\mathbf{R}\, I_m(R)$.

The luminosities of elliptical galaxies range over a factor of 10^8, from almost $10^{12}\, L_\odot$ for the very luminous galaxies found at the centers of massive clusters of galaxies, to $\lesssim 10^4\, L_\odot$ for the dimmest dwarf galaxies. The **luminosity function** $\phi(L)$ describes the relative numbers of galaxies of different luminosities, and is defined so that $\phi(L)\,\mathrm{d}L$ is the number of galaxies in the luminosity interval $L \to L + \mathrm{d}L$ in a representative unit volume of the universe. A convenient analytic approximation to $\phi(L)$ is the **Schechter law** (BM §4.1.3),

$$\phi(L)\,\mathrm{d}L = \phi_\star \left(\frac{L}{L_\star}\right)^\alpha \exp(-L/L_\star)\frac{\mathrm{d}L}{L_\star}, \tag{1.18}$$

where $\phi_\star \simeq 4.9\times10^{-3}\, h_7^3\,\mathrm{Mpc}^{-3}$, $\alpha = -1.1$, and $L_\star \simeq 2.9\times10^{10}\, h_7^{-2}\, L_\odot$ in the R band (Brown et al. 2001). The concept of a "universal" luminosity function embodied in the Schechter law is no more than a good first approximation: in fact the luminosity function is known to depend on both galaxy type and environment (BM §4.1.3).

The average R-band luminosity density derived from equation (1.18) is

$$j_R = \int \mathrm{d}L\, L\phi(L) = \phi_\star L_\star \int_0^\infty \mathrm{d}x\, x^{\alpha+1}\mathrm{e}^{-x} = (\alpha+1)!\,\phi_\star L_\star, \tag{1.19}$$

where the factorial function is defined for non-integer arguments in Appendix C.2. For the parameters given above, $j_R = 1.5\times10^8 h_7\, L_\odot\,\mathrm{Mpc}^{-3}$, with an uncertainty of about 30%.

Most luminous elliptical galaxies exhibit little or no rotation, even those with large ellipticity; this is in contrast to stars or other gravitating gas masses, which must be spherical if they do not rotate and flattened when rotating. Among dimmer elliptical galaxies, however, rotation and flattening do appear to be correlated (see §4.4.2c and Faber et al. 1997). This distinction between luminous and dim ellipticals may arise because the most recent mergers of luminous galaxies have been "dry," that is, between progenitors containing little or no gas, while the recent mergers of low-luminosity ellipticals have involved gas-rich systems. Whether or not this interpretation is correct, the different rotational properties of high-luminosity and low-luminosity elliptical galaxies illustrate that stellar systems can exhibit a much greater variety of equilibria than gaseous systems such as stars.

Each star in an elliptical galaxy orbits in the gravitational field of all the other stars and dark matter in the galaxy. The velocities of individual stars

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

can be measured in only a few nearby galaxies, but in more distant galaxies the overall distribution of stellar velocities along the line of sight can be determined from the Doppler broadening of lines in the integrated spectrum of the galaxy. The most important parameter describing this distribution is the RMS line-of-sight velocity $\sigma_\parallel$, sometimes called simply the velocity dispersion (eq. 4.25).

The luminosity, velocity dispersion, and size of elliptical galaxies are correlated. Astronomers usually plot this correlation using not the luminosity but the average surface brightness within the effective radius, which is simply $\overline{I}_{\rm e} \equiv \frac{1}{2}L/(\pi R_{\rm e}^2)$. Then if we plot the positions of a sample of elliptical galaxies in the three-dimensional space with coordinates $\log_{10}\overline{I}_{\rm e}$, $\log_{10} R_{\rm e}$, and $\log_{10}\sigma_\parallel$, they are found to lie on a two-dimensional surface, the **fundamental plane** (see BM §4.3.4 and §4.9.2), given by

$$\log_{10} R_{\rm e} = 1.24\log_{10}\sigma_\parallel - 0.82\log_{10}\overline{I}_{\rm e} + \text{constant}, \tag{1.20}$$

with an RMS scatter of 0.08 in $\log_{10} R_{\rm e}$ or 0.07 in $\log_{10}\sigma_\parallel$ (Jørgensen et al. 1996).

The properties of galaxies are determined both by the fundamental plane and by their distribution within that plane. Let us think of the space with coordinates $(\log_{10}\overline{I}_{\rm e}, \log_{10} R_{\rm e}, \log_{10}\sigma_\parallel)$ as a fictitious three-dimensional space, and imagine observing the distribution of galaxies from a distance. If the line of sight to the observer in this fictitious space lies close to the fundamental plane, the observer will find that galaxies lie close to a line in the two-dimensional space normal to the line of sight. This distribution of galaxies can be thought of as a projection of the distribution in the fundamental plane. The most important of these projections are:

(i) The **Faber–Jackson law** (BM §4.3.4),

$$\log_{10}\left(\frac{\sigma_\parallel}{150\,\mathrm{km\,s^{-1}}}\right) \simeq 0.25\log_{10}\left(\frac{L_R}{10^{10}h_7^{-2}\,L_\odot}\right). \tag{1.21}$$

Thus the velocity dispersion of an $L_\star$ galaxy is $\sigma_\parallel \simeq 200\,\mathrm{km\,s^{-1}}$. The RMS scatter in the Faber–Jackson law is about 0.1 in $\log_{10}\sigma_\parallel$ (Davies et al. 1983).

(ii) The **Kormendy relation**

$$\log_{10}\left(\frac{\overline{I}_{\rm e,R}}{1.2\times10^3\,L_\odot\,\mathrm{pc^{-2}}}\right) = -0.8\log_{10}\left(\frac{R_{\rm e}}{h_7^{-1}\,\mathrm{kpc}}\right). \tag{1.22a}$$

Here $\overline{I}_{\rm e,R}$ denotes the mean R-band surface brightness interior to $R_{\rm e}$. The RMS scatter is less than 0.25 in $\log_{10}\overline{I}_{\rm e}$. The Kormendy relation implies that

$$\log_{10}\left(\frac{L_R}{7.7\times10^9h_7^{-2}\,L_\odot}\right) = 1.2\log_{10}\left(\frac{R_{\rm e}}{h_7^{-1}\,\mathrm{kpc}}\right). \tag{1.22b}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Thus more luminous galaxies are larger, but have lower surface brightness.

Careful dynamical modeling (§4.9.2) allows us to determine the mass-to-light ratio Υ in elliptical galaxies. These studies show that at radii less than $\sim R_{\rm e}$ the mass-to-light ratio is not strongly dependent on radius, and consistent with the mass-to-light ratio that we would expect from the observed stellar population (Cappellari et al. 2006). Thus the contribution of dark matter to the mass inside $R_{\rm e}$ is $\lesssim 30\%$. The mass-to-light ratio is also tightly correlated with $\sigma_{\rm e}$, the luminosity-weighted velocity dispersion within $R_{\rm e}$:

$$\Upsilon_I = (3.80 \pm 0.2)\,\Upsilon_\odot \times \left(\frac{\sigma_{\rm e}}{200\,{\rm km\,s^{-1}}}\right)^{0.84\pm0.07} \tag{1.23}$$

with an intrinsic scatter of only 13%.

Just as stars are found in gravitationally bound systems such as galaxies, many galaxies are found in bound systems called **groups** or **clusters of galaxies** (see §1.1.5). The largest clusters of galaxies are several Mpc in radius and contain thousands of galaxies. The most luminous galaxy in a large cluster—more often called a **rich cluster**—is often exceptional, in that it is (i) several times more luminous than any other cluster galaxy, and much more luminous than one would expect from the Schechter law (1.18) ($L/L_\star \sim$ 3–10); (ii) at rest in the center of the cluster; (iii) surrounded by a dim stellar halo that extends out to $\sim 1\,$Mpc. Galaxies with these unique characteristics are called **brightest cluster galaxies**; the nearest example is M87 in the Virgo cluster (Plate 4).[8] The existence of an extended dim halo is also the defining property of **cD galaxies** (BM §4.3.1); in practice, the terms "brightest cluster galaxy" and "cD galaxy" are often used interchangeably. The halo probably arises from stars that have been stripped from individual cluster galaxies by tidal forces and now orbit independently in the cluster's gravitational field. Brightest cluster galaxies are believed to form during the hierarchical assembly of the cluster from smaller subunits (Dubinski 1998).

The dimmest elliptical galaxies are also unusual. In general, dim ellipticals have higher surface brightness than luminous ellipticals, a manifestation of the Kormendy relation (1.22). However, at luminosities $\lesssim 10^9\,L_\odot$ a distinct family of **diffuse dwarf elliptical** or **dwarf spheroidal** galaxies appears, with much larger effective radii and lower surface brightnesses than "normal" ellipticals of the same luminosity (Mateo 1998). Dwarf spheroidal galaxies are difficult to detect because their surface brightness is much less than that of the night sky; nearby dwarf spheroidals are discovered because their brightest stars produce a slight enhancement in star counts that are otherwise dominated by foreground stars belonging to our own Galaxy.

[8] The most luminous galaxy in the Virgo cluster is actually the E2 galaxy M49=NGC 4472, rather than M87. The Virgo cluster has a complex structure, consisting of two main concentrations, a dominant one near M87 and a smaller one near M49. These probably represent two merging sub-clusters, each with its own brightest cluster galaxy.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

There are at least 20 dwarf spheroidal galaxies within 200 kpc, and given the limited sky coverage of existing surveys the actual number may be 50–100 (Belokurov et al. 2007). All appear to be satellites orbiting the Galaxy.Galaxy, satellites of Their luminosities range from $2 \times 10^7\, L_\odot$ to $\lesssim 10^4\, L_\odot$. The dwarf spheroidals offer a unique probe of dark matter in galaxies, for the following reason. In more luminous galaxies, both baryonic matter (stars and gas) and dark matter contribute comparable amounts to the total mass within the visible stellar system; thus, disentangling their effects to isolate the properties of the dark-matter distribution at small radii is difficult. In some dwarf spheroidal galaxies, however, dark matter contributes 90% or more of the total mass, even at the center of the galaxy, so the dynamics is determined entirely by the gravitational field of the dark matter.

The distribution of mass in the dark halos of ellipticals can be constrained by several methods, including: (i) The kinematics of tracer particles such as globular clusters or planetary nebulae, which typically sample radii from 10–30 kpc (Côté et al. 2003; Romanowsky et al. 2003). This approach relies on the statistical analysis of the positions and velocities of hundreds or thousands of objects, assuming they are found at random orbital phases. (ii) Diffuse X-ray emission from hot gas around the galaxy (Mathews & Brighenti 2003). Measurements of the emissivity and temperature distribution, combined with the plausible assumption that the gas is in hydrostatic equilibrium, can be used to constrain the distribution of the dark matter out to ~ 30 kpc in isolated galaxies. The same technique can be applied to brightest cluster galaxies out to much larger radii, but in this case we are measuring the combined dark-matter distribution of the galaxy and the cluster. (iii) Kinematics of satellite galaxies. This technique is similar in principle to the use of globular clusters or planetary nebulae; the satellite galaxies have the advantage that they sample much larger radii, from 100–400 kpc, but the disadvantage that generally no more than one satellite is detected around a given galaxy, so the method yields only an average of the dark-matter distribution over many galaxies (Prada et al. 2003). (iv) Weak gravitational lensing, in which the gravitational field of a nearby galaxy distorts the images of distant background galaxies (Schneider 2006); once again, this method requires averaging over a large sample of lensing galaxies.

The preliminary conclusion from these studies is that luminous, isolated elliptical galaxies contain dark halos that are much larger—both in size and in mass—than the stellar systems they surround. Within uncertainties of at least a factor of two, the halos extend to ~ 300 kpc and contain ~ 10 times the mass in stars.

(b) Spiral galaxies These are galaxies, like the Milky Way and M31, that contain a prominent disk composed of stars, gas, and dust. The disk contains **spiral arms**, filaments in which stars are continuously being formed. The same spiral arms are seen in the old stars that dominate the mass of the

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

disk (see Figure 6.1 and §6.1.2). The spiral arms vary greatly in their shape, length and prominence from one galaxy to another but are always present.

In low-density regions of the universe, about 60% of all luminous galaxies are spirals, but the fraction drops to $\lesssim 10\%$ in dense regions such as the cores of galaxy clusters (BM §4.1.2).

The surface brightness in spiral galaxy disks, which traces the radial distribution of stars, obeys the exponential law (1.7) (de Jong 1996). A typical disk scale length is $R_{\rm d} \simeq 2h_7^{-1}\,{\rm kpc}$, but scale lengths range from $1h_7^{-1}\,{\rm kpc}$ to more than $10h_7^{-1}\,{\rm kpc}$. The typical central surface brightness is $I_{\rm d} \sim 100\,L_\odot\,{\rm pc}^{-2}$ (BM Figure 4.52). The interstellar gas in spiral galaxy disks often extends to much larger radii than the stars, probably because star formation is suppressed when the gas surface density falls below a critical value (see Plates 5, 6 and BM §8.2.8).

Using the 21-cm line of interstellar neutral hydrogen, the circular-speed curves $v_{\rm c}(R)$ of spiral galaxies can be followed out to radii well beyond the outer edge of the stellar distribution. The circular-speed curves of luminous spirals are nearly flat out to the largest radii at which they can be measured, often a factor of two or more larger than the edge of the stellar disk (BM §8.2.4). If most of the mass of the galaxy were in stars, we would expect the circular-speed curve at these large radii to fall as $v_{\rm c}(R) = (GM/R)^{1/2}$ where M is the total stellar mass (see Figure 2.17). The inescapable conclusion is that the mass of the galaxy at these large radii is dominated by the dark halo rather than the stars.

Typical circular speeds of spirals are between 100 and $300\,{\rm km\,s^{-1}}$. Just as the velocity dispersion of elliptical galaxies is related to their luminosity by the Faber–Jackson law (1.21), the rotation rate of spirals in the flat part of the circular-speed curve is related to their luminosity by the **Tully–Fisher law** (BM §7.3.4; Sakai et al. 2000),

$$\log_{10}\left(\frac{L_R h_7^2}{10^{10}\,L_\odot}\right) = 3.5\log_{10}\left(\frac{v_{\rm c}}{200\,{\rm km\,s^{-1}}}\right) + 0.5; \qquad (1.24)$$

the RMS scatter in this relation is about 0.14 in $\log_{10} L_R$. The slope of the Tully–Fisher law is a function of the wavelength band in which the luminosity is measured, ranging from $\simeq 3$ in the B band centered at $0.45\,\mu$m to $\simeq 4$ in the K band centered at $2.2\,\mu$m. Applied to our Galaxy, using $v_{\rm c} = (220 \pm 20)\,{\rm km\,s^{-1}}$ from equation (1.8), we find $L_R = (4.4 \pm 1.5) \times 10^{10}\,L_\odot$, consistent with Table 1.2.

Like the Milky Way, most spiral galaxies contain a bulge, a centrally concentrated stellar system that has a smooth or amorphous appearance—quite unlike that of the disk, which exhibits spiral arms, dust lanes, concentrations of young stars, and other structure. The origin of bulges is not well understood: some resemble small elliptical galaxies and presumably formed in the same way, while others resemble thickened disks, and may have formed from

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

the disk through dynamical processes (see §6.6.2 and Kormendy & Kennicutt 2004). Bulges and elliptical galaxies are sometimes called **spheroidal stellar systems** or **spheroids**, even though their shapes are not necessarily close to mathematical spheroids (page 76)—in particular, many of them are probably not axisymmetric.

The luminosity of the bulge relative to that of the disk is correlated with many other properties of the galaxy, such as the fraction of the disk mass in gas, the color of the disk, and how tightly the spiral arms are wound. This correlation is the basis of a sub-division in the Hubble classification system, which breaks up spiral galaxies into four classes or types, called Sa, Sb, Sc, Sd (Sandage & Bedke 1994). Along the sequence Sa→Sd, (i) the relative luminosity of the bulge decreases; (ii) the spiral arms become more loosely wound; (iii) the relative mass of gas increases; and (iv) the spiral arms become more clumpy, so individual patches of young stars and HII regions become more prominent. This sequence is illustrated by comparing the images of M104 (Plate 7), which is classified Sa; M81 (Plate 8), which is classified Sab (i.e., between Sa and Sb); the Sb galaxy M31 (Plate 3); the Sbc galaxies M51, M63, and M100 (Plates 1, 9, and 17); the Sc galaxy M101 (Plate 18), and the Scd galaxy M33 (Plate 19). The Milky Way is type Sbc.

The Hubble classification also divides spiral galaxies into "normal" and "barred" categories. The bar is an elongated, smooth stellar system that is reminiscent of a rigid paddle or stirrer rotating at the center of the galactic disk. The bar can be thought of as a triaxial bulge, and in practice there is no clear distinction between these two categories of stellar system: for example, a "bar" in a face-on galaxy might well be classified as a "bulge" if the galaxy were viewed edge-on. Further properties of bars are described in §6.5.

A classic barred galaxy is NGC 1300 (Plate 10) although most bars are less prominent than the one in this galaxy. Other barred galaxies are shown in Figures 6.27 and 6.28. Our own Galaxy and its neighbor, the Large Magellanic Cloud (Plate 11), are both barred. About half of all spirals are barred, and bars appear in all of the Hubble classes Sa, Sb, Sc, Sd, where their presence is indicated by inserting the letter "B" into the notation (SBa, SBb, etc.). Elliptical galaxies do not have bars.

The first evidence for dark halos in spiral galaxies came from circular-speed curves; as we have discussed, neutral-hydrogen rotation curves in some spirals remain flat out to as much as 10 times the scale length of the stellar disk, which implies that the mass at large radii is dominated by dark matter rather than stars. At much larger radii, $\gtrsim 100\,\mathrm{kpc}$, the distribution of dark matter can be measured by the same techniques that are used for ellipticals, in particular satellite galaxy kinematics and weak gravitational lensing. Within the large uncertainties, the data are consistent with the hypothesis that the size and mass of the dark halos that surround luminous spiral galaxies are the same as those surrounding isolated ellipticals—about 300 kpc in radius and containing ~ 10 times the stellar mass of the galaxy.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For most spiral galaxies, the relative contributions of dark and luminous matter within the visible stellar system are difficult to disentangle (§6.3.3). However, in some low-luminosity and low surface-brightness spirals, dark matter appears to dominate the mass at all radii (Swaters et al. 2003). Like the dwarf spheroidals, these galaxies provide valuable probes of the properties of dark halos on small scales.

(c) Lenticular galaxies These are transition objects between elliptical and spiral galaxies: like spirals, they contain a rapidly rotating disk, a bulge, and sometimes a bar, and the disk obeys the exponential surface-brightness law (1.7) characteristic of spirals. Like ellipticals, they have little or no cool gas or recent star formation, are smooth and featureless in appearance, and exhibit no spiral structure. The absence of young stars is a consequence of the absence of gas, since this is the raw material from which stars are formed.

Lenticulars are rare in low-density regions, but comprise almost half of the galaxies in the high-density centers of galaxy clusters (BM Figure 4.10). This correlation suggests that lenticulars may be spirals that have been depleted of interstellar gas by interactions with the hot gas in the cluster (van Gorkom 2004).

Lenticulars are labeled in the Hubble classification by the notation S0, or SB0 if barred. The transition from ellipticals to lenticulars to spirals is smooth and continuous, so there are S0 galaxies that might well be classified as E7 and others that could be Sa (Sandage & Bedke 1994).

(d) Irregular galaxies Along the sequence from Sc to Sd, galaxies become progressively less luminous and their spiral structure becomes less well defined. These trends continue beyond Sd: we find low-luminosity ("dwarf") disk galaxies in which the young stars are arranged chaotically rather than in spirals. These are called "irregular" galaxies and are denoted in the Hubble classification by Sm or Im, the prototypes of these two classes being the Large and Small Magellanic Clouds.

Irregular galaxies are extremely common—more than a third of our neighbors are of this type—but they do not feature prominently in most galaxy catalogs, because any flux-limited catalog is biased against intrinsically dim systems.

In irregulars the circular speed is a linear function of radius (corresponding to a constant angular speed) over most of the stellar disk, reaching a maximum of ~ 50–$70\,\mathrm{km\,s^{-1}}$ near the edge of the disk. These properties are in sharp contrast to luminous spiral galaxies, in which the circular speed is much higher and the circular-speed curve is nearly flat.

Much of the luminosity of irregular galaxies is emitted by massive young stars and large HII regions. These systems are extremely gas-rich: the interstellar gas in their disks often contains more than 30% of the mass in stars. Their irregular appearance arises partly because the optical emission is dominated by a relatively small number of luminous young stars and HII regions, and partly because the circular speed in the disk is not that much larger than the turbulent velocities in the interstellar gas ($\sim 10\,\mathrm{km\,s^{-1}}$).

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

A minority of galaxies are assigned to the "irregular" bin simply because they fit nowhere else: these include spiral or elliptical galaxies that have been violently distorted by a recent encounter with a neighbor (see Plate 12), galaxies in the last stages of merging, and galaxies that are undergoing an intense burst of star formation that overwhelms the stellar population that usually determines the classification.

It is convenient to think of the Hubble classification as a sequence E→S0→Sa→Sb→Sc→Sd→Sm→Im. Galaxies near the beginning of this sequence are called **early**, while those near the end are **late**. Thus the term "early-type galaxies" refers to ellipticals and lenticulars; an Sa galaxy is an "early-type spiral," while an Sc or Sd galaxy is a "late-type spiral," etc. This terminology is a fossil of the initial incorrect belief that the Hubble sequence was an evolutionary or time sequence.

1.1.4 Open and globular clusters

A typical galaxy contains many small stellar systems of between 10^2 and 10^6 stars. These systems are called **star clusters** and can be divided into two main types.

Open clusters are irregular stellar systems that contain $\sim 10^2$ to 10^4 stars (see Table 1.3, Plate 13, and BM §6.2). New open clusters are formed continuously in the Galactic disk, and most of the ones we see are younger than 1 Gyr (Figure 8.5). Older clusters are rare because most have been disrupted, probably by gravitational shocks from passing interstellar gas clouds (§8.2.2c). There are over 1000 cataloged open clusters out of an estimated 10^5 throughout the Galaxy. It is likely that most of the stars in the Galactic disk formed in open clusters that have since dissolved.

Globular clusters are much more massive stellar systems, containing 10^4–10^6 stars in a nearly spherical distribution (BM §§4.5 and 6.1, and Plate 14; see Ashman & Zepf 1998 and Carney & Harris 2001 for reviews). Globular clusters do not contain gas, dust, or young stars. Our Galaxy contains about 150 globular clusters, but large elliptical galaxies such as M87 can contain as many as 10 000 (Plate 4). Unlike open clusters, the Galaxy's globular clusters are old, and are believed to be relics of the formation of the Galaxy itself.[9] The metallicity appears to be the same for all the stars in a given cluster—presumably because the cluster formed from a well-mixed gas cloud—but different clusters have a wide range of metallicity, from only $0.005Z_\odot$ to nearly solar. The spatial distribution and the kinematics of a group of clusters are correlated with the metallicity, and for many purposes the clusters in our Galaxy can be divided into two groups (Zinn 1985): a roughly spherical population that contains 80% of the clusters, shows little

[9] It is a mystery why young globular clusters are absent in our Galaxy but common in many others, such as M31, the Large Magellanic Cloud, and galaxies that have undergone recent mergers.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

or no rotation and has metallicity $Z < 0.1Z_\odot$, and is associated with the stellar halo; and a flattened population that contains the remaining 20%, has $Z > 0.1Z_\odot$, exhibits rapid rotation, and is associated with the disk and bulge. This bimodal distribution of metallicities is present in the globular-cluster systems of other galaxies as well (Gebhardt & Kissler–Patig 1999).

The stellar density in the center of a globular cluster is extremely high: a typical value is $10^4\,\mathcal{M}_\odot\,\mathrm{pc}^{-3}$, compared with $0.05\,\mathcal{M}_\odot\,\mathrm{pc}^{-3}$ in the solar neighborhood. Because globular clusters have strong or high **central concentration** (the central density is much larger than the mean density) three different measures of the radius are usually quoted for globular clusters: the **core radius**, where the surface brightness has fallen to half its central value; the median or half-light radius, the radius of a sphere that contains half of all the luminosity; and the **limiting** or **tidal radius**, the outer limit of the cluster where the density drops to zero. Typical values of these and other cluster parameters are given in Table 1.3.

Luminous globular clusters emit as much light as dwarf spheroidal galaxies. However, a dwarf spheroidal galaxy is a very low surface-brightness object with a half-light radius of $\sim 300\,\mathrm{pc}$, while a luminous globular cluster has a much smaller radius ($\sim 3\,\mathrm{pc}$) and a correspondingly higher surface brightness. A handful of exceptionally luminous globular clusters, such as ω Centauri in our Galaxy and G1 in M31, may be the dense centers of tidally disrupted galaxies (Freeman 1993).

Globular clusters are among the simplest stellar systems: they are spherical, they have no dust or young stars to obscure or confuse the observations, they appear to have no dark matter other than low-luminosity stars, and they are **dynamically old**: a typical star in a globular cluster has completed many orbits ($\sim 10^4$) since the cluster was formed. Thus globular clusters provide the best physical realization we have of the **gravitational N-body problem**, which is to understand the evolution of a system of N point masses interacting only by gravitational forces (Chapter 7).

1.1.5 Groups and clusters of galaxies

Galaxies are not distributed uniformly in the universe. They belong to a rich hierarchy of structure that includes binary galaxies, small groups of a few galaxies in close proximity, enormous voids in which the number density of galaxies is greatly depleted, filaments and walls stretching for tens of Mpc, and rare giant clusters containing thousands of galaxies (see §9.2.2 and Mulchaey, Dressler, & Oemler 2004). Only on scales $\gtrsim 100\,\mathrm{Mpc}$ is the distribution of galaxies statistically homogeneous.

Associations that contain only a handful of galaxies are called groups while bigger associations are called clusters of galaxies (see Plates 15 and 16). The dividing line between groups and clusters is arbitrary, since the distribution of properties is continuous from one class to the other.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Table 1.3 Parameters of globular and open clusters

	globular	open
central density ρ_0	$1 \times 10^4\,\mathcal{M}_\odot\,\mathrm{pc}^{-3}$	$10\,\mathcal{M}_\odot\,\mathrm{pc}^{-3}$
core radius r_c	1 pc	1 pc
half-mass radius r_h	3 pc	2 pc
tidal radius r_t	35 pc	10 pc
central velocity dispersion σ_0	$6\,\mathrm{km\,s}^{-1}$	$0.3\,\mathrm{km\,s}^{-1}$
crossing time r_h/σ_0 (line-of-sight)	0.5 Myr	7 Myr
mass-to-light ratio Υ_R	$2\Upsilon_\odot$	$1\Upsilon_\odot$
mass M	$2 \times 10^5\,\mathcal{M}_\odot$	$300\,\mathcal{M}_\odot$
lifetime	10 Gyr	300 Myr
number in the Galaxy	150	10^5

NOTES: Values for globular clusters are medians from the compilation of Harris (1996). Values for open clusters are from Figure 8.5, Piskunov et al. (2007), and other sources.

The galaxies within ~ 1 Mpc are members of the **Local Group**. The two dominant members of this group are the Galaxy and M31. Dozens of smaller galaxies, mostly satellites of the two dominant galaxies, are also members (see BM §4.1.4 and van den Bergh 2000). The Local Group is believed to be a physical system rather than a chance superposition because the density of galaxies in this region is substantially higher than average, and because the Galaxy and M31 are approaching one another rather than receding with the Hubble flow. It is believed that the gravitational attraction between these two galaxies slowed and then reversed their recession, and that they will eventually merge into a single giant stellar system (see Box 3.1 and Figure 8.1).

Like star clusters, groups and clusters of galaxies may be regarded for many purposes as assemblies of masses orbiting under their mutual gravitational attraction, except that now the masses are galaxies rather than stars. However, there are two important differences between the dynamics of star clusters and galaxy groups or clusters. First, groups and clusters of galaxies are **dynamically young**: a typical galaxy in even the largest and most populous clusters has completed only a few orbits since the cluster formed, and in many smaller groups, including the Local Group, galaxies are still falling towards the group center for the first time. Second, the fractional volume of a group or cluster that is occupied by galaxies ($\gtrsim 10^{-3}$) is much larger than the fractional volume of a star cluster that is occupied by stars ($\approx 10^{-19}$). Thus collisions between galaxies in a cluster are much more frequent than collisions between stars in a star cluster (see Chapter 8).

Clusters of galaxies are the largest equilibrium structures in the universe. They arose from the gravitational collapse of rare high peaks in the fluctuating density field of dark matter in the early universe (§9.2). Conse-

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

quently their properties provide a sensitive probe of cosmological parameters. Clusters also offer a unique probe of the distribution of dark matter on large scales. The mass distribution in clusters of galaxies can be measured by many complementary methods, including (i) statistical analysis of the velocities and positions of large numbers of galaxies in the cluster; (ii) measurements of the X-ray emissivity and temperature of hot gas in the cluster; (iii) distortion of the images of background galaxies by weak gravitational lensing; (iv) strong gravitational lensing, which can produce multiple images of background galaxies near the center of the cluster; (v) the **Sunyaev–Zeldovich effect**, which is a slight depression in the measured temperature of the cosmic microwave background at the locations of clusters, caused by Compton scattering of photons by electrons in the hot cluster gas.

The biggest clusters have masses $\sim 10^{15}\,\mathcal{M}_\odot$ within 2 Mpc of their centers and velocity dispersions of $\sim 1000\,\mathrm{km\,s^{-1}}$. The mass-to-light ratios are

$$\Upsilon_R \simeq (200 \pm 50) h_7 \Upsilon_\odot, \tag{1.25}$$

(Fukugita, Hogan, & Peebles 1998), with no detectable dependence on cluster properties such as velocity dispersion or total population.

Most of the baryons in clusters of galaxies are in the hot gas. The mass in gas is a fraction $0.11 h_7^{-3/2}$ of the total mass (Allen, Schmidt & Fabian 2002), while the mass in stars is only about $0.02 h_7^{-1}$ of the total. Thus the fraction of the total mass that resides in baryons is

$$f_\mathrm{b} = 0.13 \pm 0.02, \tag{1.26}$$

with the remaining 87% comprised of WIMPs or other non-baryonic dark matter. In structures as large as clusters it is difficult to imagine how baryons and non-baryonic dark matter could be segregated (in contrast to individual galaxies, where the baryons have concentrated at the center of the dark halo to form the visible stars). Thus the baryon-to-total mass ratio f_b that is found in clusters should be a fair sample of the universe as a whole.

1.1.6 Black holes

Dynamical studies of the centers of galaxies reveal that they often contain "massive dark objects"—concentrations of 10^6–$10^9\,\mathcal{M}_\odot$ contained within a few pc of the center (BM §11.2.2). The best-studied of these objects, the one at the center of the Galaxy, has a mass of $(3.9 \pm 0.3) \times 10^6\,\mathcal{M}_\odot$ contained within a radius less than 0.001 pc. Astronomers believe that these objects must be black holes, for two main reasons. First, dynamical arguments show that no long-lived astrophysical system other than a black hole could be so massive and so small (§7.5.2). Second, many galaxies contain strong sources of non-stellar radiation at their centers, called **active galactic nuclei** or

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

AGN; the most luminous and rare of these, the quasars, can achieve luminosities of $10^{13}\,L_\odot$ and outshine their host galaxies by two orders of magnitude (see BM §4.6.2 and Krolik 1999). By far the most plausible power source for AGN is accretion onto a massive black hole, and the demography of massive dark objects in galaxy centers is roughly consistent with the hypothesis that these are dormant AGN.

It appears that most galaxies—or at least most early-type galaxies—contain a central black hole. The mass of the black hole typically amounts to ≈ 0.001–0.002 of the total mass of the stars in the host galaxy (Häring & Rix 2004). Another correlation that is more directly observable is between the black-hole mass $M_\bullet$ and the velocity dispersion $\sigma_\parallel$ near the center of the host galaxy,

$$\log_{10}\left(\frac{M_\bullet h_7}{10^8\,\mathcal{M}_\odot}\right) = (4\pm0.3)\log_{10}\left(\frac{\sigma_\parallel}{200\,\mathrm{km\,s^{-1}}}\right) + (0.2\pm0.1). \qquad (1.27)$$

The RMS scatter in this relation is $\lesssim 0.3$ in $\log_{10} M_\bullet$ (Tremaine et al. 2002).

Massive black holes are probably formed at the centers of galaxies. When galaxies merge, their black holes are dragged to the center of the merged galaxy by dynamical friction (§8.1.1a). If the resulting binary black hole is so tightly bound that it continues to decay by gravitational radiation, the two black holes will eventually merge. The final stages of this merger could provide a powerful source of gravitational radiation (§8.1.1e; Begelman, Blandford & Rees 1980).

1.2 Collisionless systems and the relaxation time

There is a fundamental difference between galaxies and the systems that are normally dealt with in statistical mechanics, such as molecules in a box. This difference lies in the nature of the forces that act between the constituent particles. The interaction between two molecules is short-range: the force is small unless the molecules are very close to each other, when it becomes strongly repulsive. Consequently, molecules in a diffuse gas are subject to violent and short-lived accelerations as they collide with one another, interspersed with much longer periods when they move at nearly constant velocity. In contrast, the gravitational force that acts between the stars of a galaxy is long-range.

Consider the force from the stars in the cone shown in Figure 1.4 on a star at the apex of the cone. The force from any one star falls off with distance r as r^{-2}, but if the density of stars is uniform, the number of attracting stars per unit length of the cone increases as r^2. Let us call a factor of two interval in radius an **octave**, by analogy with the musical octave. Then each octave in radius, from r to $2r$, has a length proportional to r, so each octave attracts

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

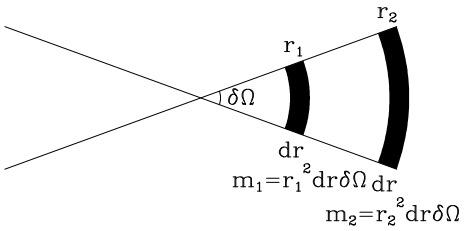

Figure 1.4 If the density of stars were everywhere the same, the stars in each of the shaded segments of a cone would contribute equally to the force on a star at the cone's apex. Thus the acceleration of a star at the apex is determined mainly by the large-scale distribution of stars in the galaxy, not by the star's nearest neighbors.

the star at the apex with a force proportional to $r^{-2} \times r^2 \times r = r$. This simple argument shows that the force on the star at the apex is dominated by the most distant stars in the system, rather than by its closest neighbors. Of course, if the density of attracting stars were exactly spherical, the star at the apex would experience no net force because it would be pulled equally in all directions. But in general the density of attracting stars falls off in one direction more slowly than in the opposing direction, so the star at the apex is subject to a net force, and this force is determined by the structure of the galaxy on the largest scale. Consequently—in contrast to the situation for molecules—the force on a star does not vary rapidly, and each star may be supposed to accelerate smoothly through the force field that is generated by the galaxy as a whole. In other words, for most purposes we can treat the gravitational force on a star as arising from a smooth density distribution rather than a collection of mass points.

1.2.1 The relaxation time

We now investigate this conclusion more quantitatively, by asking how accurately we can approximate a galaxy composed of N identical stars of mass m as a smooth density distribution and gravitational field. To answer this question, we follow the motion of an individual star, called the **subject star**, as its orbit carries it once across the galaxy, and seek an order-of-magnitude estimate of the difference between the actual velocity of this star after this interval and the velocity that it would have had if the mass of the other stars were smoothly distributed. Suppose the subject star passes within distance b of another star, called the **field star** (Figure 1.5). We want to estimate the amount $\delta\mathbf{v}$ by which the encounter deflects the velocity $\mathbf{v}$ of the subject star. In §3.1d we calculate $\delta\mathbf{v}$ exactly, but for our present purposes an approximate estimate is sufficient. To make this estimate we shall assume that $|\delta\mathbf{v}|/v \ll 1$, and that the field star is stationary during the encounter. In this case $\delta\mathbf{v}$ is perpendicular to $\mathbf{v}$, since the accelerations parallel to $\mathbf{v}$ average to zero. We may calculate the magnitude of the velocity change, $\delta v \equiv |\delta\mathbf{v}|$, by assuming that the subject star passes the field star on a straight-line trajectory, and integrating the perpendicular force $F_\perp$ along this trajectory. We place the origin of time at the instant of closest approach of the two stars,

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

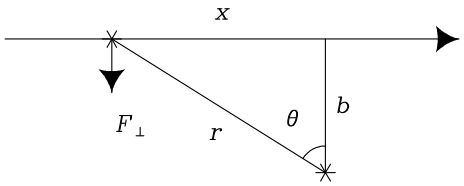

Figure 1.5 A field star approaches the subject star at speed v and impact parameter b. We estimate the resulting impulse to the subject star by approximating the field star's trajectory as a straight line.

and find in the notation of Figure 1.5,

$$F_\perp = \frac{Gm^2}{b^2+x^2}\cos\theta = \frac{Gm^2 b}{(b^2+x^2)^{3/2}} = \frac{Gm^2}{b^2}\left[1+\left(\frac{vt}{b}\right)^2\right]^{-3/2}. \tag{1.28}$$

But by Newton's laws

$$m\dot{\mathbf{v}} = \mathbf{F} \quad \text{so} \quad \delta v = \frac{1}{m}\int_{-\infty}^{\infty} \mathrm{d}t\, F_\perp, \tag{1.29}$$

and we have

$$\delta v = \frac{Gm}{b^2}\int_{-\infty}^{\infty}\frac{\mathrm{d}t}{[1+(vt/b)^2]^{3/2}} = \frac{Gm}{bv}\int_{-\infty}^{\infty}\frac{\mathrm{d}s}{(1+s^2)^{3/2}} = \frac{2Gm}{bv}. \tag{1.30}$$

In words, δv is roughly equal to the acceleration at closest approach, Gm/b^2, times the duration of this acceleration $2b/v$. Notice that our assumption of a straight-line trajectory breaks down, and equation (1.30) becomes invalid, when $\delta v \simeq v$; from equation (1.30), this occurs if the impact parameter $b \lesssim b_{90} \equiv 2Gm/v^2$. The subscript 90 stands for a 90-degree deflection—see equation (3.51) for a more precise definition.

Now the surface density of field stars in the host galaxy is of order $N/\pi R^2$, where N is the number of stars and R is the galaxy's radius, so in crossing the galaxy once the subject star suffers

$$\delta n = \frac{N}{\pi R^2}2\pi b\,\mathrm{d}b = \frac{2N}{R^2}b\,\mathrm{d}b \tag{1.31}$$

encounters with impact parameters in the range b to $b+\mathrm{d}b$. Each such encounter produces a perturbation $\delta\mathbf{v}$ to the subject star's velocity, but because these small perturbations are randomly oriented in the plane perpendicular to $\mathbf{v}$, their mean is zero.[10] Although the mean velocity change is zero, the mean-square change is not: after one crossing this amounts to

$$\sum \delta v^2 \simeq \delta v^2 \delta n = \left(\frac{2Gm}{bv}\right)^2 \frac{2N}{R^2}b\,\mathrm{d}b. \tag{1.32}$$

[10] Strictly, the mean change in velocity is zero only if the distribution of perturbing stars is the same in all directions. A more precise statement is that the mean change in velocity is due to the smoothed-out mass distribution, and we ignore this because the goal of our calculation is to determine the *difference* between the acceleration due to the smoothed mass distribution and the actual stars.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Integrating equation (1.32) over all impact parameters from $b_{\rm min}$ to $b_{\rm max}$, we find the mean-square velocity change per crossing,

$$\Delta v^2 \equiv \int_{b_{\rm min}}^{b_{\rm max}} \sum \delta v^2 \simeq 8N \left(\frac{Gm}{Rv}\right)^2 \ln\Lambda, \tag{1.33a}$$

where the factor

$$\ln\Lambda \equiv \ln\left(\frac{b_{\rm max}}{b_{\rm min}}\right) \tag{1.33b}$$

is called the **Coulomb logarithm**. Our assumption of a straight-line trajectory breaks down for impact parameters smaller than b_{90}, so we set $b_{\rm min} = f_1 b_{90}$, where f_1 is a factor of order unity. Our assumption of a homogeneous distribution of field stars breaks down for impact parameters of order R, so we set $b_{\rm max} = f_2 R$. Then

$$\ln\Lambda = \ln\left(\frac{R}{b_{90}}\right) + \ln(f_2/f_1). \tag{1.34}$$

In most systems of interest $R \gg b_{90}$ (for example, in a typical elliptical galaxy $R/b_{90} \gtrsim 10^{10}$), so the fractional uncertainty in $\ln\Lambda$ arising from the uncertain values of f_1 and f_2 is quite small, and we lose little accuracy by setting $f_2/f_1 = 1$.

Thus encounters between the subject star and field stars cause a kind of diffusion of the subject star's velocity that is distinct from the steady acceleration caused by the overall mass distribution in the stellar system. This diffusive process is sometimes called **two-body relaxation** since it arises from the cumulative effect of myriad two-body encounters between the subject star and passing field stars.

The typical speed v of a field star is roughly that of a particle in a circular orbit at the edge of the galaxy,

$$v^2 \approx \frac{GNm}{R}. \tag{1.35}$$

If we eliminate R from equation (1.33a) using equation (1.35), we have

$$\frac{\Delta v^2}{v^2} \approx \frac{8\ln\Lambda}{N}. \tag{1.36}$$

If the subject star makes many crossings of the galaxy, the velocity $\mathbf{v}$ will change by roughly Δv^2 at each crossing, so the number of crossings $n_{\rm relax}$ that is required for its velocity to change by of order itself is given by

$$n_{\rm relax} \simeq \frac{N}{8\ln\Lambda}. \tag{1.37}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

The **relaxation time** may be defined as $t_{\rm relax} = n_{\rm relax} t_{\rm cross}$, where $t_{\rm cross} = R/v$ is the **crossing time**, the time needed for a typical star to cross the galaxy once. Moreover $\Lambda = R/b_{90} \approx Rv^2/(Gm)$, which is $\approx N$ by equation (1.35). Thus our final result is

$$t_{\rm relax} \simeq \frac{0.1N}{\ln N} t_{\rm cross}. \tag{1.38}$$

After one relaxation time, the cumulative small kicks from many encounters with passing stars have changed the subject star's orbit significantly from the one it would have had if the gravitational field had been smooth. In effect, after a relaxation time a star has lost its memory of its initial conditions. Galaxies typically have $N \approx 10^{11}$ stars and are a few hundred crossing times old, so for these systems stellar encounters are unimportant, except very near their centers. In a globular cluster, on the other hand, $N \approx 10^5$ and the crossing time $t_{\rm cross} \approx 1\,{\rm Myr}$ (Table 1.3), so relaxation strongly influences the cluster structure over its lifetime of 10 Gyr.

In all of these systems the dynamics over timescales $\lesssim t_{\rm relax}$ is that of a **collisionless system** in which the constituent particles move under the influence of the gravitational field generated by a smooth mass distribution, rather than a collection of mass points. Non-baryonic dark matter is also collisionless, since both weak interactions and gravitational interactions between individual WIMPs are negligible in any galactic context.

In most of this book we focus on collisionless stellar dynamics, confining discussion of the longer-term evolution that is driven by gravitational encounters among the particles to Chapter 7.

1.3 The cosmological context

This section provides a summary of the aspects of cosmology that we use in this book. For more information the reader can consult texts such as Weinberg (1972), Peebles (1993), and Peacock (1999).

To a very good approximation, the universe is observed to be homogeneous and isotropic on large scales—here "large" means $\gtrsim 100\,{\rm Mpc}$, which is still much smaller than the characteristic "size" of the universe, the **Hubble length** $c/H_0 = 4.3 h_7^{-1}\,{\rm Gpc}$ where $1\,{\rm Gpc} = 10^9\,{\rm pc} = 10^3\,{\rm Mpc}$ and c is the speed of light. Therefore a useful first approximation is to average over the small-scale structure and treat the universe as *exactly* homogeneous and isotropic. Of course, the universe does not appear isotropic to all observers: an observer traveling rapidly with respect to the local matter will see galaxies approaching in one direction and receding in another. Therefore we define a set of **fundamental observers**, who are at rest with respect to the matter around them.[11] The universe is expanding, so we may synchronize the

[11] A more precise definition is that a fundamental observer sees no dipole component in the cosmic microwave background radiation (§1.3.5).

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

clocks of the fundamental observers by setting them to the same time at the moment when the universal homogeneous density has some particular value. This procedure enables us to define a unique **cosmic time** (Gunn 1978).

The random velocities of galaxies, and the velocities of stars within galaxies (a few hundred $\mathrm{km\,s^{-1}}$), are small compared to the relative velocities of galaxies that are separated by 100 Mpc (many thousand $\mathrm{km\,s^{-1}}$). Thus, on scales large enough that the assumption of homogeneity is accurate, any observer who moves with a typical star in a galaxy is a fundamental observer.

1.3.1 Kinematics

Consider the triangle defined by three nearby fundamental observers. As the universe evolves, the triangle may change in size, but cannot change in shape or orientation—in the contrary case, it would define a preferred direction, thereby violating the isotropy assumption. Thus, if $r_{ij}(t)$ is the length of the side joining observers i and j at cosmic time t, we must have $r_{ij}(t) = r_{ij}(t_0)a(t)$, where $a(t)$ is independent of i and j. Since this argument holds for all fundamental observers, the distance between any two of them must have the form

$$r(t) = r(t_0)a(t), \tag{1.39}$$

where the **scale factor** $a(t)$ is a universal function, which we may normalize so that $a(t_0) = 1$ at the present cosmic time t_0. The relative velocity of the two observers is

$$v(t) = \frac{\mathrm{d}r}{\mathrm{d}t} = r(t_0)\dot{a}(t) = r(t)\frac{\dot{a}(t)}{a(t)} \equiv r(t)H(t), \tag{1.40}$$

where $H(t)$ is the **Hubble parameter**. At the present time, $H(t_0) \equiv H_0$ is the Hubble constant, and equation (1.40) is a statement of the Hubble law (1.13). Thus we see that (i) the Hubble law is a consequence of homogeneity and isotropy; (ii) in a homogeneous, isotropic universe the Hubble law remains true at all times but the Hubble "constant" varies with cosmic time.

Next consider a photon that at cosmic time t passes a fundamental observer, who observes it to have frequency ν. After an infinitesimal time interval $\mathrm{d}t$, the photon has traveled a distance $\mathrm{d}r = c\,\mathrm{d}t$ and hence is overtaking a second fundamental observer who is moving away from the first at speed $\mathrm{d}v = H(t)\mathrm{d}r = H(t)c\,\mathrm{d}t$. This observer will measure a different frequency for the photon because of the Doppler shift. The measured frequency will be $\nu(1-\mathrm{d}v/c) = \nu[1-H(t)\mathrm{d}t]$; the use of this first-order formula is justified because $\mathrm{d}v$ is infinitesimal. Thus the frequency of a propagating photon as measured by a local fundamental observer decreases at a rate

$$\frac{\mathrm{d}\nu}{\mathrm{d}t} = -H(t)\nu \quad \text{or} \quad \frac{\dot{\nu}}{\nu} = -\frac{\dot{a}}{a}, \quad \text{thus} \quad \nu(t) \propto \frac{1}{a(t)}. \tag{1.41}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

In words, a photon emitted by a fundamental observer at frequency $\nu_{\rm e}$ and wavelength $\lambda_{\rm e} = c/\nu_{\rm e}$, and received by a second fundamental observer at the present time t_0, will be observed to have frequency ν_0 and wavelength λ_0 given by

$$\frac{\nu_{\rm e}}{\nu_0} = \frac{\lambda_0}{\lambda_{\rm e}} = \frac{a(t_0)}{a(t_{\rm e})} = \frac{1}{a(t_{\rm e})} \equiv 1 + z. \tag{1.42}$$

Here z is the **redshift**. Redshift is often used instead of time t to describe the cosmic time of an event, since redshift is directly observable from the wavelengths of known spectral lines, whereas the relation between time and redshift depends on the cosmological model (Figure 1.7).

These derivations assume only that spacetime is locally Euclidean, and thus they are correct even in a curved spacetime, so long as it is homogeneous and isotropic.

1.3.2 Geometry

Let the position of any fundamental observer be labeled by time-independent coordinates (q_1, q_2, q_3). At a given cosmic time t, the distance dl between the observers at (q_1, q_2, q_3) and $(q_1 + {\rm d}q_1, q_2 + {\rm d}q_2, q_3 + {\rm d}q_3)$ can be written in the form

$${\rm d}r^2 = a^2(t) h_{ij} {\rm d}q_i {\rm d}q_j, \tag{1.43}$$

where we have used the summation convention (page 772), and the metric tensor h_{ij} (cf. eq. B.13) must be independent of time in a homogeneous, isotropic universe. It can be shown that homogeneity and isotropy also imply that the q_i can be chosen such that equation (1.43) takes the form of the **Robertson–Walker metric**,

$${\rm d}r^2 = a^2(t) \left[\frac{{\rm d}x^2}{1 - kx^2/x_{\rm u}^2} + x^2 \left({\rm d}\theta^2 + \sin^2\theta \, {\rm d}\phi^2 \right) \right]. \tag{1.44}$$

Here θ and ϕ are the usual angles in spherical coordinates (Appendix B.2), x is a radial coordinate, $x_{\rm u}$ is a constant called the **radius of curvature**, and k is $+1$, 0 or -1. Since x remains fixed as the fundamental observers recede from one another, it is called a **comoving coordinate**.

In the case $k = 0$, the metric (1.44) corresponds to ordinary spherical polar coordinates (cf. eq. B.32), so at a given cosmic time the geometry of the universe is that of ordinary Euclidean or **flat** space. In the case $k = +1$, (1.44) is the three-dimensional generalization of the metric on the surface of a sphere of radius $x_{\rm u}$, ${\rm d}r^2 = {\rm d}x^2/(1 - x^2/x_{\rm u}^2) + x^2 {\rm d}\phi^2$, where x is the perpendicular distance from the polar axis to the point in question. This case is said to represent a **closed universe**, since the volume of space is finite (Problem 1.7). The case $k = -1$ has no analogous 2-surface embedded in Euclidean 3-space (e.g.,Weinberg 1972). It represents an **open universe** with infinite volume.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

1.3.3 Dynamics

The evolution of the scale factor $a(t)$ is determined by the equations of general relativity and the equation of state of the material in the universe. We shall assume that all of the major components of this material can be described as (possibly relativistic) fluids. To derive the equations governing $a(t)$ we then need only one result from relativity: that a fluid with inertial mass density ρ and pressure p has a gravitational mass density (Problem 9.5)

$$\rho' = \rho + \frac{3p}{c^2}. \tag{1.45}$$

By isotropy, the universe is spherically symmetric as viewed by any fundamental observer. Now draw a sphere of radius r around such an observer, where r is large enough that the approximation of homogeneity and isotropy is valid, but small enough that Newtonian physics applies within it. As shown at the beginning of this section, in practice this means $100\,\mathrm{Mpc} \ll r \ll 4000\,\mathrm{Mpc}$. By analogy with Newton's famous theorem that a body experiences no gravitational force from a spherical shell of matter outside it (§2.2.1), we ignore the effects of material outside the sphere. Then Newton's law of gravity tells us that a fundamental observer on the surface of the sphere is accelerated towards its center at a rate

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = -\frac{GM}{r^2}, \tag{1.46}$$

where M is the gravitational mass inside the sphere. Note that there are no pressure forces, since $\boldsymbol{\nabla} p = 0$ by homogeneity. Since $M = \rho' V$, where $V = \frac{4}{3}\pi r^3$, and $r = r_0 a(t)$, we may rewrite equation (1.46) as

$$\frac{\ddot{a}}{a} = -\frac{4\pi G \rho'}{3} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right). \tag{1.47}$$

To integrate this equation, we need to know how p and ρ vary with the scale factor $a(t)$. The internal energy of the sphere, including its rest-mass energy, is $U = \rho c^2 V$. The material satisfies $\mathrm{d}U + p\,\mathrm{d}V = 0$ (eq. F.22 with $\mathrm{d}S = 0$, since there is no heat flow in a homogeneous, isotropic universe), so

$$c^2 \mathrm{d}(\rho V) + p\,\mathrm{d}V = 0 \quad \text{or} \quad \mathrm{d}\rho + \left(\rho + \frac{p}{c^2}\right)\frac{\mathrm{d}V}{V} = 0. \tag{1.48}$$

Since $V \propto a^3(t)$, we have $\mathrm{d}V/V = 3\,\mathrm{d}a/a$, and equations (1.47) and (1.48) can be combined to eliminate p:

$$\frac{\ddot{a}}{a} = \frac{4\pi G}{3}\left(2\rho + a\frac{\mathrm{d}\rho}{\mathrm{d}a}\right). \tag{1.49}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

After multiplying by $a\dot{a}$ this equation can be integrated to yield

$$\dot{a}^2 - \frac{8\pi G\rho}{3}a^2 = 2E, \tag{1.50}$$

where E is a constant of integration, analogous to the Newtonian energy.

These equations can also be derived directly from general relativity. The relativistic derivation also connects the geometry to the energy density, by relating the parameters of the Robertson–Walker metric (1.44) to the integration constant E:

$$2E = -\frac{kc^2}{x_{\rm u}^2}. \tag{1.51}$$

Equations (1.44), (1.50), and (1.51) specify the **Friedmann–Robertson–Walker** or **FRW** model of the universe.

When $k = 0$, space is flat, $E = 0$, and the density equals the **critical density**

$$\rho_{\rm c}(t) \equiv \frac{3\dot{a}^2}{8\pi G a^2} = \frac{3H^2(t)}{8\pi G}. \tag{1.52}$$

If we define the **density parameter**

$$\Omega(t) \equiv \frac{\rho(t)}{\rho_{\rm c}(t)} = \frac{8\pi G\rho(t)}{3H^2(t)}, \tag{1.53}$$

then equation (1.50) can be written

$$\Omega^{-1} - 1 = \frac{3E}{4\pi G\rho a^2} = -\frac{3kc^2}{8\pi G\rho a^2 x_{\rm u}^2}. \tag{1.54}$$

This result implies that if $\Omega < 1$ at any time, it always remains so; this case corresponds to a universe that is open ($k = -1$) and infinite. In contrast, if Ω exceeds unity it always remains so, and we have a universe that is closed ($k = +1$) and finite. Finally, if $\Omega = 1$ at any instant it is unity for all time, and the universe is always flat. Thus the geometry of the universe is determined by its mass content, and an open universe cannot turn into a closed one or vice versa.

The present value of the critical density is

$$\rho_{\rm c0} = \rho_{\rm c}(t_0) = 9.204 \times 10^{-27}\, h_7^2\,{\rm kg\,m^{-3}} = 1.3599 \times 10^{11}\, h_7^2\, \mathcal{M}_\odot\,{\rm Mpc^{-3}}. \tag{1.55}$$

We have parametrized galaxies and other stellar systems by their mass-to-light ratio. Since the universe is homogeneous, the mass-to-light ratio measured on sufficiently large scales must be the same everywhere at a given cosmic time. The present R-band luminosity density j_R is given by equation (1.19), so the R-band mass-to-light ratio Υ_R is related to the density parameter by

$$\Upsilon_R = \frac{\rho_{\rm c0}\Omega_0}{j_R} = (900 \pm 300)\, h_7\Omega_0\, \Upsilon_\odot; \tag{1.56}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

the subscript "0" on Ω indicates the density parameter at the present epoch.

An obvious next step is to compare this result to observed mass-to-light ratios and thereby estimate Ω_0. Unfortunately, the total mass-to-light ratios of individual galaxies are quite uncertain, because the total mass contained in their dark halos is difficult to determine. However, as we argued after equation (1.26), it is likely that the mixture of baryonic and non-baryonic dark matter in rich clusters of galaxies is representative of the universe as a whole, so we might hope that the mass-to-light ratios of rich clusters are a reasonable approximation to the total mass-to-light ratios of galaxies. Taking $\Upsilon_R \simeq (200 \pm 50)\, h_7 \Upsilon_\odot$ from equation (1.25), we conclude that $\Omega_0 = 0.22 \pm 0.09$. This argument is subject to at least two possible biases: first, the galaxy population in rich clusters has a higher fraction of ellipticals than average, and hence fewer luminous young stars; second, most of the baryons in rich clusters are in the form of hot gas, which does not contribute to the R-band luminosity, but in isolated galaxies this gas might cool to form additional stars. Both of these effects should increase the mass-to-light ratio in clusters relative to isolated galaxies, and hence lead us to overestimate Ω_0. Thus we can conclude from this argument only that

$$\Omega_{\rm m0} \lesssim 0.3. \tag{1.57}$$

The subscript "m" on Ω_0 is a reminder that this value refers only to the density in non-relativistic matter that can collapse along with the baryons into clusters of galaxies. Any uniformly distributed component of the density, such as a population of relativistic particles or vacuum energy, is excluded.

The inequality (1.57) encapsulates one of the fundamental conclusions of modern cosmology: the most "natural" model, a matter-dominated flat universe in which $\Omega = 1$ at all times, is excluded by the observations.

To solve the differential equations that describe FRW models, we need to know the equation of state relating the pressure p to the density ρ for each component of the universe. For our purposes a sufficiently general parametrization is

$$p = w\rho c^2, \tag{1.58}$$

where w is a constant. If the equation of state has this form, equation (1.48) can be integrated to yield

$$\rho \propto V^{-1-w} \propto a^{-3(1+w)}. \tag{1.59}$$

Three major components contribute to the dynamics of the universe:

(i) Non-relativistic matter. This has $p \ll \rho c^2$ so $w = 0$. We label the corresponding density $\rho_{\rm m}(t)$, and for brevity we simply call this component "matter." In this case there is no distinction between the inertial mass density $\rho_{\rm m}$ and the gravitational density $\rho'_{\rm m} = \rho_{\rm m}$ (eq. 1.45). From equation (1.59) $\rho_{\rm m} \propto a^{-3}$, as expected from conservation of mass.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

(ii) Radiation and other massless or highly relativistic particles. We label this density $\rho_\gamma(t)$, and call this component "radiation." In this case $p = \frac{1}{3}\rho c^2$ so $w = \frac{1}{3}$, and from equation (1.45) the gravitational attraction is twice as strong as non-relativistic matter with the same density: $\rho'_\gamma = 2\rho_\gamma$. As the universe expands, the radiation density declines as $\rho_\gamma \propto a^{-4}$ from equation (1.59). Physically, this dependence arises because the number of photons is conserved so the number density declines as a^{-3}, and their frequency and thus the energy per photon decay as a^{-1} (eq. 1.42).

(iii) A hypothetical energy density ρ_Λ associated with the vacuum (Carroll, Press, & Turner 1992). This must be accompanied by a *negative* pressure $p = -\rho_\Lambda c^2$ (i.e., a tension) because the energy-momentum tensor of the vacuum must be proportional to the Minkowski metric if the vacuum is to appear the same to all observers, regardless of their relative motion. In the parametrization of equation (1.58), vacuum energy therefore has $w = -1$. Equation (1.59) shows that as the universe expands, ρ_Λ is independent of the scale factor, as expected since it is a universal constant.

A remarkable feature of vacuum energy is that it exerts repulsive gravitational forces—equations (1.45) and (1.58) show that gravity is repulsive for any medium with $w < -\frac{1}{3}$. Consequently, the gravity from such a medium tends to accelerate rather than decelerate the expansion of the universe.

Vacuum energy plays a significant role in cosmology only if the vacuum-energy density ρ_Λ is comparable to the current critical density $\rho_{\rm c}$ (eq. 1.55). There is no motivation from fundamental physics for a vacuum-energy density of this magnitude: the theoretical prejudice is that either ρ_Λ has a very large value, or else is exactly zero on account of some unidentified symmetry. There is no known mechanism that would favor a value of ρ_Λ comparable to $\rho_{\rm c}$. Moreover, because the critical density evolves with time while the vacuum-energy density does not, any approximate coincidence between ρ_Λ and $\rho_{\rm c}$ must be a special feature of the present epoch. These difficulties have led physicists to explore quantum fields with more general behavior than vacuum energy, under the general heading of **dark energy**.

By analogy with equation (1.53), we define

$$\Omega_{\rm m0} \equiv \frac{\rho_{\rm m0}}{\rho_{\rm c0}} \quad ; \quad \Omega_{\gamma 0} \equiv \frac{\rho_{\gamma 0}}{\rho_{\rm c0}} \quad ; \quad \Omega_{\Lambda 0} \equiv \frac{\rho_{\Lambda 0}}{\rho_{\rm c0}} \tag{1.60}$$

to be the present densities of matter, radiation, and vacuum energy in units of the critical density. With this notation $\Omega_{\rm m0} + \Omega_{\gamma 0} + \Omega_{\Lambda 0} = \Omega_0$. Then equation (1.50) can be rewritten as

$$\dot{a}^2 = H_0^2 \left[1 + \Omega_{\rm m0}(a^{-1} - 1) + \Omega_{\gamma 0}(a^{-2} - 1) + \Omega_{\Lambda 0}(a^2 - 1)\right], \tag{1.61}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

which can be integrated to yield a formula for the time dependence of the scale factor $a(t)$:

$$H_0 t = \int_0^{a(t)} \frac{a\,\mathrm{d}a}{\sqrt{\Omega_{\gamma 0} + \Omega_{\mathrm{m}0} a + (1 - \Omega_{\mathrm{m}0} - \Omega_{\gamma 0} - \Omega_{\Lambda 0})a^2 + \Omega_{\Lambda 0} a^4}}. \quad (1.62)$$

This integral can be evaluated analytically or numerically for arbitrary values of $\Omega_{\mathrm{m}0}$, $\Omega_{\gamma 0}$, and $\Omega_{\Lambda 0}$, but it is more illuminating to examine special cases:

(i) A flat, matter-dominated universe (the **Einstein-de Sitter universe**) has $\Omega_{\gamma 0} = \Omega_{\Lambda 0} = 0$, $\Omega_{\mathrm{m}0} = 1$, so

$$a(t) \propto t^{2/3} \quad ; \quad \rho_{\mathrm{m}}(t) = \frac{1}{6\pi G t^2}; \quad (1.63)$$

the second equation follows when the first is substituted into equation (1.50) with $E = 0$.

(ii) A flat, radiation-dominated universe has

$$a(t) \propto t^{1/2} \quad ; \quad \rho_{\gamma}(t) = \frac{3}{32\pi G t^2}. \quad (1.64)$$

(iii) A flat universe dominated by vacuum energy has

$$a(t) \propto \exp(H_0 t) = \exp\left[(\tfrac{8}{3}\pi G \rho_\Lambda)^{1/2} t\right] \quad ; \quad \rho_\Lambda = \frac{3H_0^2}{8\pi G} = \text{constant}. \quad (1.65)$$

At the present time, $\Omega_{\gamma 0} \simeq 10^{-4}$ (eq. 1.72), so the evolution of the universe is determined by $\Omega_{\mathrm{m}0}$ and $\Omega_{\Lambda 0}$ except at very early times. Thus the properties of the universe can be parametrized on a diagram such as Figure 1.6. Lines on this figure mark the boundary between models that have open or closed geometries ($k = -1$ or $k = +1$), and between models that expand forever and those that collapse at some future time. We have also marked off models that have no initial singularity: as we follow these "bounce" models back in time from the present, the repulsion from the vacuum energy becomes so strong that the expansion rate $\dot{a}$ reaches zero at some time t_{b} and is negative for $t < t_{\mathrm{b}}$. Physically, this means that the universe was contracting for $t < t_{\mathrm{b}}$, coasted to a halt because of increasing repulsion by the vacuum energy, and then began the expansion that continues at the present time. Such models are excluded by observations because they predict a maximum redshift, $z \simeq 2$, much smaller than the largest observed redshifts, $z \gtrsim 6$.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

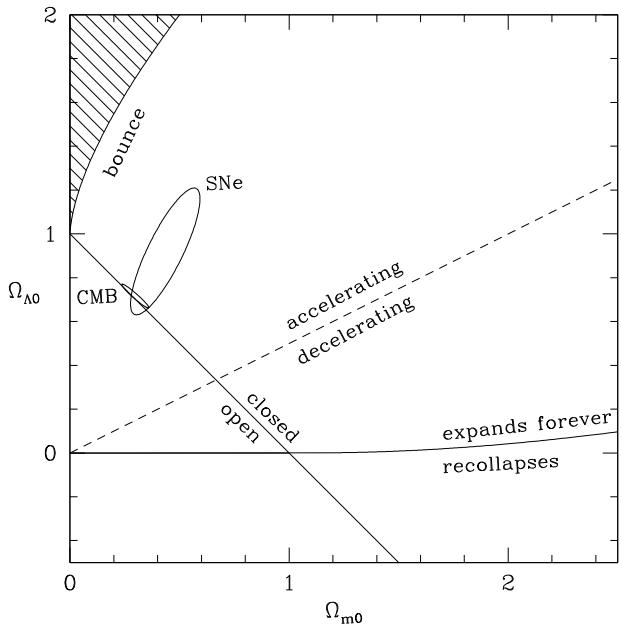

Figure 1.6 Characteristics of FRW models of the universe in which the current radiation density $\Omega_{\gamma 0}$ is negligible. The solutions are parametrized by the current matter density $\Omega_{\rm m0}$ and vacuum energy $\Omega_{\Lambda 0}$, both relative to the critical density (1.52). The lines divide models in which the geometry is open from those with closed geometry; models that will expand forever from those that will eventually collapse; and models in which the expansion is accelerating ($\ddot{a} > 0$) from those in which it is decelerating. The shaded region denotes models with no initial singularity or Big Bang; these bounced from a collapsing to an expanding state at some non-zero value $a(t) < 1$ of the scale factor. The large oval marked SNe is the 1-σ error ellipse from measurements of distant supernovae (Riess et al. 2004), and the small oval labeled CMB is the 1-σ error ellipse from measurements of fluctuations in the cosmic microwave background combined with local measurements of the Hubble constant (Spergel et al. 2007). For further discussion see page 50.

1.3.4 The Big Bang and inflation

There is strong evidence that the universe was much hotter and denser in the past—this evidence includes the existence of the cosmic microwave background and the primordial abundances of the light elements (see §1.3.5). This is consistent with the discussion of the preceding paragraph, which shows that all FRW models that are consistent with observations begin from an initial singularity or **Big Bang**. Immediately after the Big Bang the universe satisfied several striking constraints:

(i) A matter- or radiation-dominated FRW universe always evolves away from $\Omega = 1$: equation (1.54) shows that $|\Omega^{-1} - 1|$ grows in proportion to $1/\rho a^2$, which grows as $a(t)$ or $a^2(t)$, respectively. At present Ω is not

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

far from unity, so soon after the Big Bang, when $a(t)$ was much smaller than now, Ω must have been extremely close to unity. This fine-tuning of the initial conditions is called the "flatness problem."

(ii) The universe is homogeneous on large scales. It is natural to ask whether this property could be the result of physical processes occurring shortly after the Big Bang. Since information cannot propagate faster than the speed of light, the size of the largest causally connected region is given by the distance a photon can propagate since the Big Bang. Light travels at speed $c = \mathrm{d}r/\mathrm{d}t$, where $\mathrm{d}r$ is the distance. In the Robertson–Walker metric (1.44), the comoving coordinate of a photon that is moving towards the origin therefore satisfies

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{c}{a(t)}\left(1 - \frac{kx^2}{x_{\mathrm{u}}^2}\right)^{1/2}. \tag{1.66}$$

We have seen that in the early universe, $|\Omega - 1|$ must have been very small, so the geometry is nearly flat and we may set $k = 0$. Thus a photon that is emitted at t_{i} and arrives at the origin at t has come from a comoving coordinate

$$x = c\int_{t_{\mathrm{i}}}^{t} \frac{\mathrm{d}t}{a(t)}. \tag{1.67}$$

The comoving radius of the region that has been in causal contact since the Big Bang is called the **particle horizon** $x_{\mathrm{h}}(t)$, and is obtained from equation (1.67) by letting t_{i} shrink to zero. At early times the universe is expected to be radiation-dominated, since $\rho_\gamma(t) \propto a^{-4}(t)$ while $\rho_{\mathrm{m}} \propto a^{-3}$ and ρ_Λ is constant. In a flat radiation-dominated universe, the scale factor is $a(t) = bt^{1/2}$, where b is a constant (eq. 1.64). Thus we obtain

$$x_{\mathrm{h}}(t) = \frac{2ct^{1/2}}{b}, \tag{1.68}$$

which shows that the comoving horizon shrinks to zero as $t \to 0$; this in turn implies that right after the Big Bang different parts of the universe are not causally connected, so the large-scale homogeneity of the universe must be imposed as an initial condition. This uncomfortable situation is called the "horizon problem."

(iii) The rich structure in the present universe—galaxies, clusters of galaxies, etc.—has grown by gravitational instability. The spectrum of perturbations that seeded this growth must be inserted as an initial condition (the "structure problem").

Remarkably, all three of these problems can be resolved at one stroke by a single powerful assumption: that the early universe underwent a phase of accelerated expansion or **inflation**. Inflation arises when the dominant contributor to the mass density has an equation of state with $w < -\frac{1}{3}$ (eq. 1.58).

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For example, suppose that there is an early inflationary phase in which the universe is dominated by a large vacuum energy, which has $w = -1$. Then the scale factor grows exponentially with time, as described by equation (1.65), but with much larger constants H_0 and ρ_Λ. As the universe inflates, the density of non-relativistic matter falls as $a^{-3}(t)$, and the density of photons and other relativistic matter falls as $a^{-4}(t)$, while the vacuum-energy density remains constant. Hence the dynamics rapidly becomes completely dominated by the vacuum energy, equation (1.65) applies with ever greater precision, and the density parameter Ω tends to unity. More precisely, if the inflationary phase lasts for n e-foldings of the scale factor, then at the end of this phase $|\Omega - 1|$ will be smaller by a factor $\sim \exp(-2n)$ than it was at the beginning. Thus inflation naturally produces Ω very close to unity, thereby solving the flatness problem. Moreover, since $|\Omega - 1|$ is zero at the end of inflation with exponential precision, it is plausible to assume that it is exactly zero for all practical purposes, even today. Thus inflation strongly suggests that $\Omega_0 = 1$; in other words, $\Omega_{\gamma 0} + \Omega_{\rm m0} + \Omega_{\Lambda 0} = 1$. Since $\Omega_{\gamma 0}$ is negligible, the universe must lie on the line $\Omega_{\rm m0} + \Omega_{\Lambda 0} = 1$ that separates open from closed universes in Figure 1.6, just as the observations seem to indicate.

Next, inflation solves the horizon problem: equation (1.65) tells us that $\dot{a} \propto a$, so the integral in equation (1.67) is $\int \mathrm{d}t/a = \int \mathrm{d}a/(a\dot{a}) \propto \int \mathrm{d}a/a^2$, which diverges as $a \to 0$. Thus the region that is in causal contact becomes arbitrarily large if the inflationary phase begins early enough, so all regions in the observable universe could have been in causal contact in the interval between the Big Bang and the onset of inflation.

Inflation also predicts that when quantum fluctuations in the matter density are inflated past the event horizon, they are frozen into classical density fluctuations that provide the initial conditions for the growth of structure in the universe. These fluctuations enjoy a number of properties that greatly simplify the study of structure formation in the later universe: (i) they are nearly scale-invariant, in the sense that the RMS fluctuations in the gravitational potential are independent of scale; (ii) they form a Gaussian random field; (iii) they are adiabatic, that is, the entropy per particle is constant (see §9.1 for further discussion of these concepts).

Finally, inflation solves the **monopole problem**: point topological defects known as monopoles arise naturally in grand unified theories of particle physics, and the predicted density of these objects is far larger than allowed by observational constraints. Inflation solves this problem by diluting the density of monopoles—like non-relativistic matter, this density falls as $a^{-3}(t)$—to an undetectably small value.

Particle physics has no difficulty embracing particle fields that could have caused inflation. Inflation is thought to end when a phase transition converts the inflating matter to ordinary matter and radiation (in this context, "ordinary" includes WIMPs or other non-baryonic dark matter). Any

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

matter or radiation present before inflation was diluted to negligible densities by this time. The newly formed matter and radiation will be in thermal equilibrium, with density parameter $\Omega = 1$ and density fluctuations that are Gaussian, adiabatic, and nearly scale-invariant—precisely the conditions we need to explain many of the properties of the observed universe.

Thus the inflationary hypothesis not only solves the horizon, flatness, and monopole problems but also provides simple, well-defined initial conditions that enable quantitative predictions about the growth of structure in the universe. It should be kept in mind, however, that despite its central role in modern cosmology, the inflationary hypothesis is still unsupported by any evidence from other arenas in theoretical or experimental physics.

1.3.5 The cosmic microwave background

Following inflation, the universe was radiation-dominated, so the scale factor grew as $a(t) \propto t^{1/2}$ and the density fell as $\rho \propto a^{-4}$ (eq. 1.64). All particle species were in thermal equilibrium at a temperature $T(t)$, which declined approximately as a^{-1}. Once the temperature dropped below 100 MeV, 10^{-4} s after the Big Bang, the constituents of this hot plasma consisted of relativistic electrons, positrons, neutrinos, and photons, and non-relativistic protons, neutrons, and perhaps WIMPs.

As the universe continued to expand and cool, the collision time between particles grew faster than the expansion time, so particles began to drift out of thermal equilibrium. In particular, weakly interacting particles such as neutrinos dropped out of thermal equilibrium at $T \simeq 10^{10}\,\mathrm{K} = 0.86\,\mathrm{MeV}$, about 1 s after the Big Bang. At this point the neutron/proton ratio, which had been kept in equilibrium by weak interactions, was frozen at about 0.2. The free neutrons then began to decay, with e-folding time 886 s. However, long before this decay process was completed, nucleosynthesis began: below $10^9\,\mathrm{K}$, at $t \simeq 100\,\mathrm{s}$, $k_\mathrm{B}T$ was much smaller than the deuteron binding energy, so deuterium began to accumulate. Eventually its abundance grew large enough for deuterium to burn to tritium and then helium. Nucleosynthesis was essentially complete 200 s after the Big Bang, leaving most of the nucleons as hydrogen (75% by mass) or ^{4}He (25% by mass), with traces of deuterium, ^{3}He and ^{7}Li. The abundance of deuterium, in particular, depends sensitively on the density of baryons at a given temperature, which is determined by the baryon-to-photon ratio η. Thus measurements of the abundance of deuterium in primordial astrophysical systems such as intergalactic clouds can be used to determine η; measurements of the cosmic microwave background radiation (see below) determine the current photon density, and from these two quantities we can determine the baryon density. This method yields a current density parameter for baryons (Yao et al. 2006)

$$\Omega_{\mathrm{b}0} = (0.042 \pm 0.004)\, h_7^{-2}. \tag{1.69}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

As the universe expanded further, the density in radiation and relativistic matter (photons and neutrinos) continued to decline as a^{-4}, while the density in non-relativistic matter (mostly protons, electrons, and helium nuclei) declined as a^{-3}. Eventually, at redshift and time

$$z_{\gamma\mathrm{m}} \simeq 3100, \quad t_{\gamma\mathrm{m}} \simeq 6 \times 10^4\,\mathrm{yr}, \tag{1.70}$$

the density of matter equaled that of radiation (see Figure 1.7). At this point the matter was still fully ionized. However, as the universe continued to expand, at

$$z_{\mathrm{d}} \simeq 1100, \quad t_{\mathrm{d}} \simeq 4 \times 10^5\,\mathrm{yr}, \tag{1.71}$$

the electrons and protons combined to form neutral atomic hydrogen. This **decoupling** or **recombination epoch** was a milestone in the history of the universe for two reasons: (i) before recombination, the ionized baryonic plasma was locked to the photons by Thomson scattering, while after decoupling, the baryonic matter could move relative to the photons so the assembly of bound baryonic structures such as galaxies could begin; (ii) the universe became transparent.

Recombination occurred rather quickly: the fractional RMS dispersion in the scale factor at which photons suffer their last scattering is less than 10%. Thus we can imagine ourselves to be surrounded by an opaque **last-scattering surface** that hides the Big Bang from us.[12]

At recombination, the photons had a black-body spectrum, and this spectrum was preserved even after the universe became transparent: the photon frequencies all decline as $a^{-1}(t)$ (eq. 1.42) so the spectrum remained black-body with a temperature that declined in the same way.

The relic black-body radiation from the last-scattering surface, the **cosmic microwave background** or **CMB**, was discovered in 1965. It dominates the night sky at wavelengths in the range millimeters to centimeters. The spectrum is accurately black-body, with temperature $T = (2.725 \pm 0.001)\,\mathrm{K}$, and this finding provides compelling evidence that the universe arose from a hot, dense initial state—it is only in such a state that the photons can be thermalized in less than a Hubble time. The energy density of the CMB photons corresponds to a density parameter $5.04 \times 10^{-5} h_7^{-2}$; to compute the total density in radiation we must add to this the energy density contributed by relic neutrinos (this cosmic neutrino background has not yet been detected, but is a firm prediction of Big Bang cosmology). Thus the current energy density in radiation is

$$\Omega_{\gamma 0} = 8.48 \times 10^{-5}\, h_7^{-2}. \tag{1.72}$$

[12] The last-scattering surface is analogous to a stellar photosphere, except we are in a cavity in the middle of optically thick material rather than outside a sphere of optically thick material.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

The CMB is remarkably close to isotropic: apart from a dipole term that arises from the velocity of the solar system with respect to a fundamental observer ($368 \pm 2\,\mathrm{km\,s^{-1}}$), the largest fractional RMS anisotropies are $\lesssim 10^{-4}$. These are believed to arise from the primordial fluctuations introduced by inflation, and the power spectrum of these anisotropies provides an exquisitely sensitive probe of many of the fundamental parameters of the universe.

In particular, assuming a FRW universe currently dominated by non-relativistic matter and vacuum energy, and $\Omega_0 = 1$ as predicted by inflation, the power spectrum of CMB fluctuations strongly constrains the Hubble constant, the density in baryons, and the matter density (Spergel et al. 2007):

$$h_7 = 1.05 \pm 0.05, \quad \Omega_{\mathrm{b}0} = (0.0455 \pm 0.0015)\, h_7^{-2}, \quad \Omega_{\mathrm{m}0} = 0.237 \pm 0.034. \tag{1.73}$$

The vacuum-energy density is then $\Omega_{\Lambda 0} = 1 - \Omega_{\mathrm{m}0} = 0.763 \pm 0.034$. This result implies that the dynamics of the universe is currently dominated by vacuum energy—in other words, the universe appears to have entered a second period of inflation.

With these parameters, the density in non-relativistic matter is much larger than the density in baryons. Thus there must be non-baryonic dark matter that contributes roughly 19% of the critical density. Note also that the present density in vacuum energy exceeds the density in matter; since the former is independent of scale factor and the latter scales as $a^{-3} = (1+z)^3$, equality occurred quite recently (Figure 1.7), at

$$z_{\mathrm{m}\Lambda} = \left(\frac{\Omega_{\Lambda 0}}{\Omega_{\mathrm{m}0}}\right)^{1/3} - 1 = 0.48 \pm 0.09. \tag{1.74}$$

The fluctuation spectrum deduced from the CMB measurements is also approximately scale-invariant, again as predicted by inflation.

The parameters in equation (1.73) are consistent with a wide variety of astronomical measurements, including the following: (i) The Hubble constant is consistent with the best direct estimate of the distance scale, using Cepheid distances, which yields $h_7 = 1.03 \pm 0.11$ (eq. 1.14). Moreover, if measurements of the CMB power spectrum are combined with this measurement of the Hubble constant, then the assumption that $\Omega_0 = 1$ can be eliminated, and the data yield $\Omega_0 = 1.014 \pm 0.017$, consistent with the prediction of inflation that $\Omega = 1$. (ii) The fraction of the total mass composed of baryons is

$$f_{\mathrm{b}} = \frac{\Omega_{\mathrm{b}0}}{\Omega_{\mathrm{m}0}} = 0.17 \pm 0.01, \tag{1.75}$$

in reasonable agreement with the estimate from clusters of galaxies, 0.13 ± 0.02 (eq. 1.26). (ii) The baryon density $\Omega_{\mathrm{b}0}$ is consistent with the result from nucleosynthesis, equation (1.69). (iii) The matter density $\Omega_{\mathrm{m}0}$ is consistent with measurements of the geometry of the universe from supernovae, which

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

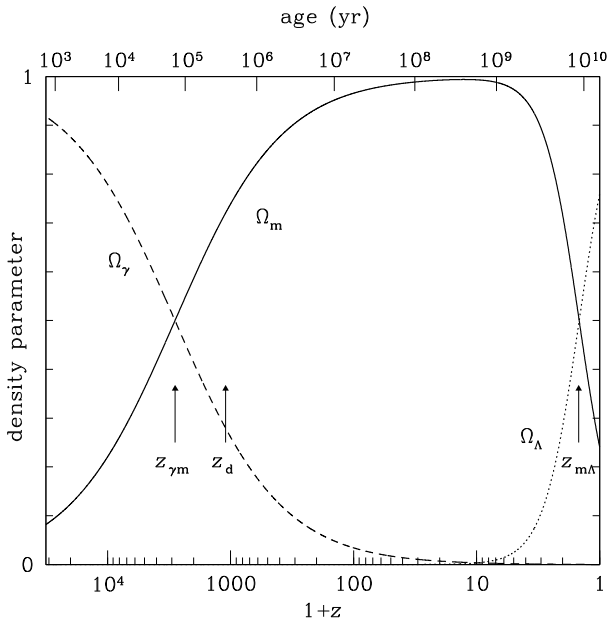

Figure 1.7 The fraction of the critical density provided by radiation (Ω_γ), matter ($\Omega_{\rm m}$), and vacuum energy (Ω_Λ) as a function of redshift z or scale factor $a(t) = (1+z)^{-1}$ (bottom axis) or age (top axis). The values shown are for a flat FRW universe with present matter density $\Omega_{\rm m0} = 0.24$, vacuum-energy density $\Omega_{\Lambda 0} = 0.76$ (eq. 1.73), and radiation density $\Omega_{\gamma 0}$ determined from the temperature of the CMB (eq. 1.72). The redshifts of equal densities of matter and radiation (eq. 1.70), decoupling (eq. 1.71), and equal densities of matter and vacuum energy (eq. 1.74) are marked.

yield $\Omega_{\rm m0} = 0.29 \pm 0.04$ for a flat universe (see Figure 1.6 and Riess et al. 2004). (iv) The matter density $\Omega_{\rm m0}$, together with equation (1.56), predicts that the average mass-to-light ratio in the universe is

$$\Upsilon_R = (220 \pm 80)\Upsilon_\odot, \tag{1.76}$$

consistent with the observed mass-to-light ratio of large clusters of galaxies, $\Upsilon_R = (210 \pm 50)\Upsilon_\odot$. (v) Given $\Omega_{\rm m0}$, $\Omega_{\Lambda 0}$, and $\Omega_{\gamma 0}$ (eq. 1.72), the temporal evolution of the scale factor is entirely determined by equation (1.61) (see Figure 1.7). The age of the universe is thus found to be

$$t_0 = (13.73 \pm 0.16)\,{\rm Gyr}, \tag{1.77}$$

consistent with the ages of globular-cluster stars (§1.1.1) and radioactive dating of the oldest stars (Cayrel et al. 2001).

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

This approximate but encouraging agreement among quite different methods of measuring the same cosmological parameters strongly suggests that we live in a universe with a flat geometry containing three main components: vacuum energy or some other field with a similar equation of state ($\sim 76\%$); non-baryonic dark matter ($\sim 20\%$), and baryons ($\sim 4\%$)—for a detailed inventory see Fukugita, Hogan, & Peebles (1998) and Fukugita & Peebles (2004).

There are unsatisfying aspects to this picture. In particular, there is no independent experimental evidence or strong theoretical justification for three of the central ingredients of this cosmological model: vacuum energy, the particle(s) that comprise the non-baryonic dark matter, and the field that drives inflation. There is also no explanation of why we happen to live at the special epoch when the densities in matter and vacuum energy are similar. Much work remains to be done.

Problems

1.1 [2] In principle, the density of matter in the solar neighborhood can be measured from its effects on planetary orbits. Assume that the solar system is permeated by a uniform medium of density $0.1\,\mathcal{M}_\odot\,\mathrm{pc}^{-3}$ (Table 1.1). Estimate the rate of precession of the perihelion of Neptune (orbital radius $4.5\times10^{12}\,\mathrm{m}$) due to the perturbing force from this medium, and compare your result to the minimum measurable precession, which is $\approx 0.01\,\mathrm{arcsec\,yr}^{-1}$. An answer to within an order of magnitude is sufficient.

1.2 [2] (a) The **luminosity density** $j(\mathbf{r})$ of a stellar system is the luminosity per unit volume at position $\mathbf{r}$. For a transparent spherical galaxy, show that the surface brightness $I(R)$ (Box 2.1) and luminosity density $j(r)$ are related by the formula

$$I(R) = 2\int_R^\infty \mathrm{d}r\,\frac{rj(r)}{\sqrt{r^2-R^2}}. \tag{1.78}$$

(b) What is the surface brightness of a transparent spherical galaxy with luminosity density $j(r) = j_0(1+r^2/b^2)^{-5/2}$ (this is the Plummer model of §2.2.2c)?

(c) Invert equation (1.78) using Abel's formula (eq. B.72) to obtain

$$j(r) = -\frac{1}{\pi}\int_r^\infty \frac{\mathrm{d}R}{\sqrt{R^2-r^2}}\frac{\mathrm{d}I}{\mathrm{d}R}. \tag{1.79}$$

(d) Determine numerically the luminosity density in a spherical galaxy that follows the $R^{1/4}$ surface-brightness law. Plot $\log_{10} j(r)$ versus $\log_{10} r/R_\mathrm{e}$, where R_e is the effective radius.

1.3 [2] The **strip brightness** $S(x)$ is defined so that $S(x)\,\mathrm{d}x$ is the total luminosity in a strip of width $\mathrm{d}x$ that passes a distance x from the projected center of the system.

(a) Show that in a transparent, spherical system

$$S(x) = 2\int_x^\infty \mathrm{d}R\,\frac{RI(R)}{\sqrt{R^2-x^2}}, \tag{1.80}$$

where $I(R)$ is the surface brightness at radius R.

(b) Show that the luminosity density and the total luminosity interior to r are related to the strip brightness by (Plummer 1911)

$$j(x) = -\frac{1}{2\pi x}\frac{\mathrm{d}S}{\mathrm{d}x} \quad ; \quad L(r) = -2\int_0^r \mathrm{d}x\,x\frac{\mathrm{d}S}{\mathrm{d}x}. \tag{1.81}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

1.4 [2] An axisymmetric transparent galaxy has luminosity density that is constant on spheroids $R^2 + z^2/q^2$ having axis ratio q. A distant observer located on the symmetry axis of the galaxy sees an image with circular isophotes and central surface brightness $I_{\rm n}$. A second distant observer, observing the galaxy from a line of sight that is inclined by an angle i to the symmetry axis, sees an image with elliptical isophotes with axis ratio $Q < 1$ and central surface brightness I_0.

(a) What is the relation between I_0, $I_{\rm n}$, and Q? Hint: the answers are different for oblate ($q < 1$) and prolate ($q > 1$) galaxies.

(b) What is the relation between q, Q, and i?

(c) Assuming that galaxies are oriented randomly, what fraction are seen from a line of sight that lies within 10° of the symmetry axis? From within 10° of the equatorial plane?

1.5 [1] (a) Why is the estimated mass-to-light ratio of clusters of galaxies, equation (1.25), proportional to the assumed value of the Hubble constant h_7?

(b) Dark matter was discovered by Zwicky (1933), who compared the mass-to-light ratio of the Coma cluster of galaxies (as measured by the virial theorem, §4.8.3) with the mass-to-light ratios of the luminous parts of spiral galaxies as measured by circular-speed curves, and concluded that there was 400 times as much dark matter as luminous matter in the Coma cluster. However, Zwicky's conclusion was based on a Hubble constant $H_0 = 558\,{\rm km\,s^{-1}\,Mpc^{-1}}$. How would his conclusion about the ratio of dark to luminous matter have been affected had he used the correct value of the Hubble constant, which is smaller by a factor of eight?

1.6 [1] (a) Associated with the vacuum-energy density ρ_Λ is the characteristic timescale $(G\rho_\Lambda)^{-1/2}$. What is its value for the cosmological parameters in equation (1.73), and what is its physical significance?

(b) Einstein's original formulation of general relativity included a contribution from vacuum energy, which he called the **cosmological constant** and parametrized by

$$\Lambda \equiv \frac{8\pi G\rho_\Lambda}{c^2}. \tag{1.82}$$

$\Lambda^{-1/2}$ is a characteristic length. What is its value for the cosmological parameters in equation (1.73)?

1.7 [2] Prove that the volume of a closed FRW universe is $2\pi^2 a^3(t) x_{\rm u}^3$.

1.8 [1] Einstein proposed a static FRW universe, that is, one in which the scale factor $a(t) = a_0 = \textit{constant}$ and the Hubble constant $H_0 = 0$.

(a) If the radiation density is negligible, prove that the matter density in this universe equals twice the vacuum-energy density.

(b) Suppose that the scale factor is perturbed from a_0 by a small amount, $a(t) = a_0 + \epsilon a_1(t)$ with $\epsilon \ll 1$. Show that $a_1(t)$ grows exponentially, so the static universe is unstable, and derive the growth rate.

1.9 [1] Assuming a flat FRW universe with parameters given by equation (1.73) at the present time, what is the value of the Hubble parameter in the distant future? If this is less than its present value H_0, why is the universe said to be accelerating?

1.10 [1] Suppose that some of the dark matter is composed of iron asteroids of density $\rho = 8\,{\rm g\,cm^{-3}}$ and radius r that are uniformly distributed throughout intergalactic space. If the density in this form is $\Omega_{\rm a} = 0.01$, find an approximate lower limit on r from the condition that the universe is not opaque, i.e., that we can see distant quasars. Your answer need be correct only to within a factor of two or so.

1.11 [2] Reproduce Figure 1.6.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

1.12 [2] Write a function to do the following task: given values of the Hubble constant H_0, the density parameters $\Omega_{\rm m0}$, $\Omega_{\gamma 0}$, $\Omega_{\Lambda 0}$, and a specified redshift z, find the age of the universe at that redshift. For a flat universe with parameters given by equations (1.72) and (1.73), what is the age at redshifts $z = 1000$, $z = 1$, and $z = 0$?

1.13 [2] The redshift at which the densities in matter and radiation are equal is $z_{\gamma\rm m}$. For a flat FRW universe containing matter, radiation, and vacuum energy, with parameters not too far from our own, prove that (a)

$$1 + z_{\gamma\rm m} = \frac{\Omega_{\rm m0}}{\Omega_{\gamma 0}} = 1.18 \times 10^4\, \Omega_{\rm m0} h_7^2 \tag{1.83}$$

(hint: use eq. 1.72); (b) the age of the universe at $z_{\gamma\rm m}$ is

$$t_{\gamma\rm m} = \frac{2(2-\sqrt{2})}{3H_0} \frac{\Omega_{\gamma 0}^{3/2}}{\Omega_{\rm m0}^2}; \tag{1.84}$$

(c) the comoving horizon at $z_{\gamma\rm m}$ (eq. 1.67) is

$$x_{\gamma\rm m} = 2(\sqrt{2}-1)\frac{c}{H_0}\frac{\Omega_{\gamma 0}^{1/2}}{\Omega_{\rm m0}} = \frac{32.7\,{\rm Mpc}}{\Omega_{\rm m0} h_7^2}. \tag{1.85}$$

Evaluate $z_{\gamma\rm m}$, $t_{\gamma\rm m}$, and $x_{\gamma\rm m}$ for the parameters of equation (1.73).

1.14 [1] The universe was opaque before decoupling at $z > z_{\rm d} \simeq 1100$ (eq. 1.71) because the ionized baryonic plasma had a high optical depth to Thomson scattering. For $z < z_{\rm d}$ the electrons and protons recombined to form neutral atoms and the universe became transparent. Somewhere between $z \sim 20$ and $z \sim 6$ high-energy photons from newly formed quasars reionized most of the intergalactic medium. Why is the universe not opaque for $z \lesssim 6$?

For general queries, contact webmaster@press.princeton.edu