2 The Large-Scale Structure of the Universe

Our current picture of how matter and radiation are distributed in the Universe on a large scale is derived from a wide variety of different types of observation. In this chapter, we concentrate upon the large-scale distribution of matter and radiation in the Universe and discuss galaxies and clusters of galaxies in Chaps. 3 and 4 respectively. The observations described in this chapter provide much of the essential underpinning of modern cosmological research.

2.1 The Spectrum and Isotropy of the Cosmic Microwave Background Radiation

On the very largest scales, the best evidence for the overall isotropy of the Universe is provided by observations of the *Cosmic Microwave Background Radiation*. This intense diffuse background radiation in the centimetre, millimetre and submillimetre wavebands was discovered in 1965 by Penzias and Wilson whilst commissioning a sensitive maser receiver system for centimetre wavelengths at the Bell Telephone Laboratories (Penzias and Wilson, 1965). It was soon established that this radiation is remarkably uniform over the sky and that, in the wavelength range $1\,\mathrm{m} > \lambda > 1\,\mathrm{cm}$, the intensity spectrum had the form $I_\nu \propto \nu^2$, corresponding to the Rayleigh–Jeans region of a black-body spectrum at a radiation temperature of about 2.7 K.

The maximum intensity of a black-body spectrum at a radiation temperature of 2.7 K occurs at a wavelength of about 1 mm at which atmospheric emission makes precise absolute measurements of the background spectrum from the surface of the Earth very difficult indeed. During the 1970s and 1980s several high-altitude balloon experiments carrying millimetre and submillimetre spectrometers were flown and evidence found for the expected turn-over in the Wien region of the spectrum, but there were discrepancies between the experiments (Weiss, 1980). The only satisfactory approach for determining the detailed spectrum and isotropy of the Cosmic Background Radiation over the whole sky was to place the receiver system in a satellite above the Earth's atmosphere and this was achieved by the Cosmic Background Explorer (COBE) of NASA which was launched in November 1989. This mission was dedicated to studies of the background radiation, not only in the millimetre and submillimetre wavebands, but also throughout the infrared waveband from 2 to 1000 μm.

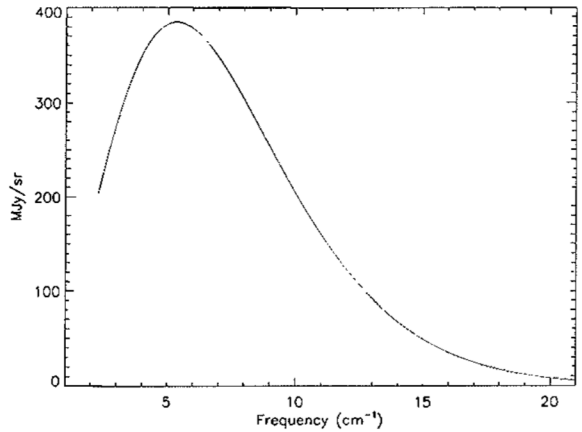


Fig. 2.1. The final spectrum of the Cosmic Microwave Background Radiation as measured by the COBE satellite (Fixsen et al., 1996). The units of the abscissa are inverse centimetres, so that 10 units corresponds to 1 mm and 5 to 2 mm. Uncertainties are a small fraction of the line thickness. Within the quoted errors, the spectrum is precisely that of a perfect black-body at radiation temperature $T = 2.728 \pm 0.002$ K

2.1.1 The Spectrum of the Cosmic Microwave Background Radiation

The Far Infrared Absolute Spectrophotometer (FIRAS) of COBE measured the *spectrum* of the Cosmic Microwave Background Radiation in the wavelength range 0.5 to 2.5 mm with very high precision during the first year of the mission. The FIRAS detectors and a reference black-body source were cooled to liquid helium temperatures and there was sufficient liquid cryogen for only one year of observation. Particular care was taken over the thermometry involved in making such absolute temperature measurements. The final spectrum shown in Fig. 2.1 is that of a perfect black-body with a radiation temperature $T = 2.728 \pm 0.002$ K (Fixsen et al., 1996). More quantitatively, the deviations from a perfect black-body spectrum in the wavelength interval $2.5 > \lambda > 0.5$ mm amounted to less than 0.03% of the maximum intensity. This is the most beautiful example I know of a naturally occurring black-body radiation spectrum.

There are two convenient ways of describing the degree to which the observed spectrum differs from that of a perfect black-body spectrum, both of them pioneered by Zeldovich and Sunyaev in the late 1960s (for details of their work, see their review of 1980 (Sunyaev and Zeldovich, 1980a)). They showed that the injection of large amounts of thermal energy in the form of hot gas into the intergalactic medium can produce various types of distortion of the black-body radiation spectrum because of Compton scattering of the background photons by hot electrons. We will not go into

the physics of these processes at this point, except to note the forms of distortion and the limits which can be set to certain characteristic parameters.

If there were early injection of thermal energy prior to the epoch when the primordial plasma recombined at a redshift of about 1000, and if the number of photons was conserved, the spectrum would relax to an equilibrium Bose–Einstein spectrum with a finite dimensionless chemical potential μ,

$$I_\nu = \frac{2h\nu^3}{c^2}\left[\exp\left(\frac{h\nu}{kT_\mathrm{r}} + \mu\right) - 1\right]^{-1} . \tag{2.1}$$

The simplest way of understanding this result is to note that the Bose–Einstein distribution is the equilibrium distribution for photons when there is a mismatch between the total energy and the number of photons over which this energy is to be distributed. In the case of a black-body spectrum, both the energy density and number density of photons are determined solely by the temperature T_r. In contrast, the Bose–Einstein distribution is determined by two parameters, the temperature T_r and the dimensionless chemical potential μ.

In the case of Compton scattering by hot electrons at late epochs, the energies of the photons are redistributed about their initial values and, to second order, there is an increase in their mean energies so that the spectrum is shifted to slightly greater frequencies.[1] In 1969, Zeldovich and Sunyaev showed that the distortion of the black-body spectrum takes the form

$$\frac{\Delta I_\nu}{I_\nu} = y\frac{x\mathrm{e}^x}{(\mathrm{e}^x - 1)}\left[x\left(\frac{\mathrm{e}^x + 1}{\mathrm{e}^x - 1}\right) - 4\right] , \tag{2.2}$$

where y is the Compton scattering optical depth $y = \int (kT_\mathrm{e}/m_\mathrm{e}c^2)\,\sigma_\mathrm{T} N_\mathrm{e}\,\mathrm{d}l$, $x = h\nu/kT_\mathrm{r}$ and σ_T is the Thomson scattering cross-section (Zeldovich and Sunyaev, 1969). In the limit of small distortions, $y \ll 1$, the intensity in the Rayleigh–Jeans region decreases as $\Delta I_\nu/I_\nu = -2y$ and the total energy under the spectrum increases as $\varepsilon = \varepsilon_0 \mathrm{e}^{4y}$.

Limits to the parameters y and μ have been derived from the very precise spectral measurements made by the FIRAS instrument. The results quoted by Page are as follows (Page, 1997):

$$|y| \leq 1.5 \times 10^{-5} , \quad |\mu| \leq 10^{-4} . \tag{2.3}$$

These are very strong limits indeed and will prove to be of astrophysical importance in the study of the physics of the intergalactic gas, as well as constraining the amount of star and metal formation which could have taken place in young galaxies.

2.1.2 The Isotropy of the Cosmic Microwave Background Radiation

Equally remarkable were the COBE observations of the *isotropy* of the distribution of the Cosmic Microwave Background Radiation over the sky. The prime instruments

[1] I have given a derivation of this result in my book *High Energy Astrophysics, Vol. 1* (Longair, 1997b).

for these studies were the Differential Microwave Radiometers which operated at frequencies of 31.5, 53 and 90 GHz, thus sampling the Rayleigh–Jeans region of the background spectrum. The angular resolution of the radiometers was 7°. The choice of observing frequency was crucial in these observations. At higher frequencies, the millimetre and submillimetre emission of diffuse Galactic dust at high galactic latitudes, often referred to as 'cirrus', confuses the picture, whilst at lower frequencies the radio synchrotron radiation of ultrarelativistic electrons gyrating in the Galactic magnetic field becomes important. The final results of the four-year mission are shown in Fig. 2.2 at increasing levels of sensitivity (Bennett et al., 1996).

Figure 2.2a illustrates the stunning result that the Cosmic Microwave Background Radiation is extraordinarily uniform over the whole sky with radiation temperature 2.728 K.

At a sensitivity level of about one part in 1000 of the total intensity, a large-scale anisotropy of dipolar form was observed over the whole sky (Fig. 2.2b). The plane of our Galaxy can also be observed as a faint band of emission along the Galactic equator. The global dipole anisotropy is naturally attributed to aberration effects associated with the Earth's motion through an isotropic radiation field. Excluding regions close to the Galactic plane, the temperature distribution was found to have precisely the expected dipole distribution, $T = T_0[1 + (v/c)\cos\theta]$, where θ is the angle with respect to the direction of maximum intensity and v is the Earth's velocity through the isotropic background radiation. The amplitude of the cosmic microwave dipole was 3.353 ± 0.024 mK with the maximum intensity in the direction towards galactic coordinates $l = 264.25° \pm 0.33°$; $b = 48.22° \pm 0.13°$ (Bennett et al., 1996). It was inferred that the Earth is moving at about 350 km s^{-1} with respect to the frame of reference in which the radiation would be 100% isotropic. It is significant that, although not designed to undertake this task, exactly the same form of large-scale dipole anisotropy was observed by the FIRAS instrument.

The measurement of the velocity of the Sun relative to the Cosmic Microwave Background Radiation is an important result for understanding the large-scale distribution of mass in the Universe. Once allowance is made for the motion of the Sun about the centre of our Galaxy, an estimate of the peculiar velocity of our Galaxy and the local group of galaxies relative to the frame of reference in which the background radiation would be perfectly isotropic can be found. This motion can be attributed to perturbations in the distribution of mass on very large scales in the relatively nearby Universe (Kolatt et al., 1995).

On angular scales of 7° and greater, Bennett and his colleagues achieved sensitivity levels better than one part in 100,000 of the total intensity from analyses of the complete microwave data set obtained over the four years of the COBE mission (Fig. 2.2c). At this sensitivity level, the radiation from the plane of the Galaxy is intense, but is confined to a broad strip lying along the Galactic equator. Away from this region, the sky appears quite smooth on a large scale, but there are significant fluctuations in intensity from beamwidth to beamwidth over the sky. These fluctuations are present at the level of only about 1 part in 100,000 of the total intensity and, when averaged over the clear region of sky at $|b| > 20°$ amount to a root-mean-square amplitude of 35 ± 2 μK on an angular scale of 7°, or to 29 ± 1 μK when

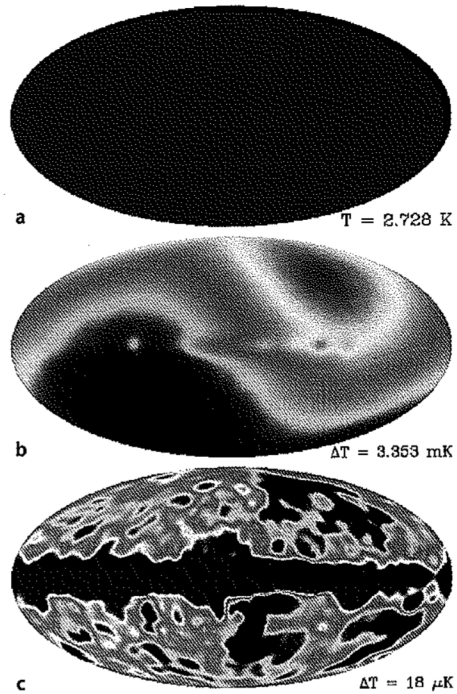


Fig. 2.2a–c. Maps of the whole sky in galactic coordinates as observed at a wavelength of 5.7 mm (53 GHz) by the COBE satellite at different sensitivity levels. **a** The distribution of total intensity over the sky. **b** Once the uniform component was removed, the dipole component associated with the motion of the Earth through the background radiation was observed, as well as a weak signal from the Galactic plane. **c** Once the dipole component was removed, radiation from the plane of the Galaxy was seen as a bright band across the centre of the picture. The fluctuations seen at high galactic latitudes were a combination of noise from the telescope and the instruments and a genuine cosmological signal. The rms value of the fluctuations at each point away from the Galactic equator amounted to 36 μK. When averaged statistically over the whole sky at high latitudes, an excess sky noise signal of cosmological origin of 30 ± 5 μK was detected (Bennett et al., 1996)

smoothed to 10° angular scale. These values were found to be frequency independent for the three COBE frequency channels at 31.5, 53 and 90 GHz. The detection of these fluctuations is a crucial result for understanding the origin of the large-scale structure of the Universe. The COBE observations allow information to be obtained about the angular spectrum of the intensity fluctuations on all scales $\theta \geq 7°$. In Chap. 15, we will deal with the important cosmological information which can be

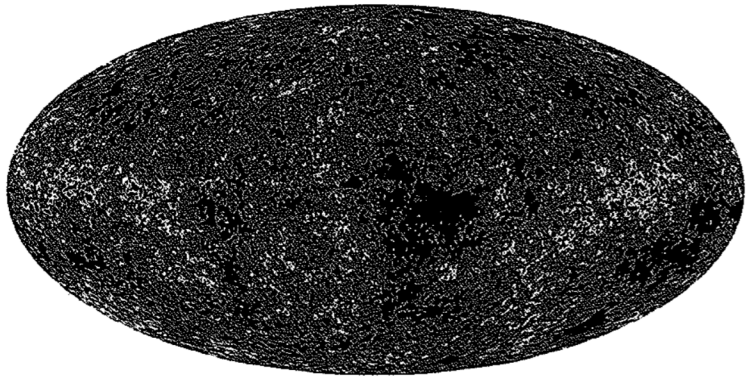


Fig. 2.3. A map of the whole sky in galactic coordinates as observed by the WMAP satellite at millimetre wavelengths (Bennett et al., 2003). The angular resolution of the map is about 20 times higher than that of Fig. 2.2c. The emissions due to Galactic dust and synchrotron radiation have been subtracted from this map

derived from observations of temperature fluctuations in the Cosmic Microwave Background Radiation on smaller angular scales.

It is interesting to compare the COBE map (Fig. 2.2c) with the more recent WMAP observations of 2003 made with about 20 times higher angular resolution (Bennett et al., 2003) (Fig. 2.3). It can be seen that the same large scale features are present on both maps. In particular, regions of strong positive and negative fluctuations agree rather well. We will have a lot more to say about Fig. 2.3 in due course.

The COBE observations are crucial for cosmology. From the point of view of the structure of the Universe on the very largest angular scales, they show that the Cosmic Microwave Background Radiation is isotropic to better than one part in 100,000. Whatever its origin, this observation in itself shows that the Universe must be extraordinarily isotropic on the large scale. As we will show, it is wholly convincing that this radiation is the cooled remnant of the very hot early phases of the Big Bang.

How is the distribution of radiation related to the distribution of matter in the Universe? We will take up this topic in much more detail in Chap. 9, but it is useful to outline here how they are related. In the standard Big Bang picture, when the Universe was squashed to only about one thousandth of its present size, the temperature of the Cosmic Microwave Background Radiation was about one thousand times greater than it is now. The temperature of the background radiation varies with redshift z as $T_r = 2.728(1 + z)\,\mathrm{K}$ and so, at a redshift $z = 1500$, the temperature of the radiation field was about 4000 K. At this temperature, there were sufficient Lyman continuum photons in the Wien region of the background spectrum to photoionise all the neutral hydrogen in the Universe. At this early epoch, known as the *epoch*

of recombination, galaxies had not formed and all the ordinary baryonic matter, which was eventually to become the visible matter of galaxies as we know them, was still in the form of remarkably smooth, partially ionised pre-galactic gas. At earlier epochs, the pre-galactic gas was fully ionised and was very strongly coupled to the background radiation by Thomson scattering.

Therefore, when we look back to these epochs, it is as if we were looking at the surface of a star surrounding us in all directions, but the temperature of the radiation we observe has been cooled by a cosmological redshift factor of 1500, so that what we observe is redshifted into the millimetre waveband. This analogy makes it clear that, because of Thomson scattering of the background radiation, we can only observe the very surface layers of our 'star'. We cannot obtain any direct information about what was going on at earlier epochs. This 'surface' at which the Universe became opaque to radiation is known as the *last scattering surface* and the fluctuations observed by COBE are interpreted as the very low intensity ripples present on that surface on angular scales of 7° and greater. These ripples grow under gravity and will eventually define some of the very largest scale structures in the local Universe.

In the interpretation of the COBE observations described in the last paragraph, it was assumed that the intergalactic gas was transparent to radiation from the epoch of recombination onwards and was not reionised and heated at some later epoch. If that were to occur, the perturbations would be further damped by Thomson scattering and this has now been detected in the WMAP observations. However, the damping is not so great that features in the power spectrum of the fluctuations are wiped out.

One important aspect of these studies is that the energy density of the Cosmic Microwave Background Radiation amounts to $\varepsilon_{\mathrm{rad}} = aT_r^4 = 4.2 \times 10^{-14}\,\mathrm{J\,m^{-3}} = 2.64 \times 10^5\,\mathrm{eV\,m^{-3}}$. This energy density of radiation pervades the whole Universe at the present epoch and provides by far the greatest contribution to the average energy density of the universal background radiation.

2.2 The Large-Scale Distribution of Galaxies

The visible Universe of galaxies is highly inhomogeneous, consisting of structures from the scale of isolated galaxies, through groups and clusters of galaxies to superclusters and giant voids in the distribution of galaxies. As we progress to larger and larger scales, the distribution of galaxies becomes smoother, but still contains significant non-random features. For many purposes, it is convenient to think of the galaxies as *the building blocks of the Universe* which define its large-scale structure.

An excellent representation of the large-scale distribution of galaxies on the sky is shown in Fig. 2.4. This remarkable picture was created from scans of 185 contiguous UK Schmidt plates, each of which covers an area of $6° \times 6°$ on the sky, the scanning being carried out by the Cambridge APM high-speed measuring machine (Maddox et al., 1990). The image is centred on the South Galactic pole and so the effect of Galactic obscuration by dust on the distribution of galaxies is negligible. Each plate was carefully calibrated and stars distinguished from galaxies by their different image profiles. Figure 2.4 contains over two million galaxies with

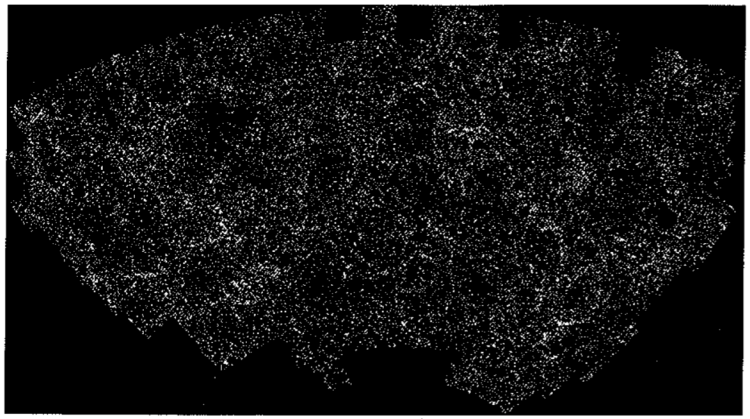


Fig. 2.4. The distribution of galaxies with $17 \leq b_j \leq 20.5$ shown in an equal area projection centred on the South Galactic pole. This image was reconstructed from machine scans of 185 UK Schmidt plates by the Cambridge APM measuring machine. There are over two million galaxies in this image. The small empty patches in the map are regions that have been excluded about bright stars, nearby dwarf galaxies, globular clusters and step wedges (Maddox et al., 1990)

apparent magnitudes in the range $17 \leq b_j \leq 20.5$ and so represents the distribution of galaxies on the sky on the grandest scale.

It is apparent that, although one bit of Fig. 2.4 does not look too different from another on a large enough scale, the distribution of galaxies is far from uniform on a small scale. There appear to be clumps of galaxies, stringy structures and holes but, of course, the eye is expert at finding such structures in random data. Despite this concern, much of the obvious clumping, the clusters, holes and stringy structures are real features of our Universe. To demonstrate the reality of these features, the three-dimensional distribution of galaxies needs to be determined and so distances have to be measured for very large samples of galaxies. Although this is a really huge task, this has now been achieved thanks to the efforts of many astronomers.

2.2.1 Two-Point Correlation Functions

We need statistical methods appropriate for describing the clustering properties of galaxies on a wide range of scales and the simplest approach is to use *two-point correlation functions*. In the cosmological case, these can be described either in terms of the distribution of galaxies on the sky, or in terms of spatial two-point correlation functions in three dimensions. On the sky, we define the *angular two-point correlation function*, $w(\theta)$, by

$$N(\theta)\,\mathrm{d}\Omega = n_\mathrm{g}[1 + w(\theta)]\,\mathrm{d}\Omega\,, \tag{2.4}$$

where $w(\theta)$ describes the excess probability of finding a galaxy at an angular distance θ from any given galaxy. The term $\mathrm{d}\Omega$ is the element of solid angle and n_g is a suitable average surface density of galaxies. The term $w(\theta)$ contains information about the clustering properties of galaxies to a given limiting apparent magnitude and can be measured with some precision from large statistical surveys of galaxies such as the Cambridge APM surveys which contain over two million galaxies (Fig. 2.4). Notice the important point that two-point correlation functions take circularly symmetric averages about each galaxy and so throw away all information about the 'stringiness' of the distribution of galaxies. Some of this more detailed structural information can be recovered using three- and four-point correlation functions, but let us begin with the two-point correlation functions.

The homogeneity of the distribution of galaxies with increasing distance can be studied by measuring the angular two-point correlation function as a function of increasing apparent magnitude. If the galaxies are sampled from a homogeneous, but clustered, distribution, the angular two-point correlation function scales with increasing limiting distance D in local Euclidean space as

$$w(\theta, D) = \frac{D_0}{D} w_0 \left(\theta \frac{D}{D_0}\right)\,, \tag{2.5}$$

where the function $w_0(\theta)$ has been determined to distance D_0. The factor $\theta(D/D_0)$ in the argument of w_0 takes account of the fact that a fixed scale subtends a smaller angle at a greater distance D and the factor (D_0/D) in front of w_0 takes account of the fact that there are more background galaxies ($\propto D^3$), but that the surface density of galaxies about any galaxy to a fixed physical scale increases only as D^2. If the galaxies extend to distances D such that redshift effects need to be taken into account, it is necessary to integrate over the luminosity function of the galaxies counted and to use a Friedman world model to determine the spatial and surface number densities of galaxies (Groth and Peebles, 1977; Scranton et al., 2002).

Such scaling analyses were carried out by Groth and Peebles who showed that the two-point correlation functions determined from a bright sample of Zwicky galaxies, from the Lick counts of galaxies and from a deep sky survey plate in an area known as the Jagellonian field scaled exactly as expected if the distribution of the galaxies displayed the same degree of spatial correlation throughout the local Universe out to $z \sim 0.1$ (Groth and Peebles, 1977, 1986). A similar result was found by comparing the two-point correlation functions found at increasing apparent magnitude limits in the machine-scanned surveys carried out by the APM group at Cambridge (Maddox et al., 1990). Figure 2.5a shows the angular two-point correlation functions $w(\theta)$ measured at increasing apparent magnitude limits in the magnitude range $17.5 < m < 20.5$. In Fig. 2.5b, these functions are scaled to the angular correlation function found from the Lick survey.

More recently, the same type of analysis has been carried out for a large sample of galaxies from the Sloan Digital Sky Survey (SDSS) which extends to apparent magnitude $r^* = 23$ by Connolly, Scranton and their colleagues. According to their estimates, the mean redshift in the magnitude interval $21 \leq r^* \leq 22$ is 0.43. Using the same scaling procedures with a proper Friedman cosmological model, they

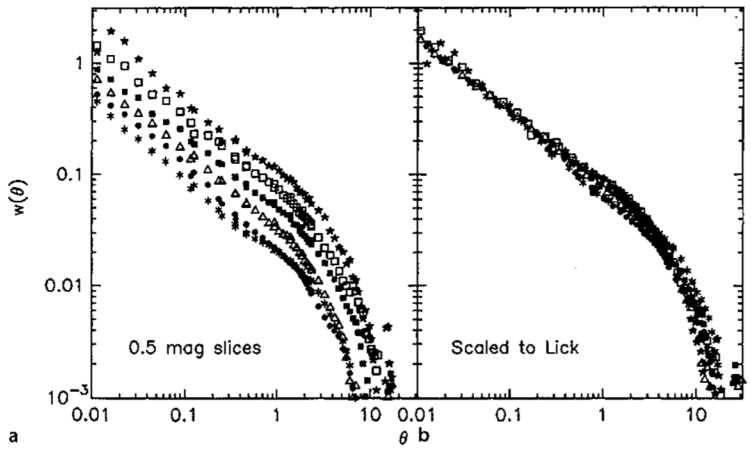


Fig. 2.5a,b. The two-point correlation function for galaxies over a wide range of angular scales. **a** The scaling test for the homogeneity of the distribution of galaxies can be performed using the correlation functions for galaxies derived from the APM surveys at increasing limiting apparent magnitudes in the range $17.5 < m < 20.5$. The correlation functions are displayed in intervals of 0.5 magnitudes. **b** The two-point correlation functions scaled to the correlation function derived from the Lick counts of galaxies (Maddox et al., 1990)

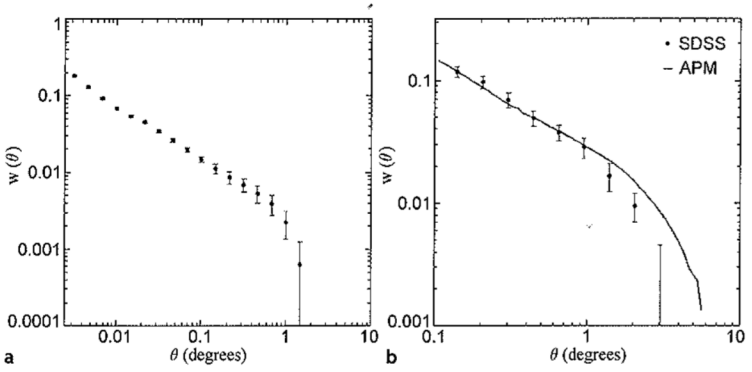


Fig. 2.6a,b. The two-point correlation function for galaxies determined from the Sloan Digital Sky Survey (SDSS) (Connolly et al., 2002; Scranton et al., 2002). **a** The angular two-point correlation function determined in a preliminary analysis of 2% of the galaxy data contained in the Sloan Digital Sky Survey. **b** Comparison of the scaled angular two-point correlation functions found by Maddox and his colleagues from the APM galaxy survey (*solid line*) with that found from the SDSS analysis

find excellent agreement with the angular two-point correlation function determined by Maddox and his colleagues (Fig. 2.6) (Connolly et al., 2002; Scranton et al., 2002). It is noteworthy that a better match of the correlation functions is found for a concordance world model with $\Omega_0 = 0.3$ and $\Omega_\Lambda = 0.7$ than for the critical model with $\Omega_0 = 1$ and $\Omega_\Lambda = 0$.

These are important results for the construction of cosmological models. As expressed by Peebles (Peebles, 1993):

> ... the correlation function analyses have yielded a new and positive test of the assumption that the galaxy space distribution is a stationary (statistically homogeneous) random process.

Figures 2.5b and 2.6b illustrate the important point that the correlation function for galaxies is smooth, meaning that clustering is found on all angular scales with no prominent features on the scales of clusters or superclusters of galaxies. Of course, it needs to be remembered that the two-point correlation function is circularly symmetric about each galaxy and so wipes out a lot of detailed information. In Figs. 2.5b and 2.6b, $w(\theta)$ can be characterised by a power law of the form

$$w(\theta) \propto \theta^{-(0.7-0.8)} \tag{2.6}$$

with a cut-off on large angular scales.

It is more meaningful physically to work in terms of the *spatial two-point correlation function* $\xi(r)$ which describes the clustering properties of galaxies in three dimensions about any galaxy:

$$N(r)\,\mathrm{d}V = N_0[1 + \xi(r)]\,\mathrm{d}V\,, \tag{2.7}$$

where $N(r)\,\mathrm{d}V$ is the number of galaxies in the volume element $\mathrm{d}V$ at distance r from any galaxy and N_0 is a suitable average space density. $\xi(r)$ describes the excess number of galaxies at distance r from any given galaxy.

In order to derive $\xi(r)$ directly from observation, we need to know the distribution of galaxies in space. If, however, we make a number of reasonable assumptions, we can derive a simple formula which relates $w(\theta)$ to $\xi(r)$. Suppose a cluster of galaxies has radial number density distribution $n(r) = n_0[1 + \xi(r)]$. Then, it is a simple calculation to show that the projected distribution is given by the integral

$$N(a) = 2\int_a^{a_{\max}} \frac{n(r)\,r}{(r^2 - a^2)^{1/2}}\,\mathrm{d}r = 2\int_a^{a_{\max}} \frac{n_0[1 + \xi(r)]\,r}{(r^2 - a^2)^{1/2}}\,\mathrm{d}r\,, \tag{2.8}$$

where a is the projected radial distance from the centre of the cluster and $a_{\max}$ is the tidal radius of the cluster. The first term in the second integral for $n(a)$ is just a constant. If we adopt a power law dependence for $\xi(r)$, $\xi(r) \propto r^{-\gamma}$, the second term in the integral can be written in dimensionless form using the substitution $r = ax$ as follows:

$$N(a) = 2n_0 a^{-(\gamma-1)} \int_1^{a_{\max}/a} \frac{x^{-(\gamma-1)}}{(x^2 - 1)^{1/2}}\,\mathrm{d}x\,. \tag{2.9}$$

Allowing the upper limit of the integral to go to infinity, we see that the integral is a constant and that $N(a)$ and $\xi(r)$ are related by

$$N(a) \propto a^{-(\gamma-1)} . \tag{2.10}$$

Consequently, in the region in which the angular two-point correlation function can be described by the power law relation (2.6), the function $\xi(r)$ can be well represented by a power law of the form

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma} , \tag{2.11}$$

where $\gamma = 1.7 - 1.8$. More detailed analyses of the relation between $\xi(r)$ and $w(\theta)$ show that the correlation function (2.11) is a good match to the data on physical scales from about $100h^{-1}$ kpc to $10h^{-1}$ Mpc in which the scale $r_0 = 5h^{-1}$ Mpc and the exponent $\gamma = 1.7 - 1.8$.[2] On scales greater than about $10h^{-1}$ Mpc the two-point correlation function decreases more rapidly than the power law (2.11). Thus, on large enough scales, the amplitude of the clustering decreases dramatically and the Universe becomes isotropic on the very largest physical scales.

We also obtain the important result that on physical scales $r \gg 5h^{-1}$ Mpc, the mean amplitude of the density perturbations is less than one and consequently density perturbations on larger scales are on average still in the linear regime $\delta\varrho/\varrho \ll 1$ at the present epoch.

2.2.2 Walls and Voids in the Distribution of Galaxies on Large Scales

The analysis of Sect. 2.2.1 provides the simplest description of the distribution of galaxies on large scales, but it cannot describe the walls and voids in the distribution of galaxies seen in Fig. 2.4. The nature of these structures has been well defined by a number of large-scale redshift surveys for galaxies.

One of the earliest complete samples of nearby galaxies is presented in Fig. 2.7 which shows the local three-dimensional distribution of galaxies derived from the Harvard–Smithsonian Astrophysical Observatory survey of over 14,000 bright galaxies (Geller and Huchra, 1989). Our own Galaxy is located at the centre of the diagram and, if the galaxies were uniformly distributed in the local Universe, the points would be uniformly distributed over the diagram, which is certainly very far from the case. There are gross inhomogeneities and irregularities in the local Universe including large 'holes' or 'voids' in which the local number density of galaxies is significantly lower than the mean, and long 'filaments' or 'walls' of galaxies, including the feature known as the 'Great Wall', which extends from right ascensions 9^h to 17^h

[2] The use of $h = H_0/(100\,\mathrm{km\,s^{-1}\,Mpc^{-1}})$ is a convenient device for adjusting the dimensions and luminosities of extragalactic objects to the reader's preferred value of Hubble's constant. If a value of $H_0 = 100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ is preferred, $h = 1$; if the value $H_0 = 50\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ is adopted, $h = 0.5$ and so on. It is now known that the value of h is $h = 0.72 \pm 0.07$.

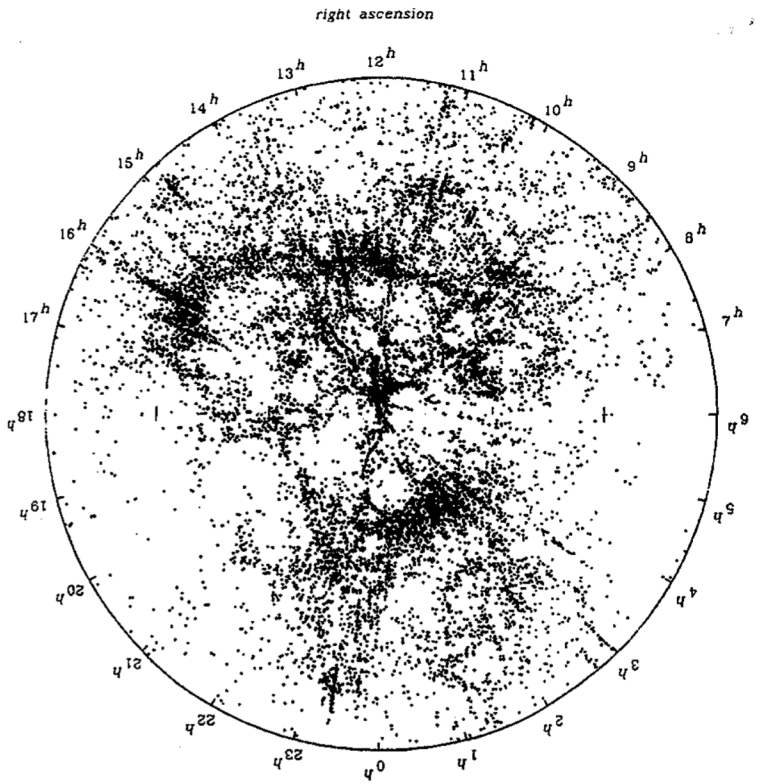
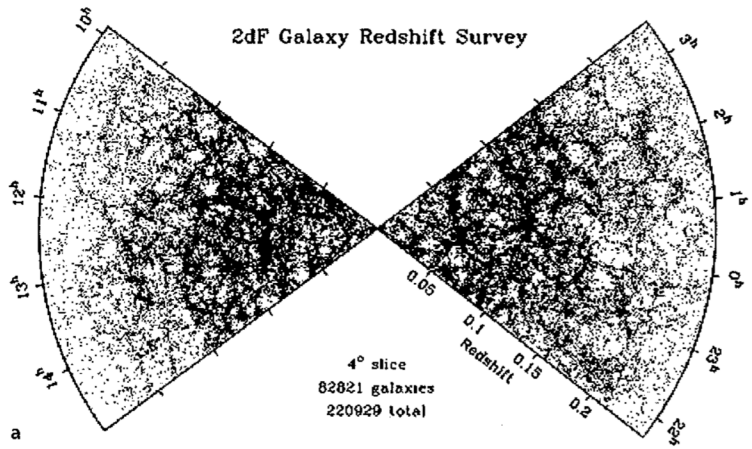


Fig. 2.7. The distribution of galaxies in the nearby Universe as derived from the Harvard–Smithsonian Center for Astrophysics survey of galaxies. The map contains over 14,000 galaxies which form a complete statistical sample around the sky between declinations $\delta = 8.5°$ and $44.5°$. All the galaxies have recession velocities less than $15{,}000\,\mathrm{km\,s^{-1}}$. Our Galaxy is located at the centre of the map and the radius of the bounding circle is $150h^{-1}$ Mpc. The galaxies within this slice have been projected onto a plane to show the large-scale features in the distribution of galaxies. Rich clusters of galaxies which are gravitationally bound systems with internal velocity dispersions of about $10^3\,\mathrm{km\,s^{-1}}$ appear as 'fingers' pointing radially towards our Galaxy at the centre of the diagram. The distribution of galaxies is highly irregular with huge holes, filaments and clusters of galaxies throughout the local Universe (Geller and Huchra, 1989)

about half-way to the limit of the survey. There are a number of 'streaks' or 'fingers' pointing towards our own Galaxy and these correspond to clusters of galaxies, the lengths of the 'streaks' corresponding to the components of the velocity dispersion of the galaxies in the clusters along the line of sight.



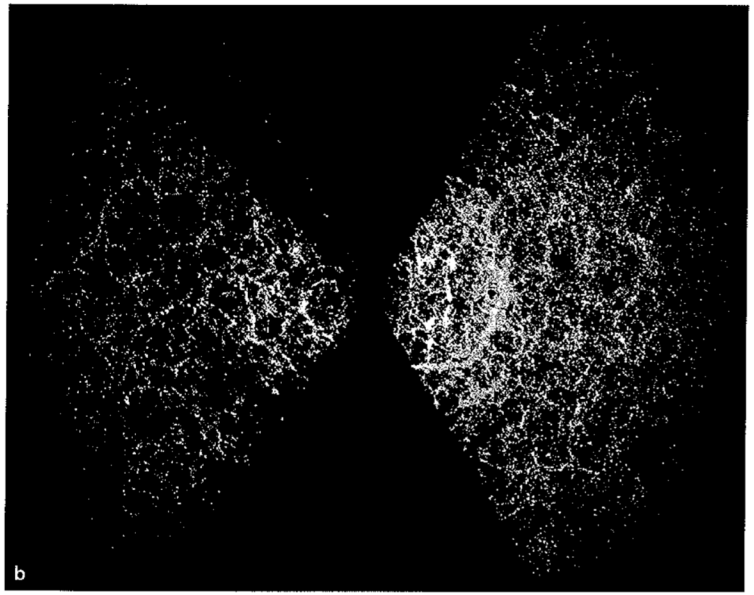


Fig. 2.8a,b. The spatial distribution of galaxies on a large-scale. In both diagrams, the distribution extends to a redshift $z \approx 0.25$. **a** A slice through the Anglo-Australian Telescope 2dF Galaxy Survey (Colless et al., 2001) showing the pronounced 'cellular' structure of the distribution of galaxies on the large scale (image courtesy of the 2dFGRS Team). **b** The distribution of galaxies in the Sloan Digital Sky Survey, showing the same 'cellular' structure observed in the AAT 2dF survey (Stoughton et al., 2002)

These surveys have now been extended to much greater distances and the same overall picture emerges. Among the first of these surveys, the Las Campanas Redshift Survey sampled the distribution of galaxies to a distance about four times that of the Harvard–Smithsonian survey and included 26,418 galaxies (Lin et al., 1996). The sizes of the voids in the galaxy distribution were on roughly the same physical scale as those in Fig. 2.7, indicating that the Universe is homogeneous on a large enough scale, consistent with the scaling arguments from the angular two-point correlation functions.

More recently, the statistics of such surveys has been increased by an order of magnitude by the Anglo-Australian Telescope Two-degree Field (2dF) survey of galaxies and the first results of the Sloan Digital Sky Galaxy Survey (Fig. 2.8a and b). In the 2dF survey, the redshifts of over 200,000 galaxies were measured. To reveal the 'cellular' structure more clearly, only a narrow wedge 4° wide is shown in Fig. 2.8a which includes 56,237 galaxies (Colless et al., 2001). The map shown in Fig. 2.8b from the Sloan galaxy survey also includes over 200,000 galaxies (Stoughton et al., 2002). In both cases, the surveys extend to redshifts of about 0.25 and it can be seen that the 'cellular structure persists out to the limits of the surveys. Notice that this means that these surveys have already mapped out a large fraction of the distribution of galaxies at the present epoch.

In Figs. 2.7 and 2.8a and b, the scales of the largest holes are about 30–50 times the scale of a cluster of galaxies, that is, up to about $50h^{-1}$ Mpc. These are the largest known structures in the Universe and one of the major cosmological challenges is to reconcile this gross irregularity in the large-scale spatial distribution of galaxies with the remarkable smoothness of the Cosmic Microwave Background Radiation seen in Fig. 2.2c. Despite the presence of the huge voids however, the amplitude of these irregularities decreases with increasing scale so that on the very largest scales, one bit of Universe looks very much like another.

It is important to have a quantitative description of the large-scale topology of the galaxy distributions shown in Figs. 2.7 and 2.8a and b. In the 1980s, Gott and his colleagues developed techniques for evaluating the topology of the distribution of voids and galaxies from large redshift surveys (Gott et al., 1986; Melott et al., 1988). As they expressed the issue delightfully in their paper:

> We would like to know whether the distribution of galaxies on large scales is best described as a hierarchy of clusters, an irregular lattice of cells or 'bubbles', a network of filaments, or a set of non-intersecting filaments. Loosely speaking, one might describe the topology of these models as respectively a 'meatball' topology, a 'swiss-cheese' topology, a 'sponge' topology and a 'spaghetti' topology.

The importance of these studies is that the topology of the distribution of galaxies is intimately related to the initial conditions from which the large-scale structure formed, in particular, to the common assumption that the perturbations were Gaussian fluctuations with random phases.

From an analysis of the CfA galaxy survey, Gott and his colleagues found that the distribution of the galaxies on the large scale is 'sponge-like', the material of

the sponge representing the location of the galaxies and the holes in the sponge corresponding to the large voids (Gott et al., 1986; Melott et al., 1988). Both the holes and the distribution of galaxies are continuously connected throughout the local Universe. This topology is possible in three dimensions but not in two. Just like a sponge, overall the distribution of material of the sponge and the holes is homogeneous, but, on a small scale, it is highly inhomogeneous.

Similar analyses have been carried out for the large AAT 2dF and SDSS Galaxy surveys by Hoyle, Vogeley, Gott and their colleagues (Hoyle et al., 2002a,b). They find that the overall topology of the two independent surveys are in remarkable agreement and similar to that of a Gaussian random field. They go further and show how further astrophysically important issues can be addressed by considering the topologies of red and blue galaxies separately and comparing these with simulations of the formation of structure in the concordance ΛCDM model. We will return to these issues in Part IV.

Another way of investigating the large-scale distribution of discrete objects in the Universe is to study the distribution of extragalactic radio sources over the sky. Unlike the optical waveband, it turns out that, when a survey of the radio sky is made, the objects which are easiest to observe are extragalactic radio sources associated with certain rare classes of active galaxy, the radio quasars and radio galaxies, at very great distances. Because they are rare objects, they sample the isotropy of the Universe on a very large scale. Figure 2.9 shows the distribution of the brightest 54,579 extragalactic radio sources at a wavelength of 6 cm in the Greenbank Catalogue of radio sources which spans most of the northern hemisphere (Gregory and Condon, 1991; Kooiman et al., 1995). Besides the hole about the North Celestial pole, there are holes in the vicinity of a few intense radio sources. There is also a small excess of sources lying along the Galactic plane but otherwise the distribution is entirely consistent with the sources being distributed uniformly at random over the sky on the large scale.

Kooiman and his colleagues report weak clustering on angular scales 1–2°, consistent with an angular two-point correlation function of the form $w(\theta) \propto \theta^{-0.8}$. A much stronger angular two-point correlation signal was found in the much deeper FIRST survey carried out with the VLA by Helfand and his colleagues (Cress et al., 1996). At a limiting flux density of 1 mJy at 1.4 GHz, 138,665 radio sources were detected and an angular two-point correlation function of the form $w(\theta) \propto \theta^{-1.1}$ measured down to angular separations of a tenth of a degree.

The radio sources are ideal for probing the large-scale distribution of discrete objects since they are readily observed at large distances. The bulk of the radio sources plotted in Fig. 2.9 lie at redshifts $z \geq 1$ and so they sample the distribution of discrete sources on the largest physical scales accessible to us at the present epoch. Notice that the extragalactic radio sources provide complementary information to that provided by the Cosmic Microwave Background Radiation, in that they refer to the large-scale distribution of discrete objects, such as galaxies, once they have formed.

On fine scales, the clustering of galaxies takes place on a very wide variety of scales from pairs and small groups of galaxies, such as the Local Group of galaxies, to giant *clusters* of galaxies, such as the Coma and Pavo clusters which can

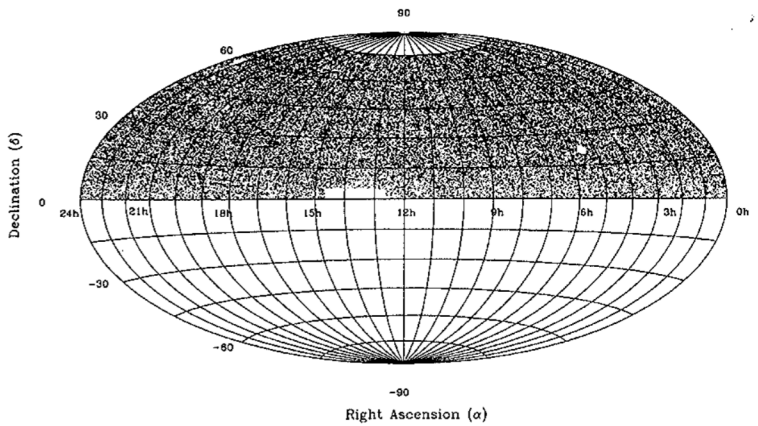


Fig. 2.9. The distribution of radio sources in the Greenbank Catalogue of radio sources at 6 cm (Gregory and Condon, 1991). The picture includes 54,579 radio sources with angular sizes less than 10.5 arcmin and flux densities $S \geq 25$ mJ. The area about the North Celestial pole was not surveyed. There are 'holes' in the distribution about a few bright radio sources and a small excess of sources associated with the Galactic plane. Otherwise, the distribution does not display any significant departure from a random distribution on a large scale, although there is evidence for clustering on the scale of 1° (Kooiman et al., 1995)

contain thousands of members – we discuss some of their properties in Chap. 4. The rich regular clusters are self-gravitating bound systems, but there are also irregular clusters which have an irregular, extended appearance and it is not so clear that these are bound systems.

The term *supercluster* is used to describe structures on scales larger than those of clusters of galaxies. They may consist of associations of clusters of galaxies, or a rich cluster with associated groups and an extended distribution of galaxies. Some authors would classify the 'stringy' structures seen in Figs. 2.7 and 2.8 as superclusters, or supercluster cells. From the physical point of view, the distinction between the clusters and the superclusters is whether or not they are gravitationally bound. Even in the rich, regular clusters of galaxies, which have had time to relax to a state of dynamical equilibrium, there has only been time for individual galaxies to cross the cluster up to about 10 times in the age of the Universe and so, on larger scales, there is scarcely time for the systems to become gravitationally bound. Our own Galaxy and the Local Group of galaxies are members of what is known as the *Local Supercluster*. This is the huge flattened distribution of galaxies centred on the Virgo cluster, which lies at a distance of about 15–20 Mpc from our own Galaxy. It can be seen very prominently in maps of the distribution of bright galaxies running more or less perpendicular to the plane of the Galaxy and is outlined by filled circles in Fig. 2.10 (Kolatt et al., 1995).

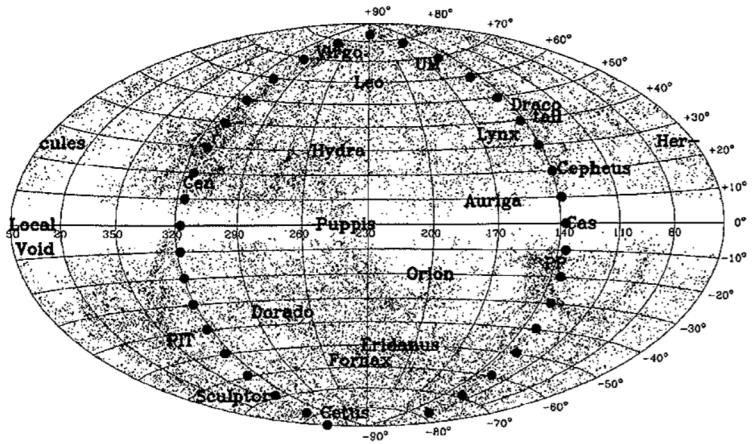


Fig. 2.10. An equal area projection drawn in galactic coordinates of the distribution of bright galaxies over the whole sky. In the north, the galaxies are taken from the UGC catalogue and, in the south, from the ESO, UGC and MCG catalogues. There is an absence of galaxies at low galactic latitudes because of extinction by interstellar dust. The prominent band of galaxies intersecting the Galactic plane at right angles at $l \sim 320°$ is the Local Supercluster of Galaxies and the Supergalactic plane is delineated by the filled circles. Some regions in which prominent clustering of galaxies is found are labelled by the constellations in which these lie (Kolatt et al., 1995)

We conclude that, on the very largest scales, the distribution of matter and radiation is remarkably isotropic and homogeneous. This greatly simplifies the construction of cosmological models.

2.3 Hubble's Law and the Expansion of the Universe

Hubble made his great discovery of the velocity–distance relation for galaxies in 1929 (Hubble, 1929). A modern version of Hubble's law, in the form of an redshift–apparent magnitude relation or Hubble diagram, is shown in Fig. 2.11 for the brightest galaxies in clusters (Sandage, 1968). It is found empirically that the brightest galaxies in nearby clusters all have more or less the same intrinsic luminosities and so their apparent magnitudes can be used to estimate relative distances by application of the inverse square law.

For a class of galaxy of fixed intrinsic luminosity L, the observed flux density S is given by the inverse square law, $S = L/4\pi r^2$ and so, converting this relation into astronomical apparent magnitudes m using the standard relation $m = \text{constant} - 2.5 \log_{10} S$, it follows that

$$m = 5 \log_{10} r + \text{constant} . \qquad (2.12)$$

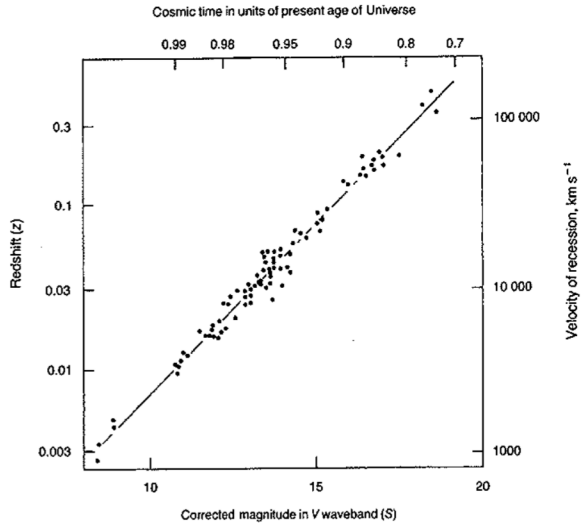


Fig. 2.11. A modern version of the velocity–distance relation for galaxies for the brightest galaxies in rich clusters of galaxies. This correlation indicates that the brightest galaxies in clusters have remarkably standard properties and that their velocities of recession from our own Galaxy are proportional to their distances (Sandage, 1968)

This was the approach adopted by Hubble and Humason in their pioneering analysis of 1934 (Hubble and Humason, 1934) – they assumed that the 5th brightest galaxy in a cluster would have more or less the same intrinsic luminosity (Fig. 1.5b). In Fig. 2.11, the corrected apparent magnitude in the V waveband is plotted against the logarithm of the redshift of the brightest galaxies in a number of rich clusters of galaxies which span a wide range of redshifts. The *redshift* z is defined by the formula

$$z = \frac{\lambda_{\text{obs}} - \lambda_{\text{em}}}{\lambda_{\text{em}}} , \qquad (2.13)$$

where λ_{em} is the emitted wavelength of some spectral feature and λ_{obs} is the wavelength at which is it observed. In the limit of small velocities, $v \ll c$, if the redshift is interpreted in terms of a recessional velocity v of the galaxy, $v = cz$ and this is the type of velocity plotted in the velocity–distance relation. It is an unfortunate tradition in optical astronomy that the splendidly dimensionless quantity, the redshift z, is converted into a velocity by multiplying it by the speed of light. As we will see below, interpreting the redshift in terms of a recessional velocity leads to confusion and misunderstanding of its real meaning in cosmology. It is best if all mention of recessional velocities are expunged in developing the framework of cosmological models.

The solid line shown in Fig. 2.11 is $m = 5\log_{10} z +$ constant, corresponding to $v \propto r$, and it runs precisely through the observed points – correlations do not come any better than this in cosmology. The velocity–distance relation is normally written $v = H_0 r$, where H_0 is known as *Hubble's constant*. The velocity–distance relation appears to hold good for all classes of extragalactic system, including the active galaxies and quasars.

We discussed in detail recent evidence for the homogeneity of the distribution of galaxies in space in Sect. 2.2. In fact, Hubble realised in the 1930s that a simple test of the homogeneity of the Universe is provided by the number counts of galaxies. As we will show ins Sect. 17.2.1, in a homogeneous Universe it is expected that the number counts of galaxies follow the law $N(\geq S) \propto S^{-3/2}$, where S is the observed flux density of the galaxy. This result is independent of the luminosity function of the sources so long as the counts do not extend to such large distances that the effects of the cosmological redshift have to be taken into account. In terms of apparent magnitudes, this relation becomes $N(\leq m) \propto 10^{0.6m}$.

Hubble found that the counts of galaxies to about 20th magnitude more or less followed this relation, although they showed some convergence at the faintest apparent magnitudes, which Hubble interpreted as evidence for the effects of space curvature at large distances (Hubble, 1936). More recent counts of galaxies are shown in Fig. 2.12 which extend to very faint apparent magnitudes (Metcalfe et al., 1996). The results are similar to those of Hubble at $m \leq 20$; the counts are slightly flatter than the Euclidean predictions, but are entirely consistent with the expectations of uniform world models once the effects of observing the populations at significant cosmological distances are taken into account. Divergences from the expectations of the uniform models occur at much fainter blue apparent magnitudes ($B \geq 22$), in the sense that there is an excess of faint blue galaxies; we will take up the origin of this excess in detail in Chap. 17.

The combination of the observed large-scale isotropy and homogeneity of the Universe with Hubble's law shows that the Universe as a whole is expanding uniformly at the present time. Let us show this formally by the following simple calculation. Consider a uniformly expanding system of points (Fig. 2.13). The definition of a uniform expansion is that the distances between any two points should increase by the same factor in a given time interval, that is, we require

$$\frac{r_1(t_2)}{r_1(t_1)} = \frac{r_2(t_2)}{r_2(t_1)} = \cdots = \frac{r_n(t_2)}{r_n(t_1)} = \cdots = \alpha = \text{constant}\,, \tag{2.14}$$

for any set of points. Let us select some galaxy at random and take the distances of the other galaxies from it to be $r_1, r_2, \ldots$. Then, the recession velocity of galaxy 1 relative to the chosen origin is

$$\begin{aligned} v_1 &= \frac{r_1(t_2) - r_1(t_1)}{t_2 - t_1} = \frac{r_1(t_1)}{t_2 - t_1}\left[\frac{r_1(t_2)}{r_1(t_1)} - 1\right] \\ &= \frac{r_1(t_1)}{t_2 - t_1}(\alpha - 1) = H_0 r_1(t_1)\,. \end{aligned} \tag{2.15}$$

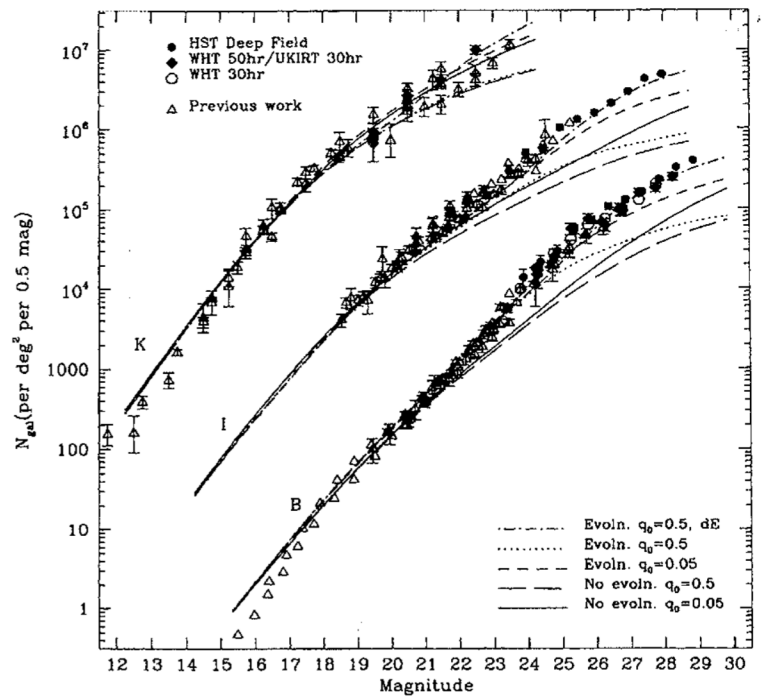


Fig. 2.12. The counts of faint galaxies observed in the B, I and K wavebands compared with the expectations of various uniform world models, as well as other models in which various forms of the evolution of the luminosity function of galaxies with redshift are assumed (Metcalfe et al., 1996). The galaxy counts follow closely the expectations of uniform world models at magnitudes less than about 22, but there is a excess of galaxies in the B and I wavebands at fainter magnitudes

Similarly, for the nth galaxy,

$$v_n = \frac{r_n(t_1)}{t_2 - t_1}(\alpha - 1) = H_0 r_n(t_1)\,. \tag{2.16}$$

Thus, a uniformly expanding distribution of galaxies automatically results in a velocity–distance relation of the form $v \propto r$.

This analysis is, however, much deeper than simply an explanation of the local velocity–distance relation. Notice that the above analysis applies to galaxies at all distances in a uniformly expanding universe. The requirements of isotropy and homogeneity mean that the same linear velocity–distance relation must hold true *at all distances*, including at distances at which the recession velocities exceed the

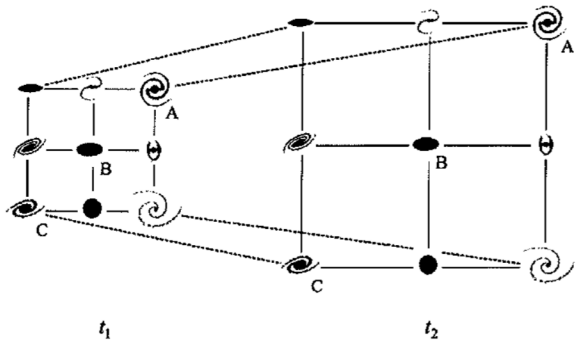


Fig. 2.13. Illustrating the origin of the velocity–distance relation for an isotropically expanding distribution of galaxies. The distribution of galaxies expands uniformly between the epochs t_1 and t_2. If, for example, we consider the motions of the galaxies relative to the galaxy A, it can be seen that galaxy C travels twice as far as galaxy B between the epochs t_1 and t_2 and so has twice the recession velocity of galaxy B relative to A. Since C is always twice the distance of B from A, it can be seen that the velocity–distance relation is a general property of isotropically expanding universes

speed of light. A good way of understanding this assertion is to consider the familiar example of the surface of an expanding balloon as an analogue for the kinematics of an expanding universe. As the balloon is blown up, the distances between points on the sphere increase and everywhere a velocity–distance relation is obtained locally. When the balloon is blown up to a huge size, widely separated points on the sphere separate at speeds which can be greater than the speed of light. There is, however, nothing unphysical about this since there is no causal connection between the points. The points are simply partaking in a geometrically uniformly expansion which takes place over enormous distances. We will return to this important point in Chap. 12.

2.4 Conclusion

The upshot of the considerations of this chapter is that the correct starting point for the construction of models for the large-scale dynamics of the Universe is that they should be isotropic and homogenous on the large scale and that they should be uniformly expanding. These are enormous simplifications and, taken in conjunction with the General Theory of Relativity, provide a set of simple world models which provide the framework within which we can study the problems of the origin of galaxies and the other large-scale structures we observe in the Universe today.

3 Galaxies

3.1 Introduction

'Galaxy formation' is the title of this book and so we should summarise what is known about the properties of galaxies. This is not a trivial business because normal galaxies are complex, many-body systems. Typically, a galaxy can consist of hundreds of millions or billions of stars, it can contain considerable quantities of interstellar gas and dust and can be subject to environmental influences through interactions with other galaxies and with the intergalactic gas. Star formation takes place in dense regions of the interstellar gas. To complicate matters further, it is certain that dark matter is present in galaxies and in clusters of galaxies and the dynamics of galaxies are largely dominated by this invisible dark component. Its nature is, however, unknown.

Until the advent of the massive surveys of galaxies, the 2dF Galaxy Survey undertaken by the Anglo-Australian Telescope and the Sloan Digital Sky Survey (SDSS), the typical properties of galaxies were defined by meticulous morphological studies of large samples of bright galaxies. Being bright and relatively nearby, the morphological classification schemes had to encompass a vast amount of detail and this was reflected in Hubble's pioneering studies as elaborated by de Vaucouleurs, Kormendy, Sandage, van den Bergh and others. In contrast, the sheer size of the new galaxy samples, which each encompass about 200,000 galaxies, has meant that classification schemes had to be based upon parameters which could be derived from computer analysis of the galaxy images and spectra. What the new approach loses in detail, it more than makes up for in the huge statistics involved and in the objective nature of the classification procedures.

These recent developments have changed the complexion of the description of the properties of galaxies. While the new samples provide basic global information about the properties of galaxies, the old schemes describe many features which need to be incorporated into the understanding of the detailed evolution and internal dynamics of particular classes of galaxy. The upshot is that, we need to develop in parallel both the traditional and more recent approaches to the classification of galaxies. We will summarise some of their more important properties, as well as elucidating some of the essential physics. In the next chapter, we will perform a similar exercise for clusters of galaxies. The books *Galaxies in the Universe: an Introduction* by Sparke and Gallagher, *Galactic Astronomy* by Binney and Merrifield and *Galactic Dynamics* by Binney and Tremaine can be thoroughly recommended

as much more thorough introductions to these topics (Sparke and Gallagher, 2000; Binney and Merrifield, 1998; Binney and Tremaine, 1987). The results of the 2dF Galaxy Survey and the Sloan Digital Sky Survey (SDSS) are too recent to have entered the textbooks. We begin with the traditional approach and then relate these studies to the more global approach adopted in recent analyses.

3.2 The Revised Hubble Sequence for Galaxies

Let us first describe the traditional approach to the classification of galaxies, noting points of contact with more recent statistical approaches. Galaxies come in a bewildering variety of different shapes and forms. In order to put some order into this diversity, classification schemes were devised on the basis of their visual appearances, or morphologies, originally on photographic plates but nowadays from digital images taken with CCD cameras. The basis of the traditional morphological schemes remains the *Hubble Sequence of Galaxies*, described in Hubble's monograph *The Realm of the Nebulae* (Hubble, 1936). The Hubble sequence, sometimes referred to as a 'tuning-fork' diagram, arranges galaxies into a continuous sequence of types with elliptical galaxies at the left-hand end and spirals at the right-hand end (Fig. 3.1). The spiral galaxies are ordered into two branches named 'normal' and 'barred' spirals. Conventionally, galaxies towards the left-hand end of the sequence are referred to as 'early-type' galaxies and those towards the right as 'late-type' galaxies, reflecting Hubble's original prejudice concerning their evolution from one type to another. Despite the fact that these ideas have long outlived their usefulness, the terms are still in common use, even in the era of the massive surveys of galaxies.

Morphological classification schemes such as the Hubble sequence become an integral part of astrophysics when independent properties of galaxies are found to correlate with the morphological classes. This has been found to be the case for

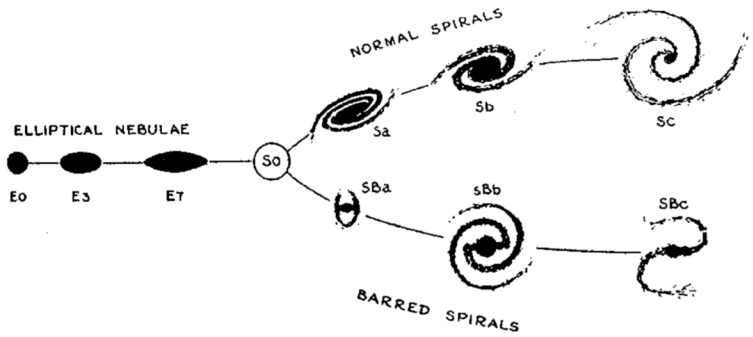


Fig. 3.1. The Hubble sequence of galaxies as presented in *The Realm of the Nebulae* (Hubble, 1936)

a number of the overall properties of galaxies such as their integrated colours, the fraction of the mass of the galaxy in the form of neutral and molecular gas, and so on (Sect. 3.8). The vast majority of galaxies can be accommodated within the *revised Hubble Sequence of Galaxies*, which is described in detail by de Vaucouleurs, Sandage, Kormendy and van den Bergh (de Vaucouleurs, 1974; Sandage, 1975; Kormendy, 1982; van den Bergh, 1998). Van den Bergh emphasised that the classical Hubble types refer primarily to intrinsically luminous galaxies and that, in addition, there exists a large population of intrinsically low luminosity *dwarf galaxies* which can only be observed relatively nearby. There are also various other categories of galaxies with special characteristics, for example, the Seyfert galaxies, cD galaxies, N galaxies, radio galaxies, starburst galaxies and so on. Many of these types of galaxy contain *active galactic nuclei*.

***Elliptical galaxies* E.** These galaxies show no structural features in their brightness distributions but have an elliptical appearance, as if they were spheroids or ellipsoids of revolution (Fig. 3.2a). In absolute magnitude, elliptical galaxies range from among the most luminous galaxies known, having $M_B \approx -24$, to dwarf ellipticals (dE), which are found in the Local Group of galaxies. In Hubble's notation, the observed ellipticity of the galaxy is included in the morphological designation according to the rule that the number $10 \times (a - b)/a$ was written after the letter E, where a and b are the observed major and minor axes of the ellipse. Thus E0 galaxies are circular and E7 galaxies, the most extreme ellipticities found in elliptical galaxies, have $b/a = 0.3$. Galaxies flatter than E7 all show a distinct disc and bulge structure and hence are classified as lenticular (S0) rather than E galaxies.

***Spiral galaxies* S, SA, SB.** The characteristic feature of spiral galaxies is their disk-like appearance with well-defined spiral arms emanating from their central regions (Fig. 3.2b and 3.3a). Very often the spiral pattern is double with a remarkable degree of symmetry with respect to the centre of the galaxy, but many more complicated configurations of spiral structure are known. The light distribution of what Hubble termed 'normal' spiral galaxies (or SA galaxies) can be decomposed into a *central bulge* or *spheroidal component*, similar in character to an elliptical galaxy, and a *disk component*, within which the spiral arms lie. In the case of the *barred spirals* (or SB galaxies), the central bulge has an elongated or ellipsoidal appearance, the spiral arms originating from the ends of the bar (Fig. 3.3a). There are as many 'barred' spiral galaxies as 'normal' spirals and, furthermore, there are just as many spirals intermediate between these two classes.

Spiral galaxies are classified as Sa, Sb, Sc according to the following criteria, in decreasing order of importance: (1) the openness of the winding of the spiral arms, (2) the degree of resolution of the arms into stars and (3) the size of the spheroidal component or central bar relative to the disk component. Thus,

- *Sa galaxies* have tightly wound spiral arms which are smooth showing no resolution into stars. The central bulge or bar is dominant, shows no structure and is unresolved into star clusters.
- *Sb galaxies* have more open spiral arms, which show resolution into stars. The central spheroidal component or bar is generally smaller than in Sa galaxies.

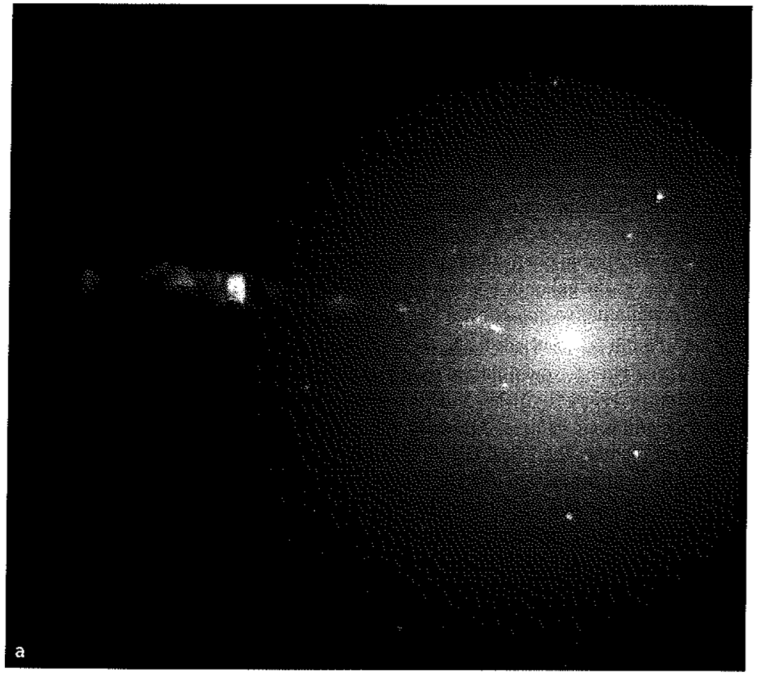


Fig. 3.2a,b. Examples of different types of normal galaxy. **a** The elliptical galaxy M87 (NGC 4486), one of the most luminous galaxies in the nearly Virgo cluster of galaxies. This image shows its famous optical jet originating in the bright nucleus of the galaxy which contains a black hole of mass $M \sim 10^9\, M_\odot$

- *Sc galaxies* have very open spiral arms which are patchy and are resolved into star clusters and regions of ionised hydrogen. The spheroidal component is very small. In barred spiral galaxies, the bar is resolved into clusters and HII regions and is not as prominent as in classes Sa or Sb.
- The revised Hubble scheme extends this classification beyond Sc to include 'nearly chaotic' structures which would have been classified as very late Sc spirals in the standard sequence but are now as classified Sd spirals.

These morphological classes are rather broad and intermediate stages along the sequence are defined as Sab, Sbc and Scd. It is found that, within a given class of spiral galaxy, the importance of the bulge can vary considerably. It is interesting to compare this approach with the more recent statistical approach in which the bulge-to-disc ratio is one of the key measurable parameters.

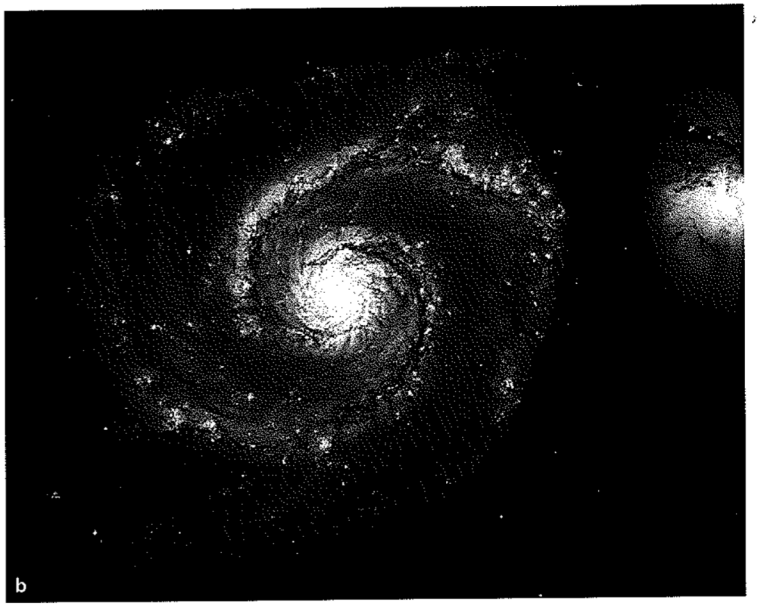


Fig. 3.2. (continued) **b** The spiral galaxy M51 (NGC 5194) and its nearby dusty companion (NGC 5195). (Images courtesy of NASA, ESA and the Hubble Heritage Team (STScI/AURA))

***Lenticular galaxies* S0 *or* L.** All galaxies with smooth light distributions and axial ratios $b/a < 0.3$ show evidence of a disk-like component and these are called *lenticular* (lens-like) or S0 galaxies (Fig. 3.3b). They are similar to spiral galaxies in that their light distributions can be decomposed into a central bulge, similar in properties to elliptical galaxies, and an extensive disk. The lenticular galaxies appear intermediate in morphological type between elliptical and spiral galaxies.

In many cases, the central bulges of the S0 galaxies have a bar-like appearance and hence, as in the case of the spirals, they can be divided into 'ordinary' and 'barred' lenticulars as well as intermediate types. In a number of lenticular galaxies, there is evidence for obscuring matter, often in the form of rings as can be seen in the example of NGC 1300 in Fig. 3.3b. In the revised Hubble classification, lenticular galaxies which are free of obscuring matter are termed 'early' $S0^-$ with stages $S0^0$ and $S0^+$ representing 'later' stages with increasing amounts of obscuring material. By the intermediate stage between lenticular and spiral galaxies, S0/a, the obscuring matter begins to show what is referred to as 'incipient spiral structure'.

Irregular galaxies. In Hubble's original classification, irregular galaxies were systems 'lacking both dominating nuclei and rotational symmetry' and the class included everything which could not be readily incorporated into the standard Hubble

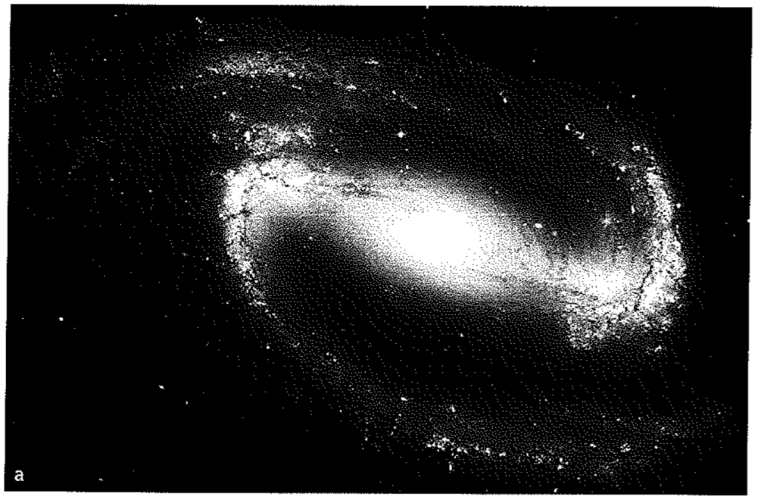


Fig. 3.3a,b. Examples of different types of normal galaxy. **a** The barred spiral galaxy NGC 1300

sequence. Many of these irregulars were similar to the companion galaxies of our own Galaxy, the Magellanic Clouds, and these were designated Irr I or Magellanic irregulars. There remained a small class of irregulars consisting of galaxies such as M82 , NGC 520 and NGC 3077, in which there was no evidence of resolution into stars; these galaxies were classified Irr II galaxies.

Evidence that the Irr I galaxies form a natural extension of the Hubble sequence was provided by de Vaucouleurs' discovery of weak but definite spiral structure in the Large Magellanic Cloud (LMC). Galaxies like the LMC can be considered to belong to stages in the Hubble sequence later than Sd and are denoted Sm. Thus, the late stages of the sequence reads: Scd, Sd, Sdm, Sm, Im. The Irr II systems find no natural place in the revised sequence and are designated I0 by de Vaucouleurs. The characteristics of the I0 irregular galaxies are that they are very rich in interstellar matter and contain young stars and active regions of star formation; a number of these would be classified as starburst galaxies.

In the revised Hubble sequence, shown in tabular form in Table 3.1, the various stages along the sequence are assigned numbers ranging from −6 to 11. All transitions along the sequence are smooth and continuous. The frequencies with which different types of galaxy are found among catalogues of bright galaxies are shown in Tables 3.2 and 3.3 (de Vaucouleurs 1963). Striking features of this table are the large percentage of lenticular galaxies and the roughly equal proportions of normal, barred and intermediate spiral galaxies. The latter statistics indicate that, in well over half the known examples of spiral galaxies, there are bar-like structures

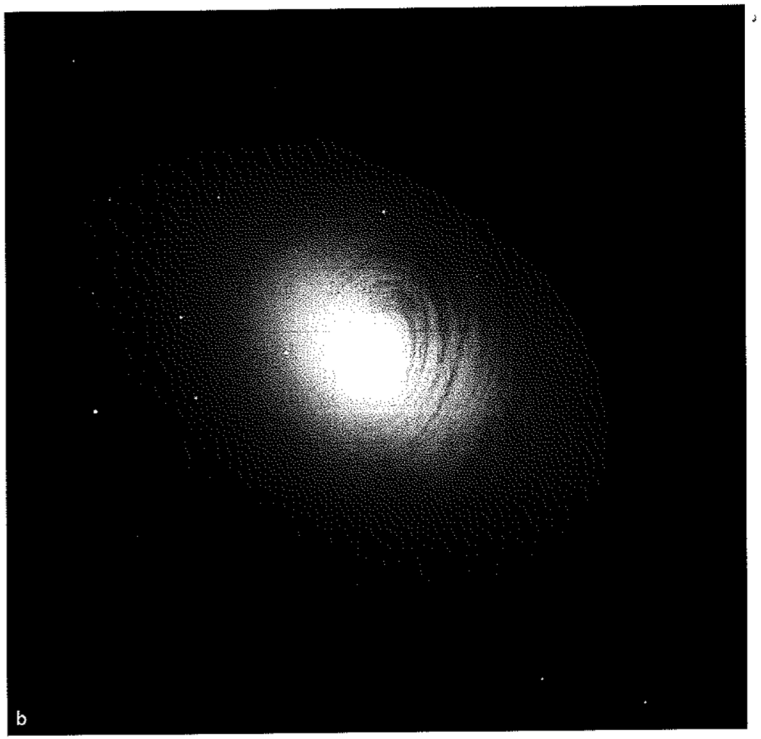


Fig. 3.3. (continued) **b** The SB0 or lenticular galaxy NGC 2787. (Images courtesy of NASA, ESA and the Hubble Heritage Team (STScI/AURA))

Table 3.1. The revised Hubble sequence of galaxies according to de Vaucouleurs' classification (de Vaucouleurs, 1974)

−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10	11
E^-	E^0	E^+	$S0^-$	$S0^0$	$S0^+$	S0a	Sa	Sab	Sb	Sbc	Sc	Scd	Sd	Sdm	Sm	Im	I0

in their central regions and this has important implications for the origin of spiral structure.

The figures given in Tables 3.2 and 3.3 must be treated with considerable caution. First of all, the galaxies included in the above statistics are those present in bright galaxy catalogues. They therefore refer to objects of a very wide range of intrinsic luminosities, from among the most luminous galaxies known, such as M87, to nearby dwarf galaxies. Secondly, as mentioned above, the classical Hubble types refer

Table 3.2. The frequencies with which galaxies of different morphological types are found among samples of bright galaxies (de Vaucouleurs, 1963)

Class of Galaxy	E	L	S	Im	I0	Pec	Total
Number	199	329	934	39	13	14	1528
Percentage	13.0	21.5	61.1	2.55	0.85	0.9	100

Table 3.3. The frequencies of different subtypes among 994 spiral galaxies (de Vaucouleurs, 1963)

Class of Galaxy	0/a	a	ab	b	bc	c	cd	d	dm	m	?	Total	Percentage
SA	17	25	25	57	57	82	30	9	3	4	2	311	31.3
SAB	13	15	23	45	50	71	35	11	3	7	1	274	27.6
SB	26	43	33	83	27	55	27	28	9	30	10	366	36.8
S	4	1	0	6	1	13	1	10	0	0	7	43	4.3

almost exclusively to intrinsically luminous galaxies. Thirdly, the above statistics include galaxies belonging to the general field, to weak groups and to rich clusters of galaxies, but the fractions of the different morphological types vary with the local galaxy number density. In an important paper, Dressler plotted the frequency of different galaxy types as a function of the number density of galaxies in which they are found (Fig. 3.4) (Dressler, 1980). Field galaxies, that is, galaxies which are not members of groups or clusters of galaxies, are located towards the left of the diagram, while rich clusters of galaxies are towards the right. It can be seen that, in rich clusters, the elliptical and S0 galaxies are much more common than the spiral galaxies, whereas in the general field, the majority of galaxies are spirals. Evidently, the environment in which a galaxy finds itself is correlated with its morphological characteristics. These relations will be quantified in more detail in Sect. 3.9.4.

3.3 Peculiar and Interacting Galaxies

The revised Hubble classification can encompass the forms of virtually all galaxies. There are, however, a number of galaxies with very strange appearances and these are referred to collectively as *peculiar galaxies*. Arp published his *Atlas of Peculiar Galaxies* in 1966 and a corresponding catalogue for the Southern Hemisphere 1987 (Arp, 1966; Arp et al., 1987). As he remarked in the introduction to the first Atlas:

> The greatest deviations from the normal are emphasised in this atlas.

A few galaxies are known, for example, in which the stellar component is in the form of a ring rather than a disc or spheroid, the Cartwheel being a beautiful example of this type of galaxy (Fig. 3.5); these are known as *ring galaxies*.

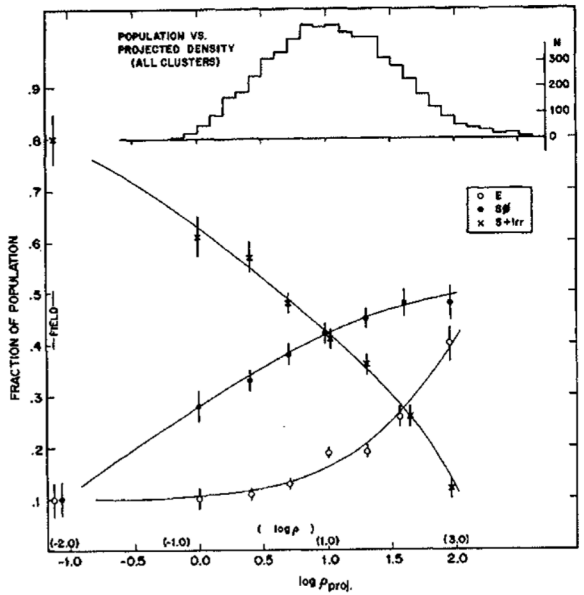


Fig. 3.4. The fractions of different morphological types of galaxy found in different galaxy environments. The local number density of galaxies is given as a projected surface density, $\varrho_{\rm proj}$ of galaxies, that is, numbers Mpc^{-2} (Dressler, 1980)

Most of these remarkable structures are due to strong gravitational interactions, or collisions, between galaxies. In the early 1970s, Toomre and Toomre carried out pioneering computer simulations of close encounters between galaxies which showed how such events could give rise to remarkable asymmetric structures (Toomre and Toomre, 1972). In Fig. 3.6, a deep image of the pair of interacting galaxies known as the Antennae is shown, revealing the extraordinary long 'tails' which seem to be emanating from a pair of closely interacting spiral galaxies in which a great deal of recent star formation has occurred.

The Toomres showed how such elongated 'tails' could be accounted for by a gravitational interaction between two spiral galaxies. In the simulation shown in Fig. 3.7, the two spiral galaxies pass close to each other on prograde orbits, that is, the rotational axes of the two discs are parallel and also parallel to the rotational axis of the two galaxies about their common centre of mass. The spiral galaxies are represented by differentially rotating discs of stars and, while they are at their distance of closest approach, the stars in the outermost rings feel the same mutual force acting upon them for a very much longer time than if the passage had been in, say, the retrograde direction. As a result, the outer rings of stars feel

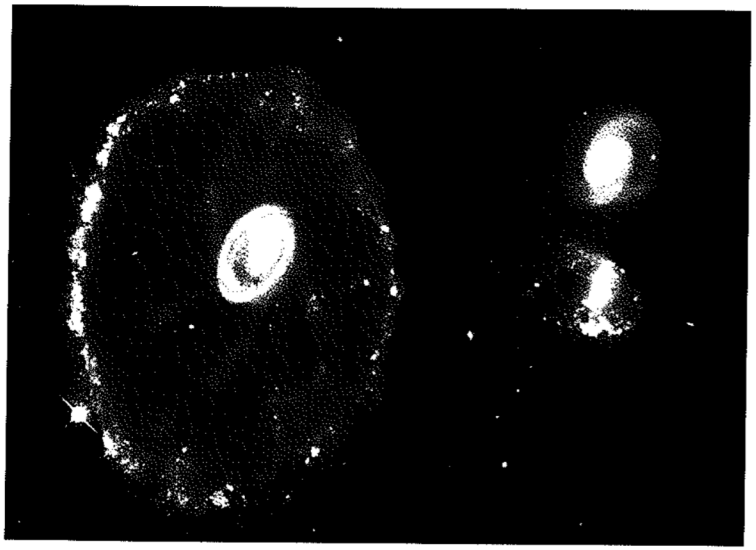


Fig. 3.5. The peculiar galaxy known as the Cartwheel as observed by the Hubble Space Telescope. (Image courtesy of NASA, ESA and the Hubble Heritage Team (STScI/AURA)) Its strange appearance is almost certainly due to a recent collision, or strong interaction, with one of its nearby companions. The simulations by Toomre and Toomre show that such a 'tidal wave' is expected if a compact mass had passed through a spiral galaxy close to its centre (Toomre, 1974)

a coherent force for an extended period and are stripped off to form the types of extended structure observed in the Antennae. Many of the features of strong gravitational interactions are described in the pioneering papers by the Toomres. Similar structures are found in more recent supercomputer simulations of colliding and interacting galaxies which can include millions of stars, as well as incorporating the dynamics of the interstellar gas in the collision (Barnes and Hernquist, 1996; Mihos and Hernquist, 1996; Springel, 2005). The compression of the interstellar media in the collision results in regions of intense star formation.

Interactions between galaxies play a central role in many aspects of galactic evolution. From the observational point of view, the IRAS satellite showed that colliding galaxies are among the most luminous extragalactic far-infrared sources. The inference is that, when galaxies collide, their interstellar media are compressed to high densities and the rate of star formation is greatly enhanced, resulting in intense far-infrared emission.

Collisions between galaxies have also assumed a central role in models of galaxy formation. In the preferred scenarios of structure formation, galaxies are built up

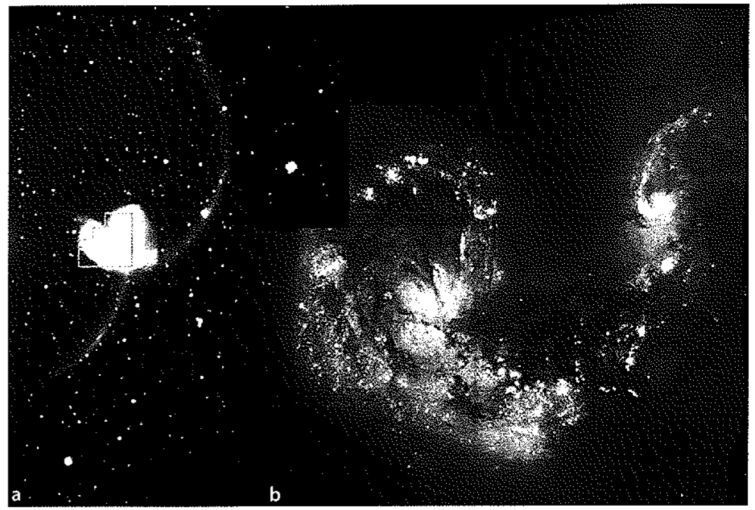


Fig. 3.6a,b. The strange pair of interacting galaxies known as the Antennae, named after the elongated structures apparently torn off in a collisions between the galaxies. **a** A wide field image showing the extended 'tails' produced in the encounter. **b** A 'true colour' Hubble Space Telescope image of the colliding galaxies. The many blue regions in the diagram contain young luminous stars, indicating that a large amount of star formation has been induced by the collision between the interstellar media of the interacting galaxies. (Image courtesy of NASA, ESA and the Hubble Heritage Team (STScI/AURA))

by the process of hierarchical clustering in which larger galaxies are formed by the coalescence of smaller galaxies. In this picture, strong gravitational encounters between galaxies are essential in forming the structures we observe today. Reference to Table 3.2 shows that the percentage of peculiar and interacting systems among the present population of galaxies is only about 1%. We will find that this percentage increases dramatically as we look further and further back in time, consistent with the hierarchical picture of structure formation.

3.4 The Light Distribution in Galaxies

Another approach to the classification of galaxies is to use their light distributions since it is found that these are somewhat different for bulge-dominated and disc-dominated systems. Let us first summarise the results of studies of bright galaxies and then show how these can be adapted for the study of large samples of galaxies.

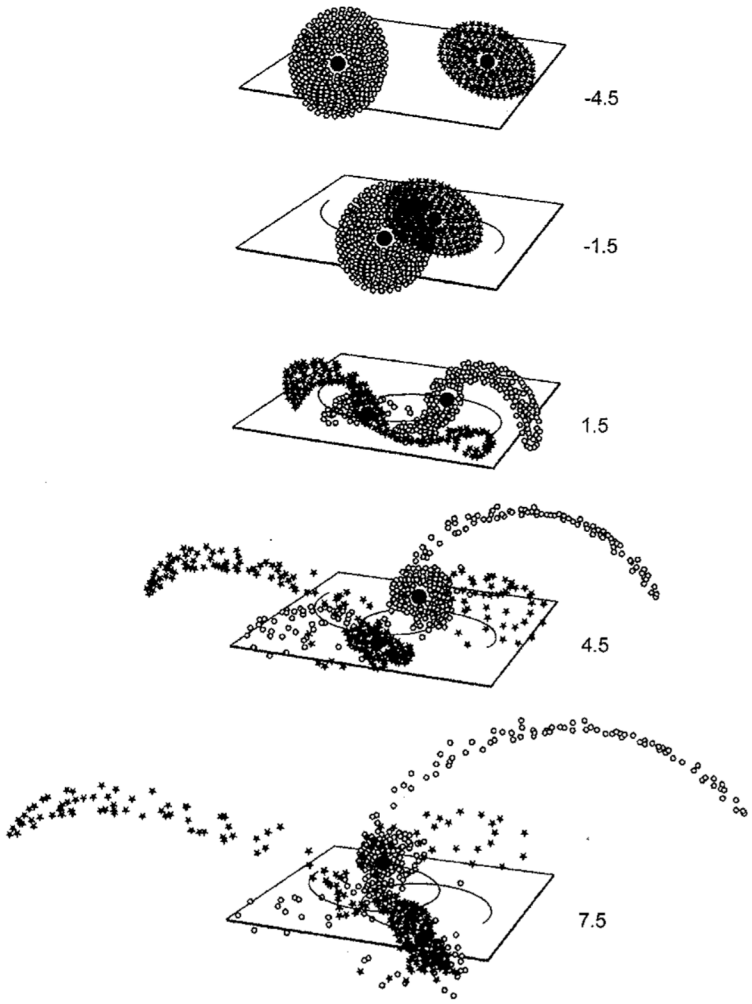


Fig. 3.7. A simulation of a close encounter between two disc galaxies which approach each other on prograde orbits. It can be seen that the outer rings of stars are torn off each galaxy, forming the remarkable 'Antenna' structures (Toomre and Toomre, 1972)

3.4.1 Elliptical Galaxies

The earliest expression for the observed surface brightness distribution of elliptical galaxies as a function of radius r is commonly referred to as *Hubble's law*

$$I(r) = I_0 \left(\frac{r}{r_c} + 1 \right)^{-2} , \tag{3.1}$$

where r_c is the *core radius* of the galaxy. This expression provides a reasonable description of the intensity distribution in the central regions of elliptical galaxies but is a poor fit in the outer regions.

A much better description of the surface brightness distribution of elliptical galaxies and the bulges of spiral galaxies is the empirical law proposed by de Vaucouleurs which is usually referred to as the $r^{1/4}$ law (de Vaucouleurs, 1948)

$$\log_{10} \left[\frac{I(r)}{I(r_e)} \right] = -3.3307 \left[\left(\frac{r}{r_e} \right)^{1/4} - 1 \right] , \tag{3.2}$$

or

$$\log_e \left[\frac{I(r)}{I(r_e)} \right] = -7.6692 \left[\left(\frac{r}{r_e} \right)^{1/4} - 1 \right] . \tag{3.3}$$

This expression provides a good representation of the luminosity profile over many decades of surface brightness. The expression has been normalised so that r_e is the radius within which half the total luminosity is emitted and $I(r_e)$ is the surface brightness at that radius. The corresponding total luminosity of the galaxy is

$$L = 7.215 \pi I_e r_e^2 \left(\frac{b}{a} \right) , \tag{3.4}$$

where b/a is the apparent axis ratio of the elliptical galaxy (Gilmore et al., 1989).

3.4.2 Spiral and Lenticular Galaxies

The light distributions in most spiral and lenticular galaxies can be decomposed into two components, a spheroidal component associated with the central bulge and a disc component. The luminosity profile of the spheroidal component is the same as that of an elliptical galaxy and may be described by the de Vaucouleurs $r^{1/4}$ law discussed above.

In almost all galaxies in which there is evidence of a disc component, including spirals, barred spirals and lenticulars, the luminosity profile of the disc may be represented by an exponential light distribution

$$I(r) = I_0 \exp(-r/h) , \tag{3.5}$$

where h is called the *disc scale length*; for our own Galaxy $h \approx 3$ kpc. I_0 is the central surface brightness. The total luminosity of the disc is then $L = 4\pi h^2 I_0$. According to Freeman, this luminosity profile is also found in very late-type galaxies, such as the Magellanic irregulars which show evidence of a disc in rotation.

In 1970, Freeman discovered the remarkable result that, although the disc components of large disc galaxies have a wide range of luminosities, there is remarkably little scatter in the value of the central surface brightness I_0 (Freeman, 1970). A mean value of $I_0 = 21.67 \pm 0.3$ B magnitudes arcsec^{-2} was found for the galaxies in the sample studied by Freeman, the differences in total luminosity being due to variations in the scale length of the light distribution h from galaxy to galaxy. There has been considerable debate about the validity of this result because the samples of galaxies studied are strongly influenced by selection effects, in particular, the galaxies have to be bright enough and large enough for precise surface photometric observations to be possible. Disney suggested that the constancy of the central surface brightness of the discs of spiral galaxies could be largely attributed to these selection effects (Disney, 1976). Van der Kruit surveyed a number of attempts to remove the effects of observational selection from the samples studied and concluded that 'non-dwarf galaxies do have a relatively narrow dispersion of central surface brightnesses, and this is not the result of selection effects' (Gilmore et al., 1989). It is certainly the case that, for low luminosity spiral galaxies, the values of central surface brightness are smaller Freeman's standard value, the most extreme cases being the low surface brightness disc galaxies which can have I_0 as low as 25.5 B magnitudes arcsec^{-2}.

3.4.3 Putting the Light Distributions Together

A convenient way of combining the light distributions of elliptical and spiral galaxies is to adopt the formulation proposed by Sérsic which can be thought of as a generalisation of de Vaucouleurs' $r^{1/4}$ law (Sérsic, 1968)

$$\log_{10}\left[\frac{I(r)}{I(r_e)}\right] = -b_n\left[\left(\frac{r}{r_e}\right)^{1/n} - 1\right]. \qquad (3.6)$$

where r_e is the radius within which half of the total light is emitted and the b_n is a normalisation constant to ensure that the total light sums to L_{tot} for a given value of n. It can be seen that the value $n = 4$ results in de Vaucouleurs' $r^{1/4}$ law and $n = 1$ in the exponential law found in the discs of spiral galaxies.

This formalism can be used to discriminate between disc-dominated and bulge-dominated galaxies. A beautiful example of the application of Sérsic's formula to a large sample of galaxies is shown in Fig. 3.8a which shows the distribution of n among 10,095 galaxies selected from the Millennium Galaxy Catalogue (Driver et al., 2006). It can be seen that the galaxy sample splits very beautifully into two populations, one centred on the value $n = 4$, corresponding to the elliptical galaxies and the bulges of spiral galaxies and the other the value $n = 1$, corresponding to the light distribution of disc galaxies. In the analysis by Driver and his colleagues, the morphological categories were checked by visual inspection of the images of the

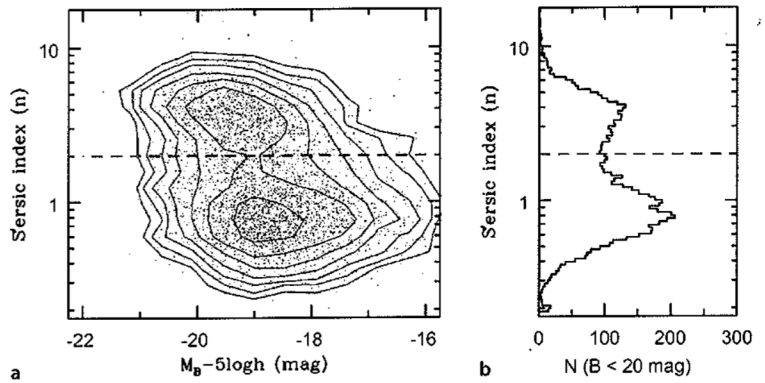


Fig. 3.8. a A plot of the observed value of the Sérsic index n as a function of the absolute blue magnitude in a sample of 10,095 galaxies from the Millennium Galaxy Catalogue. **b** The histogram showing the number of galaxies in equal logarithmic bins of Sérsic index n (Driver et al., 2006)

galaxies. The distinction between the bulge-dominated and disc-dominated systems occurs at about $n \approx 2$. As we will see, these structural parameters are strongly correlated with other properties of the galaxies, in particular, with the red and blue galaxy sequences described in Sect. 3.9.2. This procedure is particularly valuable in distinguishing the structural properties of large samples of galaxies by computer analyses.

3.5 The Masses of Galaxies

All direct methods of measuring masses in astronomy are dynamical. For systems such as star clusters, galaxies and clusters of galaxies, it can generally be assumed that they have reached some form of dynamical equilibrium and then, by measuring the velocities of the objects which make up the system and knowing its dimensions, mass estimates can be made. A key result for determining the masses of galaxies and clusters of galaxies is the *virial theorem*, first derived for star clusters by Eddington in 1916 (Eddington, 1916).

3.5.1 The Virial Theorem for Clusters of Stars, Galaxies and Clusters of Galaxies

Star clusters, galaxies and clusters of galaxies can generally be considered to be gravitationally bound configurations, meaning that the stars or galaxies have come into *dynamical equilibrium* under gravity. This assertion is supported by comparison of the *crossing time* of an object within the system with its age. The crossing time is

defined to be $t_{cr} = R/\langle v \rangle$ where R is the size of the system and $\langle v \rangle$ is the typical speed, or velocity dispersion, of the objects of which it is composed. For example, the orbital speed of the stars in our Galaxy at our distance from its centre, 8.5 kpc, is about 220 km s^{-1}. Therefore, the time it takes the stars to make one complete revolution about the centre of our Galaxy is $t = 2\pi R/v \approx 2.5 \times 10^8$ years. This is very much less than the age of the Galaxy, which is about 1.3×10^{10} years, and so the system must be gravitationally bound. Similarly, in the Coma cluster of galaxies, the crossing time is less than about one-tenth the age of the Universe, indicating that the cluster must be gravitationally bound, or else the galaxies would have dispersed long ago.

The *virial* was introduced by Rudolph Clausius in 1870 in connection with the thermal energy of gases. The virial was defined to be the quantity $\mathcal{Z}_i = -\frac{1}{2}\langle \mathbf{r}_i \cdot \mathbf{F}_i \rangle$ where the force $\mathbf{F}_i$ acts on the particle i located at position vector $\mathbf{r}_i$. The angle brackets represent the time average of the force acting on the particle and Clausius showed that $\mathcal{Z}_i$ is the system's average kinetic energy (see p. 105 in (Sparke and Gallagher, 2000)). In the astronomical context, the theorem refers to the energy balance in systems in equilibrium under gravity and it is found in a variety of different guises. In its application to the internal properties of stars, the virial theorem describes the relation between the thermal energy of the gas and its gravitational potential energy. The theorem can be extend to include rotational energy, magnetic energy, the energy in the form of convective motions or turbulence, and so on. In stellar dynamics, in which the 'gas' of stars may be taken to be collisionless, the *tensor virial theorem* relates the equilibrium state to the energies associated with the velocity distribution of the stars at each point, which will in general be anisotropic (Binney and Tremaine, 1987). In this section, we will only consider the simplest form of virial theorem for a self-gravitating system of point masses.

Suppose a system of particles (stars or galaxies), each of mass m_i, interact with each other only through their mutual forces of gravitational attraction. Then, the acceleration of the ith particle due to all other particles can be written vectorially

$$\ddot{\mathbf{r}}_i = \sum_{j \neq i} \frac{G m_j (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \,. \tag{3.7}$$

Now, take the scalar product of both sides with $m_i \mathbf{r}_i$.

$$m_i (\mathbf{r}_i \cdot \ddot{\mathbf{r}}_i) = \sum_{j \neq i} G m_i m_j \frac{\mathbf{r}_i \cdot (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \,. \tag{3.8}$$

Differentiating $(\mathbf{r}_i \cdot \mathbf{r}_i)$ with respect to time

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{r}_i \cdot \mathbf{r}_i) = 2\dot{\mathbf{r}}_i \cdot \mathbf{r}_i \,, \tag{3.9}$$

and then, taking the next derivative,

$$\tfrac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\mathbf{r}_i^2) = \frac{\mathrm{d}}{\mathrm{d}t}(\dot{\mathbf{r}}_i \cdot \mathbf{r}_i) = (\ddot{\mathbf{r}}_i \cdot \mathbf{r}_i + \dot{\mathbf{r}}_i \cdot \dot{\mathbf{r}}_i) = (\ddot{\mathbf{r}}_i \cdot \mathbf{r}_i + \dot{\mathbf{r}}_i^2) \,. \tag{3.10}$$

Therefore, (3.8) can be rewritten

$$\tfrac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}(m_i \mathbf{r}_i^2) - m_i \dot{\mathbf{r}}_i^2 = \sum_{j \neq i} G m_i m_j \frac{\mathbf{r}_i \cdot (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \,. \tag{3.11}$$

Now we sum over all the particles in the system,

$$\tfrac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sum_i m_i \mathbf{r}_i^2 - \sum_i m_i \dot{\mathbf{r}}_i^2 = \sum_i \sum_{j \neq i} G m_i m_j \frac{\mathbf{r}_i \cdot (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \,. \tag{3.12}$$

Now the double sum on the right-hand side represents the sum over all the elements of a square $n \times n$ matrix with all the diagonal terms zero. If we sum the elements ij and ji of the matrix, we find

$$G m_i m_j \left[\frac{\mathbf{r}_i \cdot (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} + \frac{\mathbf{r}_j \cdot (\mathbf{r}_i - \mathbf{r}_j)}{|\mathbf{r}_j - \mathbf{r}_i|^3} \right] = -\frac{G m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|} \,. \tag{3.13}$$

Therefore,

$$\tfrac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sum_i m_i \mathbf{r}_i^2 - \sum_i m_i \dot{\mathbf{r}}_i^2 = -\tfrac{1}{2}\sum_{\substack{j \neq i \\ i,j}} \frac{G m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|} \,, \tag{3.14}$$

where the factor $\frac{1}{2}$ on the right-hand side is included because the sum is still over all elements of the array and so the sum of each pair would be counted twice.

Now, $\sum_i m_i \dot{\mathbf{r}}_i^2$ is twice the total kinetic energy, T, of all the particles in the system, that is,

$$T = \tfrac{1}{2}\sum_i m_i \dot{\mathbf{r}}_i^2 \,. \tag{3.15}$$

The gravitational potential energy of the system is

$$U = -\tfrac{1}{2}\sum_{\substack{j \neq i \\ i,j}} \frac{G m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|} \,. \tag{3.16}$$

Therefore,

$$\tfrac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sum_i m_i \mathbf{r}_i^2 = 2T - |U| \,. \tag{3.17}$$

If the system is in statistical equilibrium

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sum_i m_i \mathbf{r}_i^2 = 0 \,, \tag{3.18}$$

and therefore

$$T = \tfrac{1}{2}|U| \,. \tag{3.19}$$

This is the equality known as the *virial theorem* in stellar dynamics.

Notice that, at no point, have we made any assumption about the orbits or velocity distributions of the particles. The velocities might be random, as is often assumed to be the case for globular clusters or spherical elliptical galaxies, but they might also have highly elongated orbits about the centre of the galaxy. In the case of the discs of spiral galaxies, the velocity vectors are highly ordered and the mean rotational speed about the centre is much greater than the random velocities of the stars. In all these cases, the virial theorem must hold if the system is to remain in dynamical equilibrium. In its simplest form, the expression (3.19) tells us nothing about the velocity distribution of the stars or galaxies within the system.

Despite the elegance of the theorem, its application to astronomical systems is not straightforward. In most cases, we can only measure directly radial velocities from the Doppler shifts of spectral lines and positions on the sky. In some cases, independent distance measures of the stars or galaxies within the system are available, but generally, within star clusters and clusters of galaxies, it is not possible to distinguish whether the objects are on the near or far side of the cluster. In some cases, the proper motions of the objects can be measured and then their three-dimensional space motions can be found. Generally, for clusters of galaxies, we need to make assumptions about the spatial and velocity distributions of galaxies in the cluster. For example, if we assume that the velocity distribution of the galaxies is isotropic, the same velocity dispersion is expected in the two perpendicular directions as along the line of sight and so $\langle v^2 \rangle = 3\langle v_\parallel^2 \rangle$, where $v_\parallel$ is the radial velocity. If the velocity dispersion is independent of the masses of the stars or galaxies, we can find the total kinetic energy $T = (1/2)\sum_i m_i \dot{\mathbf{r}}_i^2 = (3/2)M\langle v_\parallel^2 \rangle$, where M is the total mass of the system. If the velocity dispersion varies with mass, then $\langle v_\parallel^2 \rangle$ is a mass-weighted velocity dispersion. If the system is spherically symmetric, we can work out from the observed surface distribution of stars or galaxies a suitably weighted mean separation $R_{\rm cl}$, so that the gravitational potential energy can be written $|U| = GM^2/R_{\rm cl}$. Thus, the mass of the system can be found from the virial theorem

$$T = \tfrac{1}{2}|U| \qquad M = 3\langle v_\parallel^2 \rangle R_{\rm cl}/G. \tag{3.20}$$

Notice that, in general, we have to estimate some characteristic velocity, or velocity dispersion, and the size of the system in order to find its mass. This general result is widely applicable in astrophysics.

3.5.2 The Rotation Curves of Spiral Galaxies

In the case of spiral galaxies, masses can be estimated from their *rotation curves*, that is, the variation of the orbital, or rotational, speed $v_{\rm rot}(r)$ about the centre of the galaxy with distance r from its centre. Examples of the rotation curves of spiral

galaxies derived from optical and radio 21-cm line studies are shown Fig. 3.9b (Bosma, 1981). In a few galaxies, there is a well-defined maximum in the rotation curve and the velocity of rotation decreases monotonically with increasing distance from the centre. If this decrease continues to infinite distance, the total mass of the galaxy converges and is similar to that derived from the rotation curve in the central regions. In many cases, however, the rotational velocities in the outer regions of galaxies are remarkably constant with increasing distance from the centre. It is apparent from Fig. 3.9a that the flat rotation curve of our spiral neighbour M31 extends far beyond the optical image of the galaxy.

The significance of these flat rotation curves can be appreciated from application of Gauss's theorem to Newton's law of gravity. For simplicity, let us assume that the distribution of mass in the galaxy is spherically symmetric, so that we can write the mass within radius r as $M(\leq r)$. According to Gauss's law for gravity, for any spherically symmetric variation of mass with radius, we can find the radial acceleration at radius r by placing the mass within radius r, $M(\leq r)$, at the centre of the galaxy. Then, equating the centripetal acceleration at radius r to the gravitational acceleration, we find

$$\frac{GM(\leq r)}{r^2} = \frac{v_{\rm rot}^2(r)}{r} \qquad M(\leq r) = \frac{v_{\rm rot}^2(r) r}{G} \,. \tag{3.21}$$

For a point mass, say the Sun, $M(\leq r) = M_\odot$, and we recover Kepler's third law of planetary motion, the orbital period T being equal to $2\pi r/v_{\rm rot} \propto r^{3/2}$. This result can also be written $v_{\rm rot} \propto r^{-1/2}$ and is the variation of the circular rotational velocity expected in the outer regions of a galaxy if most of the mass is concentrated within the central regions.

If the rotation curve of the spiral galaxy is flat, $v_{\rm rot}$ = constant, $M(\leq r) \propto r$ and so the mass within radius r increases linearly with distance from the centre. This contrasts dramatically with the distribution of light in the discs, bulges and haloes of spiral galaxies which decrease exponentially with increasing distance from the centre

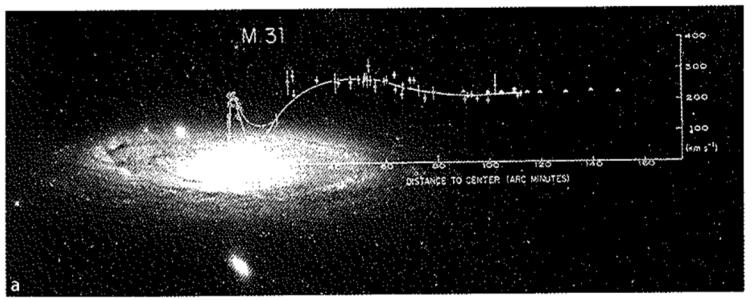
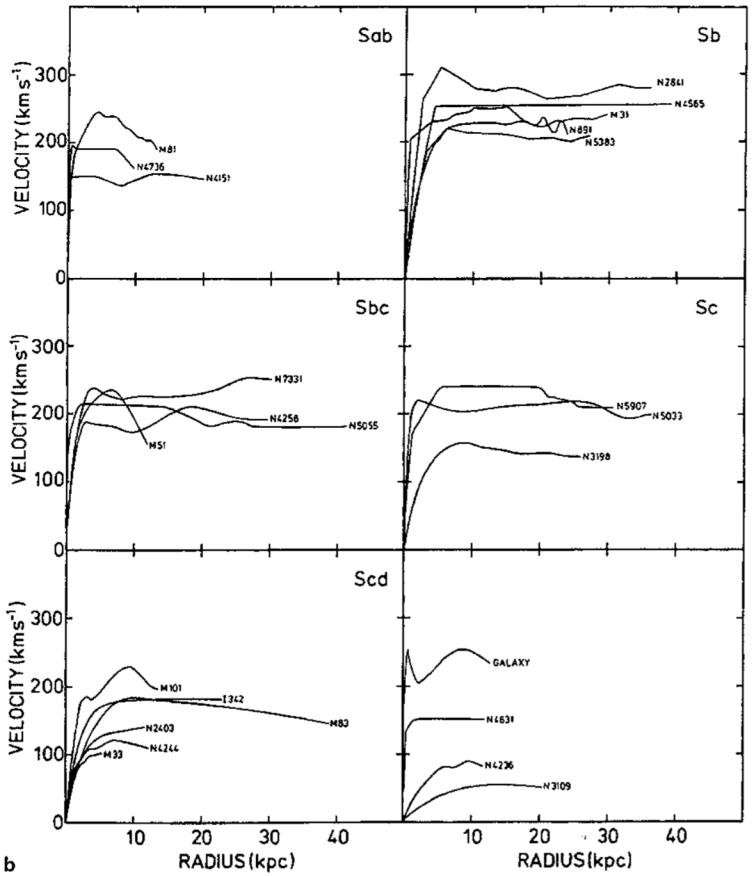


Fig. 3.9. a The rotation curve for the nearby giant spiral galaxy M31, showing the flat rotation curve extending well beyond the optical image of the galaxy (Courtesy of Dr. Vera Rubin)



b

Fig. 3.9. (continued) **b** Examples of the rotation curves of spiral galaxies from optical and neutral hydrogen observations (Bosma, 1981)

(see Sect. 3.4.2). Consequently, the local mass-to-luminosity ratio must increase in the outer regions of spiral galaxies.

It is most convenient to quote the results in terms of mass-to-luminosity ratios relative to that of the Sun. For the visible parts of spiral galaxies, for which the rotation curves are well-determined, mean mass-to-light ratios in the B waveband in the range 1–10 are found. This is similar to the value found in the solar neighbourhood; averaging over the masses and luminosities of the local stellar populations, a value of $M/L \approx 3$ is found. The M/L ratio must however increase to much larger values at large values of r. Values of $M/L \approx 10$–$20\ M_\odot/L_\odot$ are found in the outer regions of

spiral galaxies, similar to the values found for elliptical galaxies. These data provide crucial evidence for the presence of dark matter in galaxies.

There are theoretical reasons why spiral galaxies should possess dark haloes. Ostriker and Peebles showed that, without such a halo, a differentially rotating disc of stars is subject to a bar instability (Ostriker and Peebles, 1973). Their argument has been confirmed by subsequent computer simulations and suggests that dark haloes can stabilise the discs of spiral galaxies. We will return to the thorny question of the nature of the dark matter in Chap. 4.

3.5.3 The Velocity Dispersions of Elliptical Galaxies

Expression (3.20) can be used to estimate the masses of elliptical galaxies. Doppler broadening of the widths of stellar absorption lines in galaxies can be used to estimate the velocity dispersion $\langle \Delta v_\parallel^2 \rangle$ of stars along the line of sight through the galaxy. Typical mass-to-luminosity ratios for elliptical galaxies are about 10–$20\ M_\odot/L_\odot$. The trouble with this argument is that it has to be assumed that the velocity distribution of the stars in the elliptical galaxy is isotropic. As will be discussed in Sect. 3.6.3, there is compelling evidence that in general elliptical galaxies are triaxial systems and so the isotropy of the stellar velocity distribution needs to be tested directly by observation.

Evidence that there must indeed be considerable amounts of dark matter in the haloes about two of the giant elliptical galaxies in the Virgo cluster, M49 and M87, has been presented by Côté and his colleagues (Côté et al., 2001, 2003). They measured the radial velocities of a large sample of globular clusters in the haloes of these galaxies and so were able to extend the range of radii over which the velocity dispersion in these galaxies could be measured. A beautiful example of the quality of their data for M49 is shown in Fig. 3.10. Their measurements are shown by the filled circles at radii $R \geq 10$ kpc, the dotted and solid lines bracketing them showing the one and two sigma ranges of their estimates of the velocity dispersion. The points at radii less than 10 kpc show the velocity dispersion measured by other authors and it can be seen that the data are consistent with the velocity dispersion remaining remarkably constant out to radii up to 40 kpc from the centre.

Various attempts to account for the variation of the velocity dispersion with radius are indicated by the different lines on the diagram. These assume that the mass distribution follows the radial optical intensity distribution, but with various extreme assumptions about the anisotropy of the stellar velocity distribution. Even models in which the stars (or globular clusters) are on radial orbits cannot account for the fact that the line-of-sight velocity dispersion is independent of radius out to 40 kpc. Côté and his colleagues conclude that these data provide evidence that the velocity dispersion is isotropic and that there must be dark matter haloes about these galaxies. Similar conclusions can be drawn from X-ray observations of these galaxies, using the technique described in the context of clusters of galaxies in Sect. 4.4.

Physically, the fact that the velocity dispersion remains constant out to large radii has exactly the same explanation as the flatness of the rotation curves of spiral galaxies (see expression 3.21). To bind globular clusters to these massive galaxies

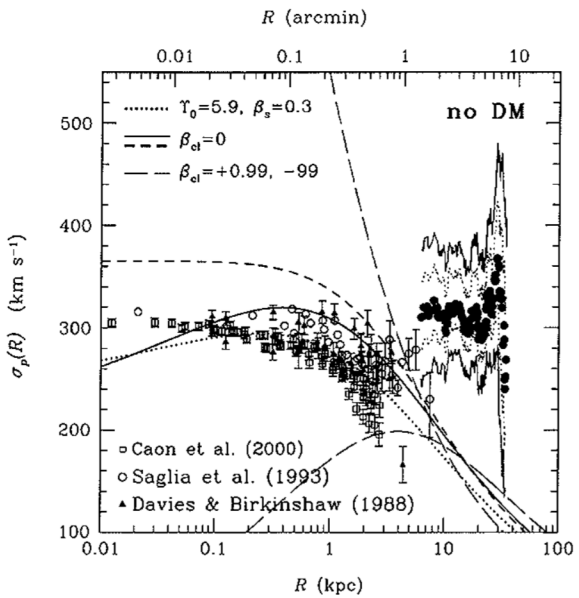


Fig. 3.10. The velocity dispersion of stars and globular clusters in the nearby giant elliptical galaxy M49 (NGC 4472). The data points at $R < 10$ kpc are obtained from the velocity width of the stellar absorption lines. The *filled circles* at radii $R > 10$ kpc are derived from the velocity dispersion of globular clusters. Various models for the velocity dispersion as a function of radius R, assuming that the mass follows the light are shown (Côté et al., 2003)

with these large velocity dispersions at large radii means that the mass within radius R must increase proportional to R.

3.6 The Properties of Spiral and Elliptical Galaxies

At first glance, it would seem that the elliptical galaxies should be among the simpler stellar systems to interpret theoretically because they can be approximated as single spheroidal stellar distributions. This turns out to be an overoptimistic expectation.

3.6.1 The Faber–Jackson Relation and the Fundamental Plane

Extensive studies have been made of correlations between various properties of elliptical galaxies, specifically, their luminosities, their sizes, as described by the de Vaucouleurs radius r_e, their central velocity dispersions, their surface brightnesses, the abundance of heavy elements, and so on. Of these, two studies are of particular

importance. The first is the analysis of Faber and Jackson who found a strong correlation between luminosity L and central velocity dispersion σ of the form $L \propto \sigma^x$ where $x \approx 4$ (Faber and Jackson, 1976). This correlation has been studied by other authors who have found values of x ranging from about 3 to 5. The significance of this relation is that, if the velocity dispersion σ is measured for an elliptical galaxy, its intrinsic luminosity can be found from the Faber–Jackson relation and hence, by measuring its observed flux density, its distance can be found.

This procedure for measuring distances was refined by Dressler and his colleagues and by Djorgovski and Davis who introduced the concept of the *fundamental plane* for elliptical galaxies (Dressler et al., 1987; Djorgovski and Davis, 1987). The fundamental plane lies in a three-dimensional space in which luminosity L is plotted against the central velocity dispersion σ and the mean surface brightness Σ_e within the half-light radius r_e, that is, $\Sigma_e = L(\leq r_e)/\pi r_e^2$. Dressler and his colleagues found an even stronger correlation than the Faber–Jackson relation when the surface brightness was included,

$$L \propto \sigma^{8/3}\, \Sigma_e^{-3/5}\,. \tag{3.22}$$

Various expressions for the fundamental plane appear in the literature, for example

$$r_e \propto \sigma^{1.4} I_e^{-0.9} \tag{3.23}$$

which is remarkably similar to (3.22).

Dressler and his colleagues found just as good a correlation if they introduced a new diameter D_n, which was defined as the circular diameter within which the total mean surface brightness of the galaxy exceeded a particular value. The surface brightness was chosen to be 20.75 B magnitudes arcsec^{-2}. The correlation found was $\sigma \propto D_n^{3/4}$, thus incorporating the dependence of both L and Σ_e into the new variable D_n.

The origin of these empirical correlations is not understood. The argument can be inverted to determine under what conditions relations such as the Faber–Jackson relation would be found. For example, since the mass of the galaxy is given by $M \propto \sigma^2 r_e$ and $L \propto I_e r_e^2$, it follows that $L \propto \sigma^4/I_e(M/L)^2$. Thus, if I_e and M/L were constant for all elliptical galaxies, we would obtain $L \propto \sigma^4$. It is not at all clear, however, why I_e and M/L should be constant for elliptical galaxies.

Despite the lack of theoretical underpinning of these correlations, Dressler and his colleagues estimate that they enable the distances of individual galaxies to be determined to about 25% and for clusters of galaxies to about 10%.

3.6.2 Ellipticals Galaxies as Triaxial Systems

It might be thought that the internal dynamics of elliptical galaxies would be relatively straightforward. Their surface brightness distributions appear to be ellipsoidal, the ratio of the major to minor axes ranging from 1:1 to about 3:1. It is natural to attribute the flattening of the elliptical galaxies to the rotation of these stellar systems and this

can be tested by observations of the mean velocities and velocity dispersions of the stars throughout the body of the galaxy. These measurements can be compared with the rotation and internal velocity dispersions expected if the flattening of the elliptical galaxies were wholly attributed to the rotation of an axisymmetric distribution of stars. In the simplest picture, it is assumed that the velocity distribution is isotropic at each point within the galaxy.

Bertola and Capaccioli in 1975 and Illingworth in 1977 first showed that elliptical galaxies rotate too slowly for centrifugal forces to be the cause of their observed flattening; in other words, the ratio of rotational to random kinetic energy is too small (Bertola and Capaccioli, 1975; Illingworth, 1977). This analysis was repeated in 1983 for a larger sample of elliptical galaxies and for the bulges of spiral galaxies by Davies and his colleagues with the results shown in Fig. 3.11 (Davies et al., 1983). The solid lines show the amount of rotation v_m necessary to account for the observed ellipticity of the elliptical galaxy relative to the velocity dispersion σ of the stars. It can be seen that, for low luminosity elliptical galaxies and for the bulges of spiral galaxies, the ellipticity of the stellar distribution can be attributed to rotation. The most luminous ellipticals with $M_B < -20.5$ generally do not possess enough rotation to account for the observed flattening of the galaxies. This means that the assumptions of an axisymmetric spatial distribution and/or an isotropic velocity distribution of stars at all points within the galaxy must be wrong. As a consequence, these massive elliptical galaxies must be *triaxial* systems, that is, systems with three unequal axes and consequently with anisotropic stellar velocity distributions. There is no reason why the velocity distribution should be isotropic because the time-scale for the exchange of energy between stars through gravitational encounters is generally greater than the age of the galaxy. Therefore, if the velocity distribution began by being anisotropic, it would not have been isotropised by now.

Further evidence for the triaxial nature of massive elliptical galaxies has come from studies of their light distributions. In many systems not only does the ellipticity of the isophotes of the surface brightness distribution vary with radius, but also the position angle of the major axis of the isophotes can change as well. All types of variation of ellipticity with radius are known. In some cases there is a monotonic change with radius but in others there can be maxima and minima in the radial variation of the ellipticity (Bertola and Galletta, 1979). The dynamics of such galaxies must be much more complicated than those of a rotating isothermal gas sphere. Another piece of evidence for the complexity of the shapes and velocity distributions within elliptical galaxies comes from the observation that, in some ellipticals, rotation takes place along the minor as well as along the major axis (Bertola et al., 1991). Thus, despite their simple appearances, some elliptical galaxies may be triaxial systems.

The theoretical position has been clarified by an elegant and original analysis by Martin Schwarzschild (Schwarzschild, 1979). By applying linear programming techniques to the determination of orbits in general self-gravitating systems, he showed that there exist stable triaxial configurations not dissimilar from those necessary to explain some of the internal dynamical properties of what appear on the surface to be simple ellipsoidal stellar distributions. His analysis showed that there exist stable orbits about the major and minor axes but not about the immediate axis of the

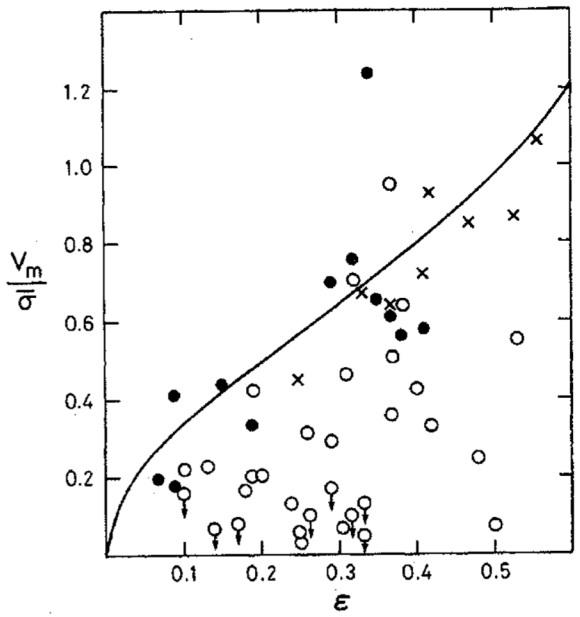


Fig. 3.11. A diagram showing the flattening of elliptical galaxies as a function of their rotational velocities. The *open circles* are luminous elliptical galaxies, the *filled circles* are lower luminosity ellipticals and the *crosses* are the bulges of spiral galaxies. If the ellipticity were entirely due to rotation with an isotropic stellar velocity distribution at each point, the galaxies would be expected to lie along the *solid line*. This diagram shows that, at least for massive ellipticals, this simple picture of rotational flattening cannot be correct (Davies et al., 1983)

triaxial figure. With this new understanding of the stellar motions in elliptical galaxies, galaxies can be characterised as oblate-axisymmetric, prolate-axisymmetric, oblate-triaxial, prolate-triaxial and so on.

3.6.3 The Tully–Fisher Relation for Spiral Galaxies

The masses of spiral galaxies can be estimated from their rotation curves as described in Sect. 3.5.2. In 1975, Tully and Fisher discovered that, for spiral galaxies, the widths of the profiles of the 21-cm line of neutral hydrogen, once corrected for the effects of inclination, are strongly correlated with their intrinsic luminosities (Tully and Fisher, 1977). In their studies, they correlated the total B luminosities with the corrected velocity width ΔV of the 21-cm line and found the relation

$$L_B \propto \Delta V^{\alpha} , \tag{3.24}$$

where $\alpha = 2.5$. A much larger survey carried out by Aaronson and Mould found a somewhat steeper slope, $\alpha = 3.5$, for luminosities measured in the optical B waveband and an even steeper slope, $\alpha = 4.3$ in the near-infrared H waveband at 1.65 μm (Aaronson and Mould, 1983). The correlation was found to be much tighter in the infrared as compared with the blue waveband, because the luminosities of spiral galaxies in the blue waveband are significantly influenced by interstellar extinction within the galaxies themselves, whereas, in the infrared waveband the dust becomes transparent. What has come to be called the *infrared Tully–Fisher relation* is very tight indeed. As a result, measurement of the 21-cm velocity width of a spiral galaxy can be used to infer its absolute H magnitude and hence, by measuring its flux density in the H waveband, its distance can be estimated. This procedure has resulted in some of the best distance estimates for spiral galaxies and has been used in programmes to measure the value of Hubble's constant.

There is an interesting interpretation of the Tully–Fisher relation for exponential discs. Suppose that the mass distribution follows the same distribution as the optical surface brightness with radius, $I = I_0 \exp(-r/h)$. Then, the total mass of the disc is

$$M = \int_0^\infty 2\pi r I_0 \mathrm{e}^{-r/h}\,\mathrm{d}r = 2\pi I_0 h^2 \int_0^\infty x\mathrm{e}^{-x}\,\mathrm{d}x = 2\pi I_0 h^2 \,. \tag{3.25}$$

Thus, most of the mass of the disc lies within radius $r \sim h$. The maximum of the rotation curve therefore corresponds roughly to the Keplerian velocity at distance h from the centre. Placing all the mass at the centre of the disc and equating the centripetal and gravitational accelerations, the maximum of the rotation curve is expected to correspond to $V_{\max}$ where

$$\frac{V_{\max}^2}{h} \approx \frac{2\pi G I_0 h^2}{h^2}\,; \qquad V_{\max} \propto (I_0 h)^{1/2} \,. \tag{3.26}$$

Eliminating h from (3.25) and (3.26), we find that $M \propto V_{\max}^4$. If we now adopt Freeman's result that the central surface brightnesses of bright spiral galaxies have a roughly constant value and assume that the mass-to-luminosity ratio is constant within the discs of spiral galaxies, we expect $L \propto V_{\max}^4$, roughly the observed Tully–Fisher relation.

3.6.4 Luminosity–Metallicity Relations

An important aspect of the physics of galaxies which will play an important role in understanding their formation is the relation between their luminosities, masses, colours and the abundances of the heavy elements, the last being referred to as their *metallicities*. The observations are well-understood, but their interpretation is subject to many caveats and uncertainties.

For elliptical galaxies, Faber showed in 1973 showed that there is a correlation between their luminosities and the strength of the magnesium absorption lines (Faber, 1973). In subsequent analyses, a similar relation was established over a wide range of luminosities and between the central velocity dispersion of the elliptical galaxy

and the strength of the Mg_2 index (Bender et al., 1993). They also showed that the Mg_2 index was strongly correlated with the (B – V) colours of the bulges of these galaxies and so the correlation referred to the properties of the galaxy as a whole. The result was a strong correlation between luminosity, colour and metallicity spanning almost a factor of 1000 in luminosity (Fig. 3.12).

A similar relation was found by Visvanathan and Sandage for elliptical galaxies in groups and clusters of galaxies in the sense that the more luminous the galaxy, the redder they were observed to be (Visvanathan and Sandage, 1977). Their prime interest was in using this correlation in groups and clusters of galaxies to estimate their distances, but the sense of the correlation is the same as that found by Faber and her colleagues since galaxies with greater metallicities have greater line blanketing in the blue and ultraviolet regions of the spectrum and hence are redder than their lower metallicity counterparts.

A similar correlation was first established for late-type and star forming galaxies by Lequeux and his colleagues (Lequeux et al., 1979). These pioneering studies involved determining the gas-phase metallicities of the galaxies and were followed by a number of studies which extended the luminosity–metallicity correlation to a range of 11 magnitudes in absolute luminosity and a factor of 100 in metallicity (Zaritsky et al., 1994). These studies laid the foundation for the analyses of the huge databases of galaxies available from the Sloan Digital Sky Survey.

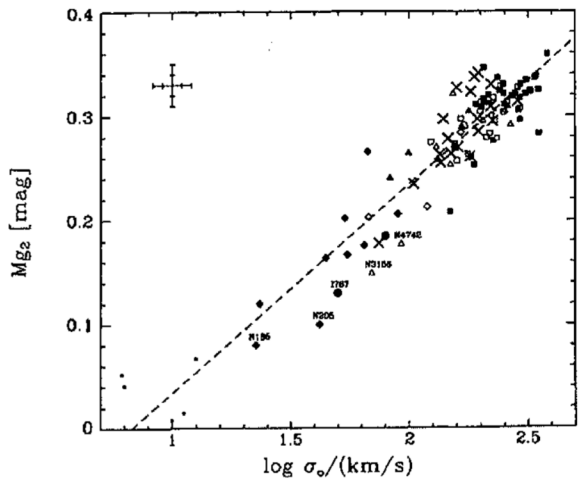


Fig. 3.12. The correlation between the luminosity of the bulges or spheroids of elliptical galaxies, what Faber and her colleagues refer to as 'Dynamically Hot Galaxies', and the strength of their absorption lines due to heavy elements, as measured by the Mg_2 index (Bender et al., 1993)

In the analysis of Tremonti and her colleagues, rather than using luminosity, they work directly with the stellar mass of the galaxy (Tremonti et al., 2004). This approach has become feasible thanks to the development of efficient and reliable codes for determining the stellar and gaseous masses of galaxies from their optical spectra (Bruzual and Charlot, 2003; Charlot and Longhetti, 2001). It turns out that the correlation with stellar mass is stronger than that with luminosity. Figure 3.13 shows the strong correlation between metallicity and the total stellar mass of star-forming galaxies. These observations provide important constraints on the physics of the evolution of galaxies. With the advent of 8–10-metre class telescopes, these studies have been extended to samples of galaxies at large redshifts and so constrain directly the evolution of the stellar and gaseous content of galaxies of different masses (Savaglio et al., 2005). These topics will be taken up in much more detail in Chaps. 17 to 19.

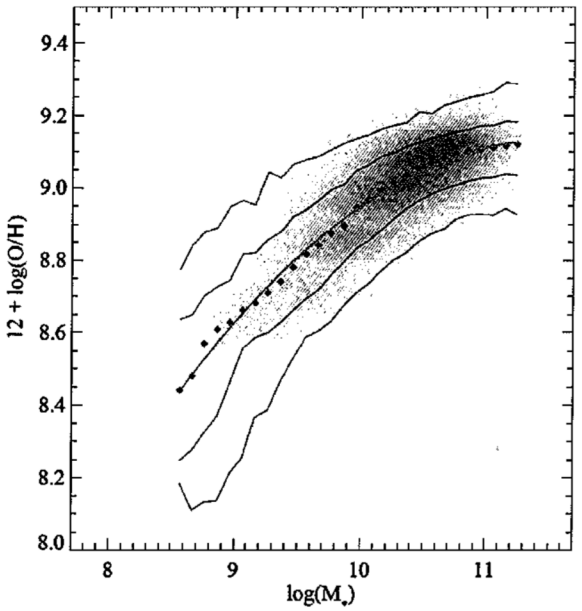


Fig. 3.13. The stellar mass–gas phase metallicity relation for 53,400 star-forming galaxies from the SDSS. The large *black points* represent the median in bins of 0.1 dex in mass which include at least 100 data points. The *thin line* through the data is a best-fitting smooth curve and the *solid lines* are the contours which enclose 68% and 95% of the data (Tremonti et al., 2004)

3.7 The Luminosity Function of Galaxies

The frequency with which galaxies of different intrinsic luminosities are found in space is described by the *luminosity function* of galaxies. The luminosity function of galaxies $\phi(L)\,\mathrm{d}L$ is defined to be the space density of galaxies with intrinsic luminosities in the range L to $L + \mathrm{d}L$. If S is the flux density (in $\mathrm{W\,m^{-1}\,Hz^{-1}}$) of a nearby galaxy, for which redshift corrections can be neglected, the luminosity of the galaxy is $L = 4\pi r^2 S$ (in $\mathrm{W\,Hz^{-1}}$), where r is the distance of the galaxy. In optical astronomy, it is traditional to work in terms of absolute magnitudes, M, rather than luminosities and so, in terms of absolute magnitudes, the luminosity function $\phi(L)\,\mathrm{d}L = \Phi(M)\,\mathrm{d}M$. The important difference between these two forms of the luminosity function is that, in terms of magnitudes, the luminosity function is presented on a logarithmic scale of luminosity. The absolute magnitude M and the luminosity L are related by the expression

$$\log\left(\frac{L}{L^*}\right) = -0.4\,(M - M^*)\,. \tag{3.27}$$

where the absolute magnitude M^* and the luminosity L^* are corresponding reference values of these quantities.

In 1977, Felten made a careful comparison of nine different determinations of the local luminosity function for nearby galaxies, reducing them all to the same value of Hubble's constant, the same magnitude system and the same corrections for Galactic extinction. In this heroic analysis, he found that the independent determinations were in remarkably good agreement (Felten, 1977). Felten's analysis is summarised in Fig. 3.14, using reduced absolute magnitudes, $M_{\mathrm{B}_{\mathrm{T}}^0}$ in de Vaucouleurs' $\mathrm{B}_{\mathrm{T}}^0$ magnitude system and using a Galactic extinction law $A_{\mathrm{B}} = 0.25\,\mathrm{cosec}\,|b|$. The solid line shows a best-fit to the data of the form of luminosity function proposed by Schechter

$$\phi(x)\,\mathrm{d}x = \phi^* x^{\alpha} \mathrm{e}^{-x}\,\mathrm{d}x\,, \tag{3.28}$$

or,

$$\phi(L)\,\mathrm{d}L = \phi^*\left(\frac{L}{L^*}\right)^{\alpha} \exp\left(-\frac{L}{L^*}\right)\frac{\mathrm{d}L}{L^*}\,, \tag{3.29}$$

where $x = L/L^*$ and L^* is the luminosity which characterises the 'break' in the luminosity function seen in Fig. 3.14 (Schechter, 1976). The form of the Schechter luminosity function is as simple as it could be: a power law with a high luminosity exponential cut-off. Its shape is characterised by two parameters, the slope of the power law α at low luminosities and the 'break' luminosity L^*.

It is traditional in optical astronomy to write the luminosity function in terms of astronomical magnitudes rather than luminosities and then the beautiful simplicity of the Schechter function is somewhat spoiled:

$$\begin{aligned}\Phi(M)\,\mathrm{d}M = {} & \tfrac{2}{5}\phi^* \ln 10\,\{\mathrm{dex}[0.4(M^* - M)]\}^{\alpha+1} \\ & \times \exp\{-\mathrm{dex}[0.4(M^* - M)]\}\,\mathrm{d}M\,,\end{aligned} \tag{3.30}$$

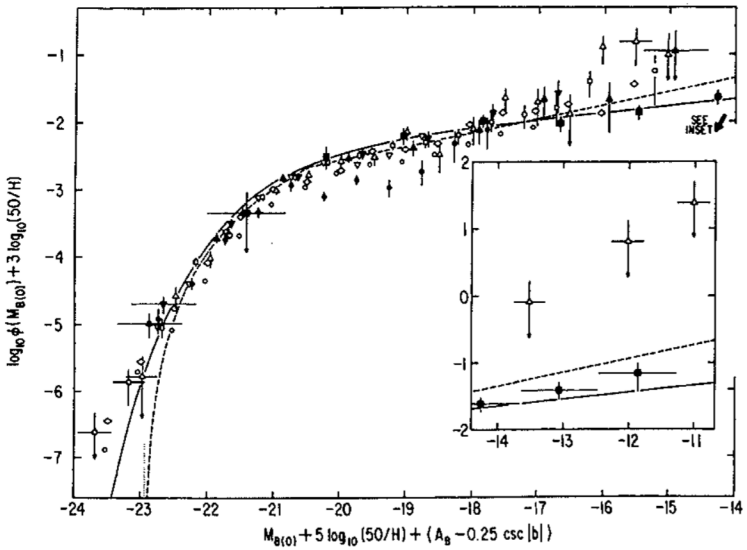
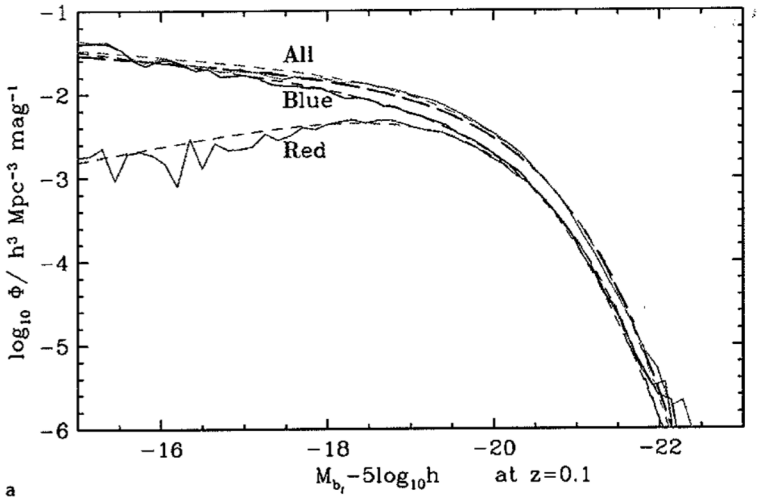


Fig. 3.14. The local luminosity function of galaxies from nine independent estimates considered by Felten fitted by a Schechter luminosity function of the form $n(x)\,\mathrm{d}x \propto x^{\alpha}\mathrm{e}^{-x}\,\mathrm{d}x$, where $x = L/L^*$ (Felten, 1977)

where M^* is the absolute magnitude corresponding to the luminosity L^*. We have used the notation dex y to mean 10^y. In his reassessment of the luminosity function for galaxies in 1985, Felten preferred the following best-fit values: $\alpha = -1.25$ and $M_{B_T^0} = -20.05 + 5\log_{10}h$ (Felten, 1985).

The normalisation factor ϕ^* determines the space density of galaxies and allowance has to be made for the fact that the galaxies used in the determination mostly lie within the local supercluster. Hence, the value of ϕ^* is an overestimate as compared with what would be found for a sample of field galaxies. Felten's preferred value of ϕ^* for the general field was $1.20 \times 10^{-2}h^3$ Mpc^{-3}.

These pioneering efforts by Felten were followed by careful studies of larger and larger samples of galaxies (see, for example, the review by Binggeli, Sandage and Tammann (Binggeli et al., 1988)), culminating in the analyses of very large samples of galaxies observed in the 2dF and Sloan Digital Sky Survey (SDSS) galaxy surveys . These very large surveys sample such large volumes of the Universe that the problems of correcting for the presence of the local supercluster are not relevant. Recent determinations of the luminosity function of galaxies from these surveys are shown in Fig. 3.15. The 2dF galaxy survey included 221,414 galaxies for all of which spectroscopic redshifts and colours were available (Fig. 3.15a). The overall luminosity function, as well as the functions for red and blue galaxies are shown on



a

Fig. 3.15. a The luminosity function of galaxies derived from a sample of 221,414 galaxies observed in the 2dF galaxy survey. The overall luminosity function and those of the red and blue galaxies in the sample have been fitted by Schechter luminosity functions (Cole et al., 2005)

Table 3.4. Parameters describing the overall luminosity function of galaxies from the 2dF and SDSS surveys. The functions are determined at a redshift of 0.1 and include K-corrections and evolutionary corrections for the observed change in form of the luminosity functions over the redshift interval $0 < z < 0.3$

Galaxy survey	Waveband	ϕ^*/h^3 Mpc^{-3}	$M^* - 5\log_{10}h$	α
2dF galaxy survey	b_J	0.0156	-19.52	-1.18
SDSS galaxy survey	r	0.0149 ± 0.0004	-20.44 ± 0.01	-1.05 ± 0.01

the diagram. In the case of the SDSS survey, redshifts were determined for 147,986 galaxies (Fig. 3.15b) (Blanton et al., 2003). The best-fit parameters describing the overall luminosity function for these two large surveys are listed in Table 3.4. It can be seen that the form of these functions are in good agreement.

3.7.1 Aspects of the Luminosity Function of Galaxies

A number of features of the luminosity function of galaxies should be noted.

Dependence upon galactic environment. With the availability of large unbiased samples of galaxies, it is possible to determine the luminosity function for galaxies of different morphological types in different environments, such as clusters, groups and void regions. The evidence of Fig. 3.15a shows that there is a clear difference in

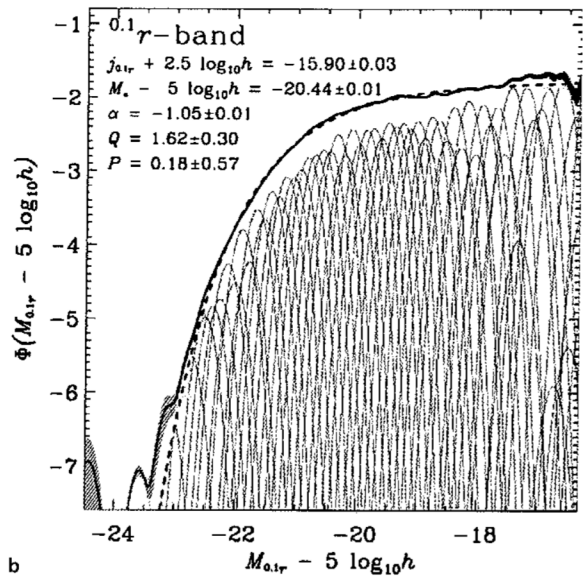


Fig. 3.15. (continued) **b** The luminosity function of galaxies derived from a sample of 147,986 galaxies observed in the Sloan Digital Sky Survey. The magnitudes are measured in the r^* waveband as observed at a redshift $z = 0.1$. The observations are very well-fitted by a Schechter luminosity function with the parameters given on the diagram and in Table 3.4 (Blanton et al., 2003)

the luminosity functions of red and blue selected galaxies, corresponding to early and late-type galaxies respectively in the Hubble sequence. In addition, there is evidence that the luminosity function of galaxies in rich clusters differs from that of galaxies in underdense regions of the Universe, the void regions. Using data from the 2dF galaxy survey, Croton and his colleagues showed clear differences in the forms of the luminosity functions for early and late-type galaxies as a function of the over or underdensity of the region relative to the mean density of galaxies (Fig. 3.16) (Croton et al., 2005). The population in the voids is dominated by late-type galaxies and shows, relative to the mean, a deficit of early-type galaxies that becomes increasingly pronounced at magnitudes fainter than $M_{\rm bj} - 5\log_{10}h = -18.5$. In contrast, clusters show a relative excess of very bright early-type galaxies with $M_{\rm bj} - 5\log_{10}h < -19$.

These facts combined with the differences in the relative numbers of galaxies of different morphological types as a function of galaxy density indicate that the approximation of a universal luminosity function for all galaxies wherever they are found in the Universe is, at best, a rough approximation.

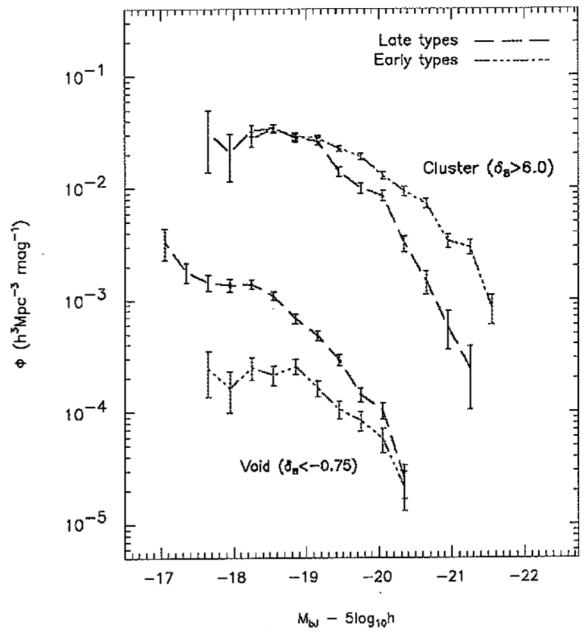


Fig. 3.16. The luminosity functions for early and late-type galaxies in rich clusters of galaxies and in large voids (Croton et al., 2005). As they express this result, the void population is composed almost exclusively of faint late-type galaxies, while in the clusters the galaxy population brighter than $M_{\rm bj} - 5\log_{10}h = -19$ consists predominantly of early types

Is L^* a standard candle? In 1962, Abell suggested that the luminosity of the break in the luminosity function of rich clusters L^*, corresponding to M^*, could be used as a 'standard candle' in the redshift–apparent magnitude relation (Abell, 1962). He found excellent agreement with the expected slope of the redshift–magnitude relation using this technique. Subsequent studies of the luminosity functions of individual clusters of galaxies have show that they are similar in form to the standard Schechter function with more or less the same parameters as those described above. Schechter found that, if only those clusters for which good fits to his proposed function were included, the dispersion in the absolute magnitude of M^* was only 0.25 magnitudes, as good a result as has been obtained from studies of the brightest galaxies in clusters (Schechter, 1976).

With the availability of the large surveys of galaxies, this proposal has to be treated with some caution since there is evidence for the evolution of the form the luminosity function, even over remarkably small redshift intervals. As shown in

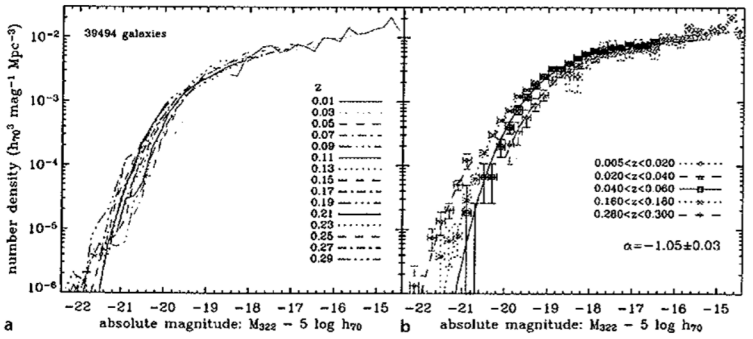


Fig. 3.17. The evolution of the luminosity function of galaxies at small redshifts. Over the redshift interval 0 to 0.3, the value of M^* becomes more positive by (0.8 ± 0.1) magnitudes, that is, the break in the luminosity function moves to fainter intrinsic luminosities as the Universe grows older (Baldry et al., 2005)

Fig. 3.17, the luminosity function as observed in the u-waveband shows significant changes over the small redshift interval $0.3 > z > 0.02$ (Baldry et al., 2005). Specifically, they find that M^* decreases by 0.8 ± 0.1 magnitudes between redshifts 0.3 and zero.

The brightest galaxies in clusters. At the very highest luminosities, the brightest galaxies in clusters do not fit smoothly onto an extrapolation of the Schechter luminosity function. These massive galaxies, a splendid example of which can be seen in Fig. 4.1, are known as *cD galaxies*, their characteristic being that they are similar to giant elliptical galaxies but in addition possess extensive stellar envelopes. They are the most luminous galaxies found in rich clusters and groups of galaxies. It appears that these are very special galaxies and not just the most luminous members of the luminosity function of galaxies, as was shown statistically by Tremaine and Richstone (Tremaine and Richstone, 1977). Evidently, there must be some physical reason why the first ranked cluster galaxies have these unique properties; we will return to this issue in Chap. 4.

The luminosity function for low luminosity galaxies. The luminosity function is quite poorly known at low luminosities, because these galaxies can only be observed in relatively nearby groups and clusters. According to Binggeli and his colleagues, the lowest luminosity regions of the luminosity function are exclusively associated with irregular and dwarf elliptical galaxies (Binggeli et al., 1988). These conclusions are confirmed by analyses of the SDSS, in particular, the analysis of a large sample of galaxies with distances in the range $10 < r < 150$ Mpc so that galaxies as faint as $M = -12.5$ can be included (Blanton et al., 2005). These data show an upturn in the slope of the luminosity function at very low luminosities, the best-fitting value of α being about -1.3. As the authors comment, however, a large number of galaxies at very low luminosities may be missing because of their low surface brightnesses

and so the true low luminosity slope may be -1.5 or even steeper. In agreement with Bingelli and his colleagues, they find that extremely low luminosity galaxies are predominantly blue, low surface brightness, exponential disks.

3.7.2 The Integrated Luminosity and the Mean Mass-to-Luminosity Ratio for Visible Matter in the Universe

An important calculation is the integrated luminosity of all the galaxies within a given volume of space. For a cluster of galaxies, the result would be the integrated optical luminosity of the cluster; if this were a typical unit volume of space, the result would be the luminosity density of the radiation due to all the galaxies in the Universe. Although the number of galaxies in the luminosity function diverges at low luminosities, the total background light remains finite. The luminosity density is

$$\begin{aligned} \varepsilon_{B(0)} &= \int_0^\infty L\,\phi(L)\,\mathrm{d}L = \phi^* L^* \int_0^\infty x^{a+1} \mathrm{e}^{-x}\,\mathrm{d}x \\ &= \phi^* L^* \Gamma(a+2)\,, \end{aligned} \tag{3.31}$$

where Γ is the gamma-function. For a cluster of galaxies, ϕ^* is the normalisation factor in the luminosity function. To estimate the luminosity density of a typical volume of space, we can use the values determined by Felten for the field luminosity function quoted above, $a = -1.25$, $\phi^* = 1.2 \times 10^{-3} h^3$ Mpc^{-3} and $M^* = -20.05 + 5 \log_{10} h$, corresponding to $1.24 \times 10^{10} h^{-2}\, L_\odot$. Then,

$$\varepsilon_{B(0)} = 1.8 \times 10^8 h L_\odot \,\mathrm{Mpc}^{-3}\,. \tag{3.32}$$

The value found from the SDSS luminosity function (Blanton et al., 2003) in the $^{0.1}r$ waveband is

$$(1.84 \pm 0.04) \times 10^8 h L_\odot \,\mathrm{Mpc}^{-3}\,. \tag{3.33}$$

These results are consistent with other estimates of the luminosity density, for example from the 2dF Galaxy Redshift Survey (Fig. 3.15a) and the Millennium Galaxy Catalogue.

A useful reference value for cosmological studies is the average mass-to-luminosity ratio for the Universe, if it is assumed to have the critical cosmological density, $\varrho_c = 3H_0^2/8\pi G = 2.0 \times 10^{-26} h^2$ kg m^{-3}. In terms of solar units, the mass-to-luminosity ratio would be

$$\frac{\varrho_c}{\varepsilon_B} = \left(\frac{M}{L}\right)_B = 1600\, h \left(\frac{M_\odot}{L_\odot}\right)_B\,. \tag{3.34}$$

Although there is some variation about this estimate, its importance lies in the fact that it is significantly greater than the typical mass-to-luminosity ratios of galaxies and clusters of galaxies, even when account is taken of the dark matter which must

be present. This result indicates that the mass present in galaxies and clusters of galaxies is not sufficient to close the Universe.

It is useful to work out typical values for the mean space density and luminosity of galaxies. Using the mean luminosity of galaxies for Felten's best estimate of the luminosity function with $a = -1.25$, we find $\langle L \rangle = 1.25L^* = 1.55 \times 10^{10} h^{-2}\, L_\odot$. Adopting the mean luminosity density of the Universe given by (3.32), the typical number density of galaxies $\bar{n} = \varepsilon_{B(0)}/\langle L \rangle = 10^{-2} h^3$ Mpc^{-3}. In other words, the typical galaxies which contribute most of the integrated light of galaxies are separated by a distance of about $5h^{-1}$ Mpc, if they were uniformly distributed in space, which we know to be very far from the truth. For reference, galaxies such as our own and M31 have luminosities $L_{Gal}(B) \approx 10^{10} L_\odot$. Evidently, if the ratio of mass-to-luminosity were the same for all galaxies, the 'mean' galaxies would also contribute most of the visible mass in the Universe.

These data also enable limits to be placed upon the average mass density in stars at the present epoch. In the simplest estimate, we can adopt a typical mass-to-luminosity ratio for the *visible* parts of galaxies of $M/L \approx 3$ and then the density parameter in stars at the present epoch would be $\Omega_* h = 2 \times 10^{-3}$. A very much more careful analysis has been carried out by Bell and his collaborators who used the combined SDSS and *Two Micron All Sky Survey* (2MASS) catalogues of galaxies (Bell et al., 2003). The benefit of including the 2MASS data is that the luminosity functions can be determined in the relatively unobscured 2 μm waveband. Their upper limit to the stellar mass density in the local Universe is

$$\Omega_* h = (2 \pm 0.6) \times 10^{-3}\,, \tag{3.35}$$

assuming the initial mass function of stars is as rich in low mass stars as is allowed by galaxy dynamics in the local Universe. This is a key result for many aspects of galaxy formation.

3.8 The Properties of Galaxies: Correlations Along the Hubble Sequence

What gives the Hubble classification physical significance is the fact that a number of physical properties are correlated with position along the sequence. Many of these were reviewed by Roberts and Haynes in an important analysis of the properties of a large sample of bright galaxies selected primarily from the Third Reference Catalogue of Bright Galaxies (de Vaucouleurs et al., 1991; Roberts and Haynes, 1994). They emphasised that, although there are clear trends, there is a wide dispersion about these correlations at any point along the sequence (Fig. 3.18).

Some of the more important findings of Roberts and Haynes' survey are as follows:

- *Total masses and luminosities.* The average masses and range of masses are roughly constant for galaxies in classes S0 to Scd. At later stages beyond Scd,

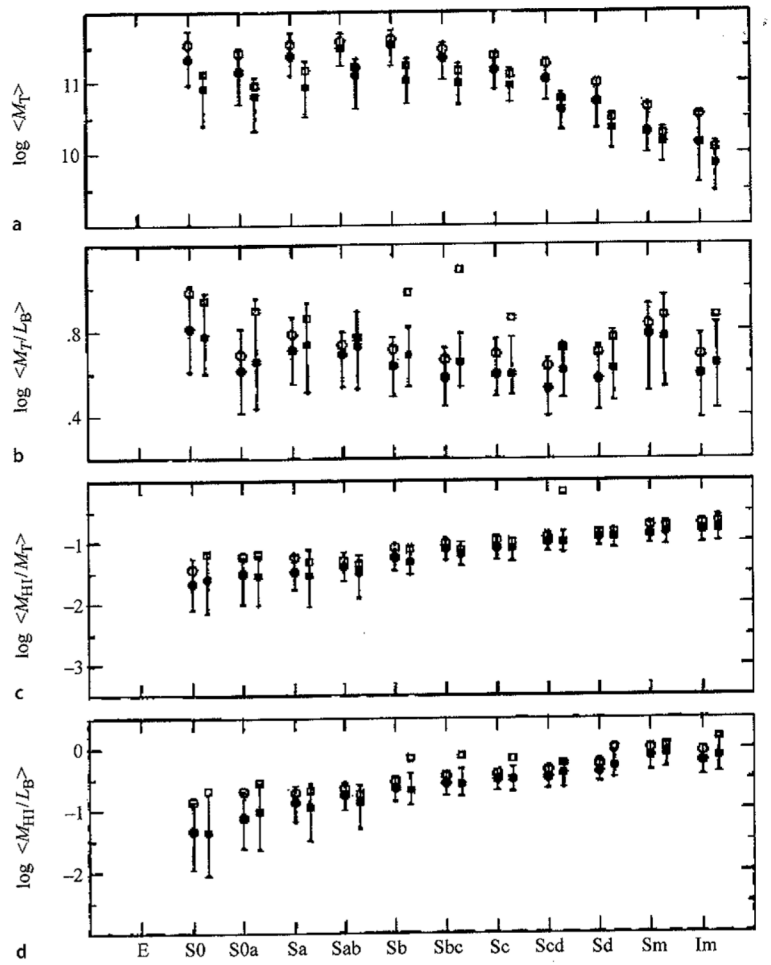


Fig. 3.18a–g. Global galaxy parameters as a function of stage along the Hubble sequence (Roberts and Haynes, 1994). The *circles* represent the galaxies in the RC3-UGC sample and the *squares* those within the local supercluster of galaxies. The *filled circles* are medians; the *open symbols* are mean values. The *error bars* represent the 25 and 75 percentiles of the distributions. **a** Total masses, M_T; **b** Total mass-to-luminosity ratio (M_T/L_B); **c** Neutral hydrogen mass to total mass (M_{HI}/M_T); **d** Neutral hydrogen mass to blue luminosity (M_{HI}/L_B)

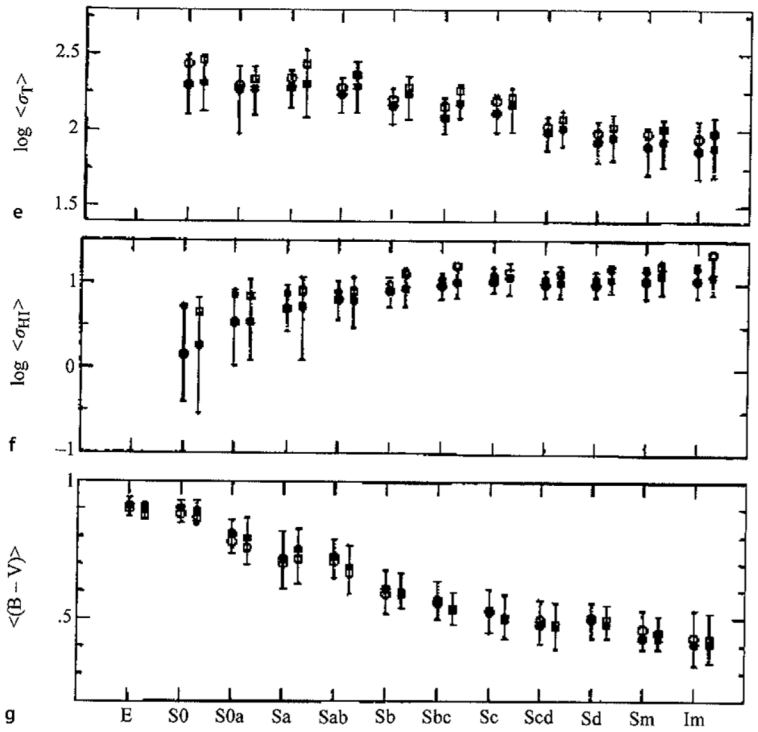


Fig. 3.18. (continued) **e** Total mass surface density (σ_T); **f** Surface mass density of neutral hydrogen (σ_{HI}); **g** Integrated (B − V) colour

the masses of the galaxies decrease monotonically (Fig. 3.18a). The mass-to-luminosity ratios of the galaxies in the sample are roughly constant (Fig. 3.18b) and so it is no surprise that the average luminosity for the S0 to Scd galaxies is roughly constant, whilst it decreases monotonically beyond Scd. These relations again quantify van den Bergh's remark that the classical Hubble types refer primarily to luminous, and consequently, massive galaxies.

- *Neutral hydrogen.* There is a clear distinction between elliptical and spiral galaxies in that very rarely is neutral hydrogen observed in ellipticals while all spiral and late-type galaxies have significant gaseous masses. The upper limit to the mass of neutral hydrogen in elliptical galaxies corresponds to $M_{HI}/M_{tot} \leq 10^{-4}$. For spiral galaxies, the fractional mass of the galaxy in the form of neutral hydrogen ranges from about 0.01 for Sa galaxies to about 0.15 at Sm, the increase being monotonic along the revised Hubble sequence (Fig. 3.18c). The fractional

hydrogen mass is more or less independent of the mass of the galaxy at a particular point along the Hubble sequence. A consequence of the constancy of the M_{tot}/L_B ratio for the galaxies in the sample is that there is also a significant trend for the ratio M_{HI}/L_B to increase along the sequence (Fig. 3.18d).

- *Total surface density and surface density of neutral hydrogen.* These quantities change in opposite senses along the Hubble sequence. The total surface density, as determined by the total mass of the galaxy and its characteristic radius, decreases monotonically along the sequence (Fig. 3.18e), whereas the surface density of neutral hydrogen increases along the sequence (Fig. 3.18f).
- *Integrated colour.* There is a strong correlation in the sense that elliptical galaxies are red whereas late-type galaxies are blue. This relation is shown quantitatively in Fig. 3.18g. Despite the systematic trend, there is a significant dispersion about the relation at each point in the sequence. For example, there are Sc galaxies which are red. As we will see, the analysis of the very large samples of galaxies provided by the SDSS and 2dF Galaxy Surveys have quantified the central importance of colour in understanding the astrophysics of galaxies (Sect. 3.9).
- *Luminosity function of HII regions.* In a pioneering study, Kennicutt and his colleagues determined the luminosity function of HII regions in different galaxy types (Kennicutt et al., 1989). Normalising to the same fiducial mass, it was found that there is a much greater frequency of HII regions in the late-type galaxies as compared with early-type galaxies and that the relation is monotonic along the sequence.

Morton and Haynes pointed out that an obvious interpretation of these correlations is that there are different rates of star formation in different types of galaxy. As they express it, the various correlations provide information about the past, current and future star formation rates in galaxies. The correlation with colour along the sequence is related to the past star formation history of the galaxy; the changes in the luminosity function of HII regions refer to star formation rates at the present epoch; the large fraction of the mass of neutral hydrogen and its large surface density at late stages in the sequence show that these galaxies may continue to have high star formation rates in the future.

To put more flesh on this argument, the integrated colours of galaxies of different Hubble types can be plotted on a (U − B, B − V) colour–colour diagram, the colours being corrected for internal and external reddening. Such a colour–colour diagram for a sample of galaxies selected from the Hubble Atlas of Galaxies is shown in Fig. 3.19 in which it can be seen that the colours of galaxies occupy a remarkably narrow region of the (U − B, B − V) plane (Larson and Tinsley, 1978). There is a monotonic variation of Hubble types along this locus, the bluest galaxies being the Sc and Sd galaxies and the reddest the elliptical galaxies, as can be seen from comparison with Fig. 3.18g. The colours of the galaxies cannot be represented by those of any single class of star which is hardly surprising since different classes of star make the dominant contribution at different wavelengths.

The integrated light of galaxies is principally the sum of the light of main sequence stars plus red giant stars, in particular, the K and M giants. To a rough

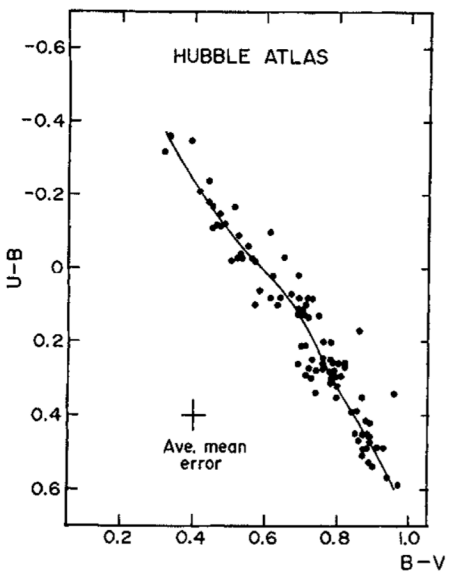


Fig. 3.19. The colour–colour (U–B, B–V) diagram for the integrated colours of galaxies of different morphological types for galaxies selected from the Hubble Atlas of Galaxies (Larson and Tinsley, 1978)

approximation, the colours of galaxies can be represented by the sum of the numbers of luminous blue stars on the main sequence and of luminous giants on the giant branch. If all the stars in galaxies formed 10^{10} years ago, the main sequence termination point would now have reached roughly the mass of the Sun, $M \approx M_{\odot}$, and the brightest main-sequence stars would have spectral properties similar to that of the Sun, that is, a G2 star. There would therefore be no bright blue stars on the main sequence and the integrated light of the galaxy would be dominated by red giants. On the other hand, if star formation has continued over 10^{10} years, or if there were a burst of star formation in the recent past, there would be a significant population of hot blue stars on the main sequence giving the galaxy a significantly bluer colour.

3.9 The Red and Blue Sequences

With the availability of the large samples of galaxies from the SDSS and the 2dF Galaxy Survey, a more quantitative approach to the classification of galaxies has had to be developed, necessitated by the need to analyse these huge samples by computer algorithms. What is lost in detail in these computer-based classifications is more than compensated for by the huge statistics of galaxies with different properties. The upshot of these studies is that what are traditionally referred to as early and late-

type galaxies are found to form two distinct sequences which are known as the *red* and *blue sequences*. In summary:

- The *red sequence* consists of non-star-forming, high mass spheroidal galaxies, or, more colloquially 'old, red and dead' galaxies.
- The *blue sequence* consists of star-forming, low mass galaxies which are disc-dominated.

These two sequences are defined by a number of the characteristic properties which have already been introduced.

3.9.1 Colour Versus Absolute Magnitude

Perhaps the most striking distinction between the two sequences appears in the plot of the colour ${}^{0.1}(g-r)$ against absolute magnitude M. Figure 3.20a shows the distribution of these properties for 144,000 galaxies from the SDSS catalogue (Blanton et al., 2003). Superimposed on the diagram are isodensity contours, the bulk of the galaxies lying within the heavy white contours. The separation into two sequences is clearly defined, the oval region at the top of the diagram being the red sequence and the broader region towards the bottom right the blue sequence.

Baldry and his colleagues have shown that the colour distribution of these galaxies can be separated into red and blue sequences which can be can be very well-described by Gaussian distributions over the magnitude range $-23.5 \leq M_r \leq -15.75$ (Baldry et al., 2004). It is striking how precisely the overall colour distribution in each bin of absolute magnitude over this wide magnitude range can be decomposed into two Gaussian distributions (Fig. 3.21). The red galaxies are the

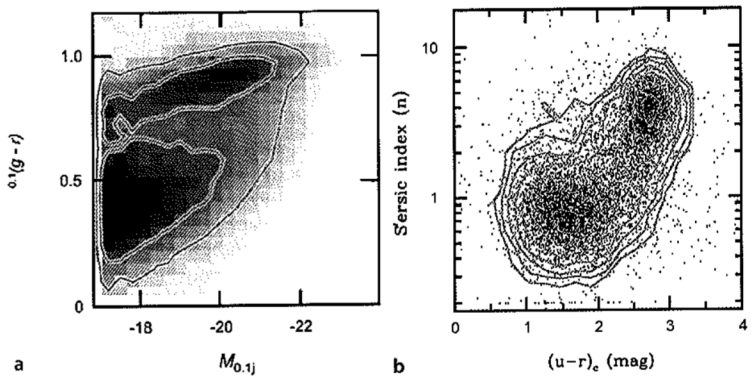


Fig. 3.20. **a** Illustrating the bimodality in the distribution of colour ${}^{0.1}(g-r)$ of galaxies as a function of optical absolute magnitude (Blanton et al., 2003). **b** A plot of Sérsic index against colour for 10,095 galaxies selected from the Millennium Galaxy Catalogue (Driver et al., 2006)

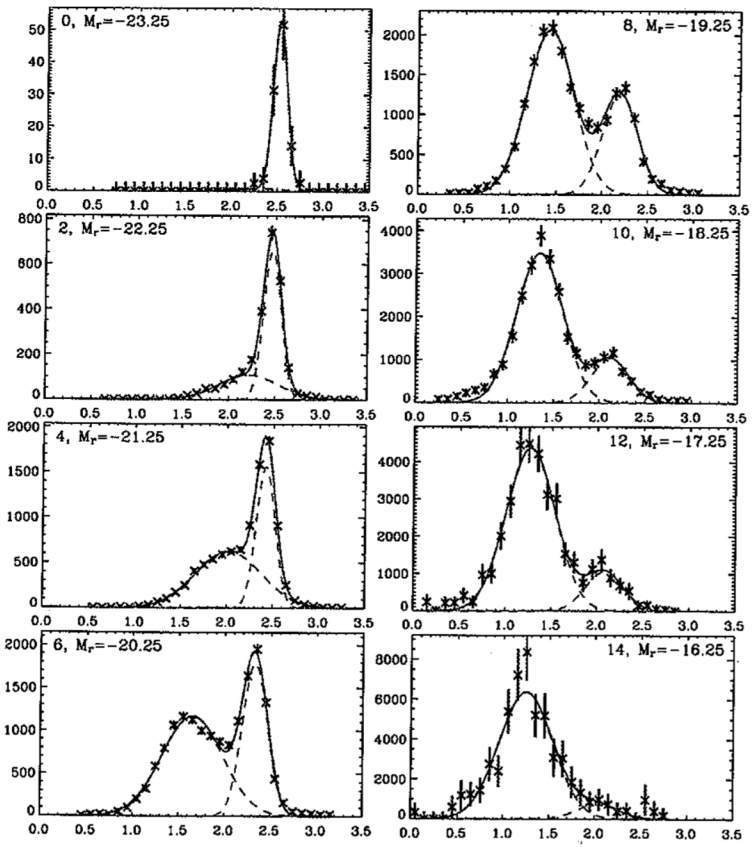


Fig. 3.21. Illustrating the bimodality in the distribution of the colours of galaxies as a function of optical absolute magnitude for a sample of 66,848 galaxies selected from the Sloan Digital Sky Survey (SDSS). The distributions of colours have been fitted by pairs of Gaussians. The data have been binned in intervals of 0.1 in the rest frame $(u - r)$ colour. The galaxy distributions are binned in 0.5 magnitude intervals. Only half of the histograms presented by the authors are shown (Baldry et al., 2004)

most luminous, while the blue galaxies form the dominant population at low absolute magnitudes, as is reflected in the different luminosity functions for red and blue galaxies (Fig. 3.15a).

3.9.2 Sérsic Index and Colour

Bimodality is also present in the structural properties of the galaxies. As seen in Fig. 3.8a, the Sérsic index n can be used to divide galaxies into spheroidal-dominated and disc-dominated galaxies and this shows up even more dramatically in a plot of colour against Sérsic index (Fig. 3.20b) (Driver et al., 2006). As discussed in Sect. 3.4.3, the spheroid-dominated systems are most commonly found with Sérsic parameter $n = 4$, whereas the disc-dominated systems typically have $n = 1$. There is a clear separation between these systems in Fig. 3.8a, but it is even more pronounced in Fig. 3.20b in which the red and blue sequences occupy quite separate regions of the diagram. The dividing line between the two sequences occurs about $n = 2$.

3.9.3 Mean Stellar Age and Concentration Index C

Another approach to separating galaxies into two sequences is to use measures of the age of their stellar populations and the degree of concentration of the light towards their centres. Kauffmann and her colleagues have used sample of 122,808 galaxies from the SDSS to study the average age of their stellar populations using the amplitude of the Balmer break, or discontinuity, at 400 nm, $D_n(4000)$, and the Balmer absorption line index $H\delta_A$. The latter measures the strengths of the Balmer absorption line which are particularly strong in galaxies which have undergone a recent burst of star formation (Kauffmann et al., 2003). They have shown that these indices provide good measures of star formation activity over the last 10^9 and $(1–10) \times 10^9$ years respectively.

The concentration index C is defined to be the ratio $C = (R90/R50)$, where $R90$ and $R50$ are the radii enclosing 90% and 50% of the Petrosian r-band luminosity of the galaxy. The concentration parameter C is strongly correlated with Hubble type, $C = 2.6$ separating early from late-type galaxies. Those galaxies with concentration indices $C \geq 2.6$ are early-type galaxies, reflecting the fact that the light is more concentrated towards their centres.

$D_n(4000)$ and $H\delta_A$ are plotted against the concentration index C and the mean stellar mass density within the half light radius μ_* in Fig. 3.22. The panels of that diagram show that the galaxy populations are divided into two distinct sequences. Kauffmann and her colleagues show that the dividing line between the two sequences occurs at a stellar mass $M \approx 3 \times 10^{10}\, M_\odot$. Lower mass galaxies have young stellar populations, low surface mass densities and the low concentration indices typical of disks. They infer that a significant fraction of the lowest mass galaxies have experienced recent starbursts. For stellar masses $M \geq 3 \times 10^{10}\, M_\odot$, the fraction of galaxies with old stellar populations increases rapidly. These also have the high surface mass densities and high concentration indices typical of spheroids or bulges.

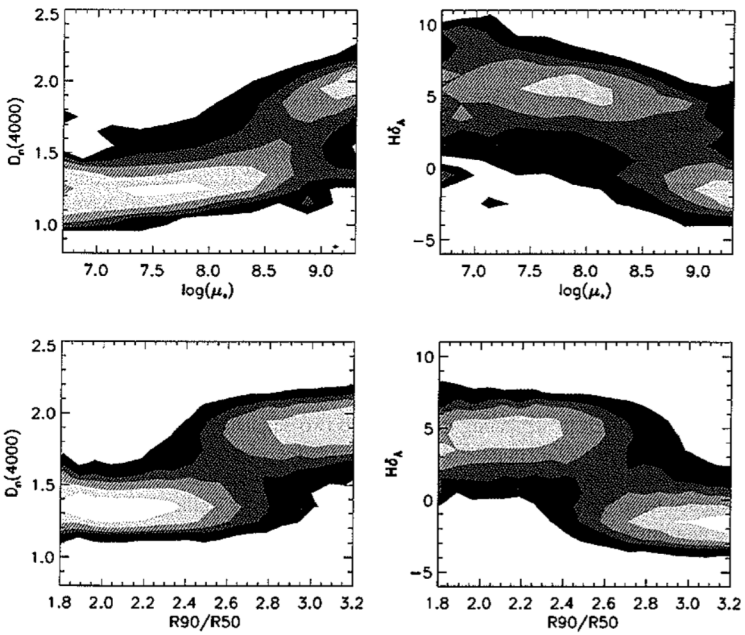


Fig. 3.22. Density distributions showing the trends of the stellar age indicators $D_n(4000)$ and $H\delta_A$ with concentration index $C = (R90/R50)$ and surface mass density μ_* (Kauffmann et al., 2003)

3.9.4 The Effect of the Galaxy Environment

The differences in morphological types of galaxies found in different galactic environments has already been illustrated in Figs. 3.4 and 3.16. Another way of presenting these data emphasising the distinction between the galaxies in the red and blue sequences was carried out by Hogg and his colleagues (Hogg et al., 2004). The sample consisted of 55,158 galaxies in the redshift interval $0.08 \leq z \leq 0.12$. The local galaxy density about any given galaxy was defined by the quantity $\delta_{1\times8}$, meaning the overdensity about any galaxy in a cylindrical volume with transverse comoving radius 1 h^{-1} Mpc and comoving half-length along the line of sight of 8 h^{-1} Mpc. Thus, a galaxy in an environment with the average density of galaxies has $\delta_{1\times8} = 0$. Values of $\delta_{1\times8} \geq 50$ are found in the cores of rich clusters.

The top row of Fig. 3.23 shows contour plots of the number density of galaxies in the colour–absolute magnitude diagram of Fig. 3.21a, but now shown separately for different overdensity environments, ranging from low excess number densities, $\delta_{1\times8} \leq 3$, to very high density environments $\delta_{1\times8} \geq 50$. These data quantify the statement that red galaxies are found preferentially in rich galaxy environments. The

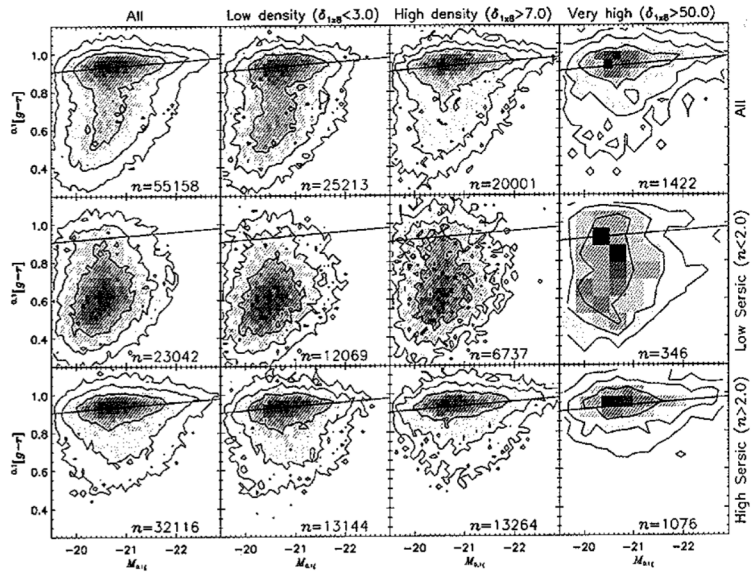


Fig. 3.23. Illustrating the bimodality in the distribution of the colours of galaxies as a function of the density of galaxies in which the galaxy is observed and as a function of their structures as parameterised by the Sérsic index n (Hogg et al., 2004)

second and third rows further split the sample of galaxies into those with Sérsic parameters greater and less than 2. These diagrams quantify the statement that red spheroidal galaxies are found in the richest cluster regions and these are avoided by the blue disc-like galaxies.

3.9.5 The New Perspective

The new way of quantifying the physical properties of galaxies developed in this section illustrates a profound difference of approach to the study of galaxies as compared with, say, ten years ago when the first edition of this book was written. The advent of the huge galaxy surveys as represented by the SDSS and 2dF galaxy surveys have provided the opportunity to quantify by computer algorithm the properties of galaxies which in the past relied somewhat upon the eye of the experienced observer. The division of galaxies into members of the blue and red sequences parallels in many ways the division into early and late-type galaxies. To a good approximation, galaxies earlier than Sa in the Hubble sequence, stage $T = 1$ in de Vaucouleurs' classification (Table 3.1), are members of the red sequence and galaxies later than $T = 1$ belong to the blue sequence.

Of particular importance is the fact that the relative number densities of galaxies of different types are now well-established with large statistics and so are ripe for comparison with the predictions of theories of galaxy formation. As an example of the usefulness of the new statistics, an important result is how the average number, luminosity and mass densities of the stellar component of galaxies in the local Universe are made up. Bell and his colleagues have shown, for example, that while the red sequence contains only 20% of the galaxies by number, these contribute 40% of the stellar luminosity density and 60% of the average stellar mass density at the present epoch (Bell et al., 2003).

3.10 Concluding Remark

This exposition has focussed upon understanding the properties of galaxies *at the present epoch* and has refrained from consideration of the vast amount of data now available on samples of galaxies at earlier epochs, or large redshifts, which provide clues to their origin and evolution. These topics will be taken up in much more detail in Chaps. 17 to 19 once the origin of large-scale structures in the Universe has been established.

4 Clusters of Galaxies

Associations of galaxies range from pairs and small groups, through the giant clusters containing over a thousand galaxies, to the vast structures on scales much greater than clusters such as the vast 'walls' seen in Figs. 2.7 and 2.8. Clustering occurs on all scales, as is demonstrated by the two-point correlation function for galaxies (Figs. 2.5 and 2.6). Few galaxies can be considered truly isolated. Rich clusters of galaxies are of particular interest because they are the largest gravitationally bound systems we know of in the Universe. They possess correspondingly deep gravitational potential wells which can be observed through the bremsstrahlung X-ray emission of hot gas which forms an atmosphere within the cluster. The hot gas can also be detected through the decrements which it causes in the Cosmic Microwave Background Radiation as a result of the Sunyaev–Zeldovich effect.

Clusters, therefore, provide laboratories for studying many different aspects of galactic evolution within rather well-defined astrophysical environments. Interactions of galaxies with each other and with the intergalactic medium in the cluster can be studied, as well as the distribution and nature of the dark matter, which dominates their dynamics. Radio source events can strongly perturb the distribution of hot gas. From the perspective of the formation of large-scale structure, the mass function for clusters of galaxies provides constraints on the development of structure on large scales and on cosmological parameters.

4.1 The Large-Scale Distribution of Clusters of Galaxies

Until relatively recently, the surveys of rich clusters of galaxies which have been the focus of most attention resulted from the pioneering efforts of George Abell. More recently, clusters have been detected by analysing the distribution of galaxies found in machine-scanned surveys of 48-inch Schmidt telescope plates, such as the APM and COSMOS cluster surveys. Most recently, rich clusters have been identified in the large catalogues of galaxies provided by the Sloan Digital Sky Survey. Another approach is to identify clusters of galaxies as extended X-ray sources at high galactic latitudes and this has proved to be an effective procedure which is independent of the need to identify the individual cluster members (Sect. 4.4).