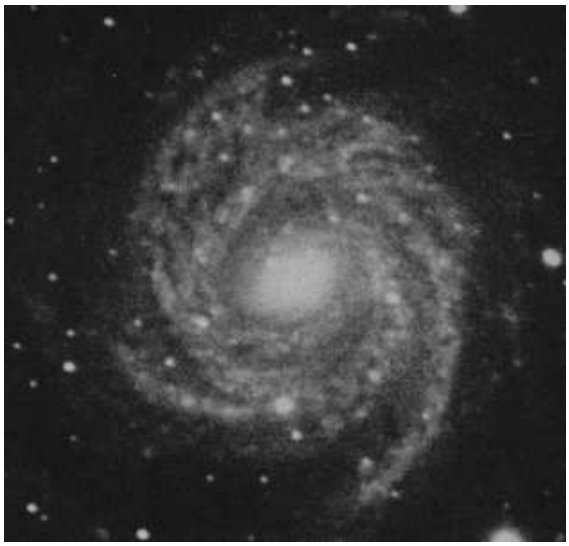
Galaxy Formation and Evolution



Houjun Mo

Department of Astronomy, University of Massachusetts
710 North Pleasant Str., Amherst, MA 01003-9305, USA

Frank van den Bosch

Department of Physics & Astronomy, University of Utah
115 South 1400 East, Salt Lake City, UT 84112-0830, USA

Simon White

Max-Planck Institute for Astrophysics
Karl-Schwarzschild Str. 1, D-85741 Garching, Germany

Contents

1 Introduction *page* 1

1.1 The Diversity of the Galaxy Population 2

1.2 Basic Elements of Galaxy Formation 5
- 1.2.1 The Standard Model of Cosmology 6
- 1.2.2 Initial Conditions 6
- 1.2.3 Gravitational Instability and Structure Formation 7
- 1.2.4 Gas Cooling 8
- 1.2.5 Star Formation 8
- 1.2.6 Feedback Processes 10
- 1.2.7 Mergers 10
- 1.2.8 Dynamical Evolution 12
- 1.2.9 Chemical Evolution 12
- 1.2.10 Stellar Population Synthesis 13
- 1.2.11 The Intergalactic Medium 13

1.3 Time Scales 14

1.4 A Brief History of Galaxy Formation 15
- 1.4.1 Galaxies as Extragalactic Objects 15
- 1.4.2 Cosmology 16
- 1.4.3 Structure Formation 18
- 1.4.4 The Emergence of the Cold Dark Matter Paradigm 20
- 1.4.5 Galaxy Formation 22

2 Observational Facts 25

2.1 Astronomical Observations 25
- 2.1.1 Fluxes and Magnitudes 26
- 2.1.2 Spectroscopy 29
- 2.1.3 Distance Measurements 32

2.2 Stars 34

2.3 Galaxies 38
- 2.3.1 The Classification of Galaxies 38
- 2.3.2 Elliptical Galaxies 42
- 2.3.3 Disk Galaxies 50
- 2.3.4 The Milky Way 56

2.3.5 Dwarf Galaxies 58
2.3.6 Nuclear Star Clusters 60
2.3.7 Starbursts 61
2.3.8 Active Galactic Nuclei 61
2.4 Statistical Properties of the Galaxy Population 62
2.4.1 Luminosity Function 63
2.4.2 Size Distribution 64
2.4.3 Color Distribution 65
2.4.4 The Mass-Metallicity Relation 66
2.4.5 Environment Dependence 67
2.5 Clusters and Groups of Galaxies 68
2.5.1 Clusters of Galaxies 68
2.5.2 Groups of Galaxies 72
2.6 Galaxies at High Redshifts 74
2.6.1 Galaxy Counts 75
2.6.2 Photometric Redshifts 75
2.6.3 Galaxy Redshift Surveys at $z \sim 1$ 77
2.6.4 Lyman-Break Galaxies 78
2.6.5 Lyα Emitters 79
2.6.6 Sub-Millimeter Sources 80
2.6.7 Extremely Red Objects and Distant Red Galaxies 80
2.6.8 The Cosmic Star Formation History 82
2.7 Large-Scale Structure 82
2.7.1 Two-Point Correlation Functions 83
2.7.2 Probing the Matter Field via Weak Lensing 86
2.8 The Intergalactic Medium 86
2.8.1 The Gunn-Peterson Test 87
2.8.2 Quasar Absorption Line Systems 87
2.9 The Cosmic Microwave Background 91
2.10 The Homogeneous and Isotropic Universe 94
2.10.1 The Determination of Cosmological Parameters 96
2.10.2 The Mass and Energy Content of the Universe 97
Bibliography 101

1
Introduction

This book is concerned with the physical processes related to the formation and evolution of galaxies. Simply put, a galaxy is a dynamically bound system that consists of many stars. A typical bright galaxy, such as our own Milky Way, contains a few times 10^{10} stars and has a diameter ($\sim 20\,\mathrm{kpc}$) that is several hundred times smaller than the mean separation between bright galaxies. Since most of the visible stars in the Universe belong to a galaxy, the number density of stars within a galaxy is about 10^7 times higher than the mean number density of stars in the Universe as a whole. In this sense, galaxies are well-defined, astronomical identities. They are also extraordinarily beautiful and diverse objects whose nature, structure and origin have intrigued astronomers ever since the first galaxy images were taken in the mid-nineteenth century.

The goal of this book is to show how physical principles can be used to understand the formation and evolution of galaxies. Viewed as a physical process, galaxy formation and evolution involve two different aspects: (i) initial and boundary conditions; and (ii) physical processes which drive evolution. Thus, in very broad terms, our study will consist of the following parts:

- Cosmology: Since we are dealing with events on cosmological time and length scales, we need to understand the space-time structure on large scales. One can think of the cosmological framework as the stage on which galaxy formation and evolution take place.
- Initial conditions: These were set by physical processes in the early Universe which are beyond our direct view, and which took place under conditions far different from those we can reproduce in earth-bound laboratories.
- Physical processes: As we will show in this book, the basic physics required to study galaxy formation and evolution includes general relativity, hydrodynamics, dynamics of collisionless systems, plasma physics, thermodynamics, electrodynamics, atomic, nuclear and particle physics, and the theory of radiation processes.

In a sense, galaxy formation and evolution can therefore be thought of as an application of (relatively) well-known physics with cosmological initial and boundary conditions. As in many other branches of applied physics, the phenomena to be studied are diverse and interact in many different ways. Furthermore, the physical processes involved in galaxy formation cover some 23 orders of magnitude in physical size, from the scale of the Universe itself down to the scale of individual stars, and about four orders of magnitude in time scales, from the age of the Universe to that of the lifetime of individual, massive stars. Put together, it makes the formation and evolution of galaxies a subject of great complexity.

From an empirical point of view, the study of galaxy formation and evolution is very different from most other areas of experimental physics. This is due mainly to the fact that even the shortest timescales involved are much longer than that of a human being. Consequently, we cannot witness the actual evolution of individual galaxies. However, because the speed of light is finite, looking at galaxies at larger distances from us is equivalent to looking at galaxies when

the Universe was younger. Therefore, we may hope to infer how galaxies form and evolve by comparing their properties, in a statistical sense, at different epochs. In addition, at each epoch we can try to identify regularities and correspondences among the galaxy population. Although galaxies span a wide range in masses, sizes and morphologies, to the extent that no two galaxies are alike, the structural parameters of galaxies also obey various scaling relations, some of which are remarkably tight. These relations must hold important information regarding the physical processes that underlie them, and any successful theory of galaxy formation has to be able to explain their origin.

Galaxies are not only interesting in their own right, they also play a pivotal role in our study of the structure and evolution of the Universe. They are bright, long-lived and abundant, and so can be observed in large numbers over cosmological distances and time scales. This makes them unique tracers of the evolution of the Universe as a whole, and detailed studies of their large scale distribution can provide important constraints on cosmological parameters. In this book we therefore also describe the large scale distribution of galaxies, and discuss how it can be used to test cosmological models.

In Chapter 2 we start by describing the observational properties of stars, galaxies and the large scale structure of the Universe as a whole. Chapters **??** through **??** describe the various physical ingredients needed for a self-consistent model of galaxy formation, ranging from the cosmological framework to the formation and evolution of individual stars. Finally, in Chapters **??** to **??** we combine these physical ingredients to examine how galaxies form and evolve in a cosmological context, using the observational data as constraints.

The purpose of this introductory chapter is to sketch our current ideas about galaxies and their formation process, without going into any detail. After a brief overview of some observed properties of galaxies, we list the various physical processes that play a role in galaxy formation and outline how they are connected. We also give a brief historical overview of how our current views of galaxy formation have been shaped.

1.1 The Diversity of the Galaxy Population

Galaxies are a diverse class of objects. This means that a large number of parameters is required in order to characterize any given galaxy. One of the main goals of any theory of galaxy formation is to explain the full probability distribution function of all these parameters. In particular, as we will see in Chapter 2, many of these parameters are correlated with each other, a fact which any successful theory of galaxy formation should also be able to reproduce.

Here we list briefly the most salient parameters that characterize a galaxy. This overview is necessarily brief and certainly not complete. However, it serves to stress the diversity of the galaxy population, and to highlight some of the most important observational aspects that galaxy formation theories need to address. A more thorough description of the observational properties of galaxies is given in Chapter 2.

(a) Morphology One of the most noticeable properties of the galaxy population is the existence of two basic galaxy types: spirals and ellipticals. Elliptical galaxies are mildly flattened, ellipsoidal systems that are mainly supported by the random motions of their stars. Spiral galaxies, on the other hand, have highly flattened disks that are mainly supported by rotation. Consequently, they are also often referred to as disk galaxies. The name 'spiral' comes from the fact that the gas and stars in the disk often reveal a clear spiral pattern. Finally, for historical reasons, ellipticals and spirals are also called early- and late-type galaxies, respectively.

Most galaxies, however, are neither a perfect ellipsoid nor a perfect disk, but rather a combination of both. When the disk is the dominant component, its ellipsoidal component is generally

called the bulge. In the opposite case, of a large ellipsoidal system with a small disk, one typically talks about a disky elliptical. One of the earliest classification schemes for galaxies, which is still heavily used, is the Hubble sequence. Roughly speaking, the Hubble sequence is a sequence in the admixture of the disk and ellipsoidal components in a galaxy, which ranges from early-type ellipticals that are pure ellipsoids to late-type spirals that are pure disks. As we will see in Chapter 2, the important aspect of the Hubble sequence is that many intrinsic properties of galaxies, such as luminosity, color, and gas content, change systematically along this sequence. In addition, disks and ellipsoids most likely have very different formation mechanisms. Therefore, the morphology of a galaxy, or its location along the Hubble sequence, is directly related to its formation history.

For completeness, we stress that not all galaxies fall in this spiral vs. elliptical classification. The faintest galaxies, called dwarf galaxies, typically do not fall on the Hubble sequence. Dwarf galaxies with significant amounts of gas and ongoing star formation typically have a very irregular structure, and are consequently called (dwarf) irregulars. Dwarf galaxies without gas and young stars are often very diffuse, and are called dwarf spheroidals. In addition to these dwarf galaxies, there is also a class of brighter galaxies whose morphology neither resembles a disk nor a smooth ellipsoid. These are called peculiar galaxies and include, among others, galaxies with double or multiple subcomponents linked by filamentary structure and highly-distorted galaxies with extended tails. As we will see, they are usually associated with recent mergers or tidal interactions. Although peculiar galaxies only constitute a small fraction of the entire galaxy population, their existence conveys important information about how galaxies may have changed their morphologies during their evolutionary history.

(b) Luminosity and Stellar Mass Galaxies span a wide range in luminosity. The brightest galaxies have luminosities of $\sim 10^{12}\,\mathrm{L}_{\odot}$, where $\mathrm{L}_{\odot}$ indicates the luminosity of the Sun. The exact lower limit of the luminosity distribution is less well defined, and is subject to regular changes, as fainter and fainter galaxies are constantly being discovered. In 2007 the faintest galaxy known was a newly discovered dwarf spheroidal Willman I, with a total luminosity somewhat below $1000\,\mathrm{L}_{\odot}$.

Obviously, the total luminosity of a galaxy is related to its total number of stars, and thus to its total stellar mass. However, the relation between luminosity and stellar mass reveals a significant amount of scatter, because different galaxies have different stellar populations. As we will see in Chapter **??**, galaxies with a younger stellar population have a higher luminosity per unit stellar mass than galaxies with an older stellar population.

An important statistic of the galaxy population is its luminosity probability distribution function, also known as the luminosity function. As we will see in Chapter 2, there are many more faint galaxies than bright galaxies, so that the faint ones clearly dominate the number density. However, in terms of the contribution to the total luminosity density, neither the faintest nor the brightest galaxies dominate. Instead, it is the galaxies with a characteristic luminosity similar to that of our Milky Way that contribute most to the total luminosity density in the present-day Universe. This indicates that there is a characteristic scale in galaxy formation, which is accentuated by the fact that most galaxies that are brighter than this characteristic scale are ellipticals, while those that are fainter are mainly spirals (at the very faint end dwarf irregulars and dwarf spheroidals dominate). Understanding the physical origin of this characteristic scale has turned out to be one of the most challenging problems in contemporary galaxy formation modeling.

(c) Size and Surface Brightness As we will see in Chapter 2, galaxies do not have well defined boundaries. Consequently, several different definitions for the size of a galaxy can be found in the literature. One measure often used is the radius enclosing a certain fraction (e.g., half) of the total luminosity. In general, as one might expect, brighter galaxies are bigger. However, even for

a fixed luminosity, there is a considerable scatter in sizes, or in surface brightness, defined as the luminosity per unit area.

The size of a galaxy has an important physical meaning. In disk galaxies, which are rotation supported, the sizes are a measure of their specific angular momenta (see Chapter **??**). In the case of elliptical galaxies, which are supported by random motions, the sizes are a measure of the amount of dissipation during their formation (see Chapter **??**). Therefore, the observed distribution of galaxy sizes is an important constraint for galaxy formation models.

(d) Gas Mass Fraction Another useful parameter to describe galaxies is their cold gas mass fraction, defined as $f_{\rm gas} = M_{\rm cold}/[M_{\rm cold} + M_\star]$, with $M_{\rm cold}$ and $M_\star$ the masses of cold gas and stars, respectively. This ratio expresses the efficiency with which cold gas has been turned into stars. Typically, the gas mass fractions of ellipticals are negligibly small, while those of disk galaxies increase systematically with decreasing surface brightness. Indeed, the lowest surface brightness disk galaxies can have gas mass fractions in excess of 90 percent, in contrast to our Milky Way which has $f_{\rm gas} \sim 0.1$.

(e) Color Galaxies also come in different colors. The color of a galaxy reflects the ratio of its luminosity in two photometric passbands. A galaxy is said to be red if its luminosity in the redder passband is relatively high compared to that in the bluer passband. Ellipticals and dwarf spheroidals generally have redder colors than spirals and dwarf irregulars. As we will see in Chapter **??**, the color of a galaxy is related to the characteristic age and metallicity of its stellar population. In general, redder galaxies are either older or more metal rich (or both). Therefore, the color of a galaxy holds important information regarding its stellar population. However, extinction by dust, either in the galaxy itself, or along the line-of-sight between the source and the observer, also tends to make a galaxy appear red. As we will see, separating age, metallicity and dust effects is one of the most daunting tasks in observational astronomy.

(f) Environment As we will see in §§2.5-2.7, galaxies are not randomly distributed throughout space, but show a variety of structures. Some galaxies are located in high density clusters containing several hundreds of galaxies, some in smaller groups containing a few to tens of galaxies, while yet others are distributed in low-density filamentary or sheet-like structures. Many of these structures are gravitationally bound, and may have played an important role in the formation and evolution of the galaxies. This is evident from the fact that elliptical galaxies seem to prefer cluster environments, whereas spiral galaxies are mainly found in relative isolation (sometimes called the field). As briefly discussed in §1.2.8 below, it is believed that this morphology-density relation reflects enhanced dynamical interaction in denser environments, although we still lack a detailed understanding of its origin.

(g) Nuclear Activity For the majority of galaxies, the observed light is consistent with what we expect from a collection of stars and gas. However, a small fraction of all galaxies, called active galaxies, show an additional non-stellar component in their spectral energy distribution. As we will see in Chapter **??**, this emission originates from a small region in the centers of these galaxies, called the active galactic nucleus (AGN), and is associated with matter accretion onto a supermassive black hole. According to the relative importance of such non-stellar emission, one can separate active galaxies from normal (or non-active) galaxies.

(h) Redshift Because of the expansion of the Universe, an object that is farther away will have a larger receding velocity, and thus a larger redshift. Since the light from high-redshift galaxies was emitted when the Universe was younger, we can study galaxy evolution by observing the galaxy population at different redshifts. In fact, in a statistical sense the high-redshift galaxies are the progenitors of present-day galaxies, and any changes in the number density or intrinsic properties of galaxies with redshift give us a direct window on the formation and evolution of the

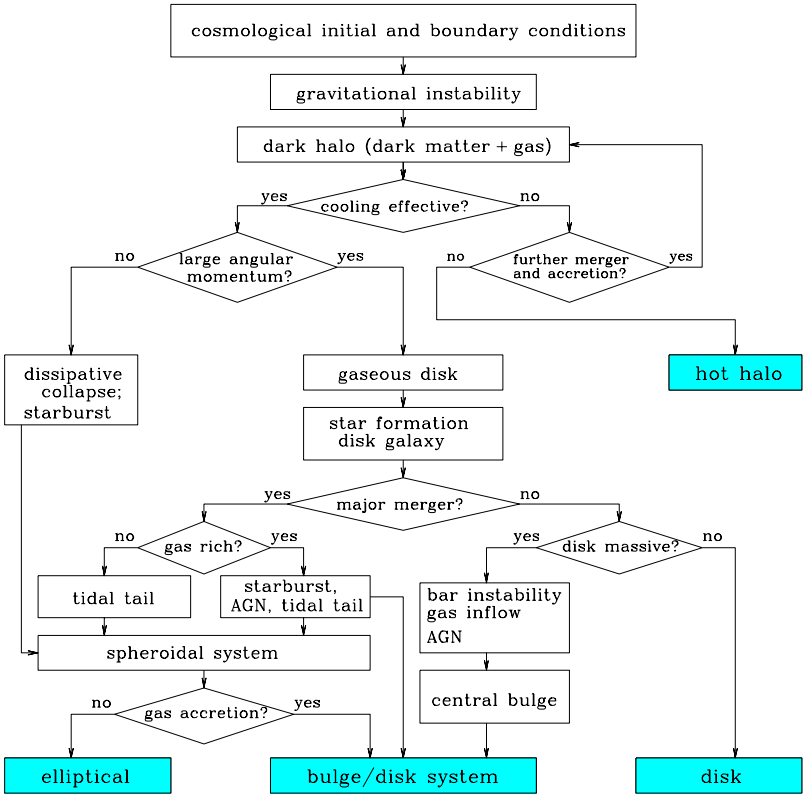


Fig. 1.1. A logic-flow chart for galaxy formation. In the standard scenario, the initial and boundary conditions for galaxy formation are set by the cosmological framework. The paths leading to the formation of various galaxies are shown along with the relevant physical processes. Note, however, that processes do not separate as neatly as this figure suggests. For example, cold gas may not have the time to settle into a gaseous disk before a major merger takes place.

galaxy population. With modern, large telescopes we can now observe galaxies out to redshifts beyond six, making possible for us to probe the galaxy population back to a time when the Universe was only about 10 percent of its current age.

1.2 Basic Elements of Galaxy Formation

Before diving into details, it is useful to have an overview of the basic theoretical framework within which our current ideas about galaxy formation and evolution have been developed. In this section we give a brief overview of the various physical processes that play a role during the formation and evolution of galaxies. The goal is to provide the reader with a picture of the relationships among the various aspects of galaxy formation to be addressed in greater detail in the chapters to come. To guide the reader, Fig. 1.1 shows a flow-chart of galaxy formation, which illustrates how the various processes to be discussed below are intertwined. It is important to stress, though, that this particular flow-chart reflects our current, undoubtedly incomplete view of galaxy formation. Future improvements in our understanding of galaxy formation and evolution may add new links to the flow-chart, or may render some of the links shown obsolete.

1.2.1 The Standard Model of Cosmology

Since galaxies are observed over cosmological length and time scales, the description of their formation and evolution must involve cosmology, the study of the properties of space-time on large scales. Modern cosmology is based upon the Cosmological Principle, the hypothesis that the Universe is spatially homogeneous and isotropic, and Einstein's theory of General Relativity, according to which the structure of space-time is determined by the mass distribution in the Universe. As we will see in Chapter **??**, these two assumptions together lead to a cosmology (the standard model) that is completely specified by the curvature of the Universe, K, and the scale factor, $a(t)$, describing the change of the length scale of the Universe with time. One of the basic tasks in cosmology is to determine the value of K and the form of $a(t)$ (hence the spacetime geometry of the Universe on large scales), and to show how observables are related to physical quantities in such a universe.

Modern cosmology not only specifies the large-scale geometry of the Universe, but also has the potential to predict its thermal history and matter content. Because the Universe is expanding and filled with microwave photons at the present time, it must have been smaller, denser and hotter at earlier times. The hot and dense medium in the early Universe provides conditions under which various reactions among elementary particles, nuclei and atoms occur. Therefore, the application of particle, nuclear and atomic physics to the thermal history of the Universe in principle allows us to predict the abundances of all species of elementary particles, nuclei and atoms at different epochs. Clearly, this is an important part of the problem to be addressed in this book, because the formation of galaxies depends crucially on the matter/energy content of the Universe.

In currently popular cosmologies we usually consider a Universe consisting of three main components. In addition to the 'baryonic' matter, the protons, neutrons and electrons† that make up the *visible* Universe, astronomers have found various indications for the presence of dark matter and dark energy (see Chapter 2 for a detailed discussion of the observational evidence). Although the nature of both dark matter and dark energy is still unknown, we believe that they are responsible for more than 95 percent of the energy density of the Universe. Different cosmological models differ mainly in (i) the relative contributions of baryonic matter, dark matter, and dark energy, and (ii) the nature of dark matter and dark energy. At the time of writing, the most popular model is the so-called ΛCDM model, a flat universe in which ~ 75 percent of the energy density is due to a cosmological constant, ~ 21 percent is due to 'cold' dark matter (CDM), and the remaining 4 percent is due to the baryonic matter out of which stars and galaxies are made. Chapter **??** gives a detailed description of these various components, and describes how they influence the expansion history of the Universe.

1.2.2 Initial Conditions

If the cosmological principle held perfectly and the distribution of matter in the Universe were perfectly uniform and isotropic, there would be no structure formation. In order to explain the presence of structure, in particular galaxies, we clearly need some deviations from perfect uniformity. Unfortunately, the standard cosmology does not in itself provide us with an explanation for the origin of these perturbations. We have to go beyond it to search for an answer.

A classical, General Relativistic description of cosmology is expected to break down at very early times when the Universe is so dense that quantum effects are expected to be important. As we will see in §**??**, the standard cosmology has a number of conceptual problems when applied to the early Universe, and the solutions to these problems require an extension of the standard

† Although an electron is a lepton, and not a baryon, in cosmology it is standard practice to include electrons when talking of baryonic matter

cosmology to incorporate quantum processes. One generic consequence of such an extension is the generation of density perturbations by quantum fluctuations at early times. It is believed that these perturbations are responsible for the formation of the structures observed in today's Universe.

As we will see in §**??**, one particularly successful extension of the standard cosmology is the inflationary theory, in which the Universe is assumed to have gone through a phase of rapid, exponential expansion (called inflation) driven by the vacuum energy of one or more quantum fields. In many, but not all, inflationary models, quantum fluctuations in this vacuum energy can produce density perturbations with properties consistent with the observed large-scale structure. Inflation thus offers a promising explanation for the physical origin of the initial perturbations. Unfortunately, our understanding of the very early Universe is still far from complete, and we are currently unable to predict the initial conditions for structure formation entirely from first principles. Consequently, even this part of galaxy formation theory is still partly phenomenological: typically initial conditions are specified by a set of parameters that are constrained by observational data, such as the pattern of fluctuations in the microwave background or the present-day abundance of galaxy clusters.

1.2.3 Gravitational Instability and Structure Formation

Having specified the initial conditions and the cosmological framework, one can compute how small perturbations in the density field evolve. As we will see in Chapter **??**, in an expanding universe dominated by non-relativistic matter, perturbations grow with time. This is easy to understand. A region whose initial density is slightly higher than the mean will attract its surroundings slightly more strongly than average. Consequently, over-dense regions pull matter towards them and become even more over-dense. On the other hand, under-dense regions become even more rarefied as matter flows away from them. This amplification of density perturbations is referred to as gravitational instability and plays an important role in modern theories of structure formation. In a static universe, the amplification is a run-away process, and the density contrast $\delta\rho/\rho$ grows exponentially with time. In an expanding universe, however, the cosmic expansion damps accretion flows, and the growth rate is usually a power law of time, $\delta\rho/\rho \propto t^{\alpha}$, with $\alpha > 0$. As we will see in Chapter **??**, the exact rate at which the perturbations grow depends on the cosmological model.

At early times, when the perturbations are still in what we call the linear regime ($\delta\rho/\rho \ll 1$), the physical size of an overdense region increases with time due to the overall expansion of the Universe. Once the perturbation reaches overdensity $\delta\rho/\rho \sim 1$, it breaks away from the expansion and starts to collapse. This moment of 'turn-around', when the physical size of the perturbation is at its maximum, signals the transition from the mildly non-linear regime to the strongly non-linear regime.

The outcome of the subsequent non-linear, gravitational collapse depends on the matter content of the perturbation. If the perturbation consists of ordinary baryonic gas, the collapse creates strong shocks that raise the entropy of the material. If radiative cooling is inefficient, the system relaxes to hydrostatic equilibrium, with its self-gravity balanced by pressure gradients. If the perturbation consists of collisionless matter (e.g., cold dark matter), no shocks develop, but the system still relaxes to a quasi-equilibrium state with a more-or-less universal structure. This process is called violent relaxation and will be discussed in Chapter **??**. Non-linear, quasi-equilibrium dark matter objects are called dark matter halos. Their predicted structure has been thoroughly explored using numerical simulations, and they play a pivotal role in modern theories of galaxy formation. Chapter **??** therefore presents a detailed discussion of the structure and formation of dark matter halos. As we shall see, halo density profiles, shapes, spins and internal substructure

all depend very weakly on mass and on cosmology, but the abundance and characteristic density of halos depend sensitively on both of these.

In cosmologies with both dark matter and baryonic matter, such as the currently favored CDM models, each initial perturbation contains baryonic gas and collisionless dark matter in roughly their universal proportions. When an object collapses, the dark matter relaxes violently to form a dark matter halo, while the gas shocks to the virial temperature, $T_{\rm vir}$ (see §**??** for a definition) and may settle into hydrostatic equilibrium in the potential well of the dark matter halo if cooling is slow.

1.2.4 Gas Cooling

Cooling is a crucial ingredient of galaxy formation. Depending on temperature and density, a variety of cooling processes can affect gas. In massive halos, where the virial temperature $T_{\rm vir} \gtrsim 10^7\,{\rm K}$, gas is fully collisionally ionized and cools mainly through Bremsstrahlung emission from free electrons. In the temperature range $10^4\,{\rm K} < T_{\rm vir} < 10^6\,{\rm K}$, a number of excitation and de-excitation mechanisms can play a role. Electrons can recombine with ions, emitting a photon, or atoms (neutral or partially ionized) can be excited by a collision with another particle, thereafter decaying radiatively to the ground state. Since different atomic species have different excitation energies, the cooling rates depend strongly on the chemical composition of the gas. In halos with $T_{\rm vir} < 10^4\,{\rm K}$, gas is predicted to be almost completely neutral. This strongly suppresses the cooling processes mentioned above. However, if heavy elements and/or molecules are present, cooling is still possible through the collisional excitation/de-excitation of fine and hyperfine structure lines (for heavy elements) or rotational and/or vibrational lines (for molecules). Finally, at high redshifts ($z \gtrsim 6$), inverse Compton scattering of cosmic microwave background photons by electrons in hot halo gas can also be an effective cooling channel. Chapter **??** will discuss these cooling processes in more detail.

Except for inverse Compton scattering, all these cooling mechanisms involve two particles. Consequently, cooling is generally more effective in higher density regions. After non-linear gravitational collapse, the shocked gas in virialized halos may be dense enough for cooling to be effective. If cooling times are short, the gas never comes to hydrostatic equilibrium, but rather accretes directly onto the central protogalaxy. Even if cooling is slow enough for a hydrostatic atmosphere to develop, it may still cause the denser inner regions of the atmosphere to lose pressure support and to flow onto the central object. The net effect of cooling is thus that the baryonic material segregates from the dark matter, and accumulates as dense, cold gas in a protogalaxy at the center of the dark matter halo.

As we will see in Chapter **??**, dark matter halos, as well as the baryonic material associated with them, typically have a small amount of angular momentum. If this angular momentum is conserved during cooling, the gas will spin up as it flows inwards, settling in a cold disk in centrifugal equilibrium at the center of the halo. This is the standard paradigm for the formation of disk galaxies, which we will discuss in detail in Chapter **??**.

1.2.5 Star Formation

As the gas in a dark matter halo cools and flows inwards, its self-gravity will eventually dominate over the gravity of the dark matter. Thereafter it collapses under its own gravity, and in the presence of effective cooling, this collapse becomes catastrophic. Collapse increases the density and temperature of the gas, which generally reduces the cooling time more rapidly than it reduces the collapse time. During such runaway collapse the gas cloud may fragment into small, high-density cores that may eventually form stars (see Chapter **??**), thus giving rise to a visible galaxy.

Unfortunately, many details of these processes are still unclear. In particular, we are still

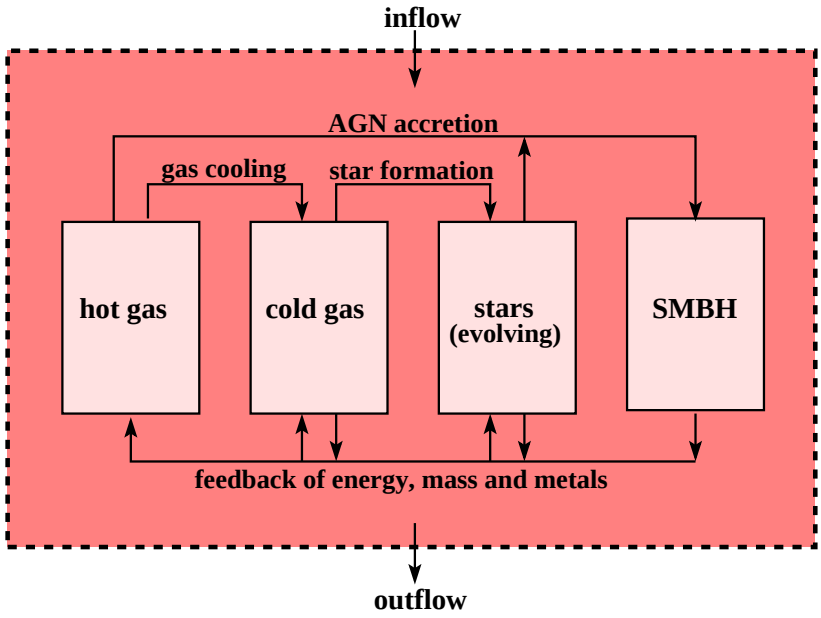


Fig. 1.2. A flow chart of the evolution of an individual galaxy. The galaxy is represented by the dashed box which contains hot gas, cold gas, stars and a supermassive black hole (SMBH). Gas cooling converts hot gas into cold gas, star formation converts cold gas into stars, and dying stars inject energy, metals and gas into the gas components. In addition, the SMBH can accrete gas (both hot and cold) as well as stars, producing AGN activity which can release vast amounts of energy which affect primarily the gaseous components of the galaxy. Note that in general the box will not be closed: gas can be added to the system through accretion from the intergalactic medium and can escape the galaxy through outflows driven by feedback from the stars and/or the SMBH. Finally, a galaxy may merge or interact with another galaxy, causing a significant boost or suppression of all these processes.

unable to predict the mass fraction of, and the time-scale for, a self-gravitating cloud to be transformed into stars. Another important and yet poorly-understood issue is concerned with the mass distribution with which stars are formed, i.e. the initial mass function (IMF). As we will see in Chapter **??**, the evolution of a star, in particular its luminosity as function of time and its eventual fate, is largely determined by its mass at birth. Predictions of observable quantities for model galaxies thus require not only the birth rate of stars as a function of time, but also their IMF. In principle, it should be possible to derive the IMF from first principles, but the theory of star formation has not yet matured to this level. At present one has to assume an IMF *ad hoc* and check its validity by comparing model predictions to observations.

Based on observations, we will often distinguish two modes of star formation: quiescent star formation in rotationally supported gas disks, and starbursts. The latter are characterized by much higher star formation rates, and are typically confined to relatively small regions (often the nucleus) of galaxies. Starbursts require the accumulation of large amounts of gas in a small volume, and appear to be triggered by strong dynamical interactions or instabilities. These processes will be discussed in more detail in §1.2.8 below and in Chapter **??**. At the moment, there are still many open questions related to these different modes of star formation. What fraction of stars formed in the quiescent mode? Do both modes produce stellar populations with the same IMF? How does the relative importance of starbursts scale with time? As we will see, these and related questions play an important role in contemporary models of galaxy formation.

1.2.6 Feedback Processes

When astronomers began to develop the first dynamical models for galaxy formation in a CDM dominated universe, it immediately became clear that most baryonic material is predicted to cool and form stars. This is because in these 'hierarchical' structure formation models, small dense halos form at high redshift and cooling within them is predicted to be very efficient. This disagrees badly with observations, which show that only a relatively small fraction of all baryons are in cold gas or stars (see Chapter 2). Apparently, some physical process must either prevent the gas from cooling, or reheat it after it has become cold.

Even the very first models suggested that the solution to this problem might lie in feedback from supernovae, a class of exploding stars that can produce enormous amounts of energy (see §**??**). The radiation and the blastwaves from these supernovae may heat (or reheat) surrounding gas, blowing it out of the galaxy in what is called a galactic wind. These processes are described in more detail in §**??** and §**??**.

Another important feedback source for galaxy formation is provided by Active Galactic Nuclei (AGN), the active accretion phase of supermassive black holes (SMBH) lurking at the centers of almost all massive galaxies (see Chapter **??**). This process releases vast amounts of energy – this is why AGN are bright and can be seen out to large distances, which can be tapped by surrounding gas. Although only a relatively small fraction of present-day galaxies contain an AGN, observations indicate that virtually all massive spheroids contain a nuclear SMBH (see Chapter 2). Therefore, it is believed that virtually all galaxies with a significant spheroidal component have gone through one or more AGN phases during their life.

Although it has become clear over the years that feedback processes play an important role in galaxy formation, we are still far from understanding which processes dominate, and when and how exactly they operate. Furthermore, to make accurate predictions for their effects, one also needs to know how often they occur. For supernovae this requires a prior understanding of the star formation rates and the IMF. For AGN it requires understanding how, when and where supermassive black holes form, and how they accrete mass.

It should be clear from the above discussion that galaxy formation is a subject of great complexity, involving many strongly intertwined processes. This is illustrated in Fig. 1.2, which shows the relations between the four main baryonic components of a galaxy, hot gas, cold gas, stars, and a supermassive black hole. Cooling, star formation, AGN accretion and feedback processes can all shift baryons from one of these components to another, thereby altering the efficiency of all the processes. For example, increased cooling of hot gas will produce more cold gas. This in turn will increases the star formation rate, hence the supernova rate. The additional energy injection from supernovae can reheat cold gas, thereby suppressing further star formation (negative feedback). On the other hand, supernova blastwaves may also compress the surrounding cold gas, so as to boost the star formation rate (positive feedback). Understanding these various feedback loops is one of the most important and intractable issues in contemporary models for the formation and evolution of galaxies.

1.2.7 Mergers

So far we have considered what happens to a single, isolated system of dark matter, gas and stars. However, galaxies and dark matter halos are not isolated. For example, as illustrated in Fig. 1.2, systems can accrete new material (both dark and baryonic matter) from the intergalactic medium, and can lose material through outflows driven by feedback from stars and/or AGN. In addition, two (or more) systems may merge to form a new system with very different properties from its progenitors. In the currently popular CDM cosmologies, the initial density fluctuations have larger amplitudes on smaller scales. Consequently, dark matter halos grow hierarchically,

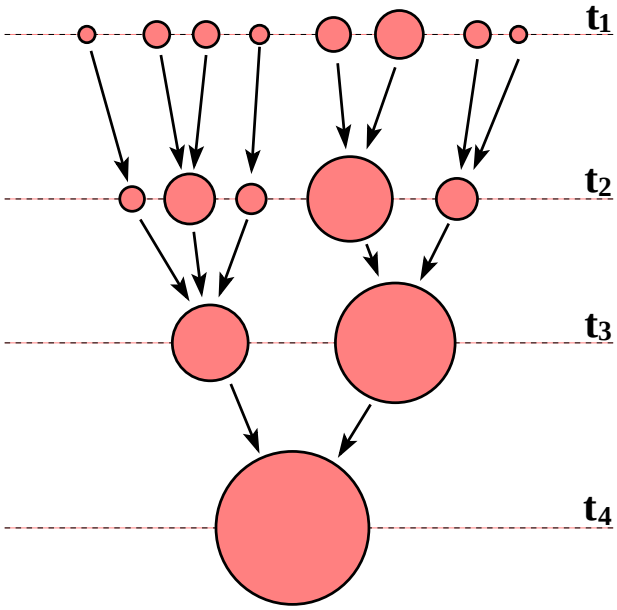


Fig. 1.3. A schematic merger tree, illustrating the merger history of a dark matter halo. It shows, at three different epochs, the progenitor halos that at time t_4 have merged to form a single halo. The size of each circle represents the mass of the halo. Merger histories of dark matter halos play an important role in hierarchical theories of galaxy formation.

in the sense that larger halos are formed by the coalescence (merging) of smaller progenitors. Such a formation process is usually called a hierarchical or 'bottom-up' scenario.

The formation history of a dark matter halo can be described by a 'merger tree' that traces all its progenitors, as illustrated in Fig. 1.3. Such merger trees play an important role in modern galaxy formation theory. Note, however, that illustrations such as Fig. 1.3 can be misleading. In CDM models part of the growth of a massive halo is due to merging with a large number of much smaller halos, and to a good approximation, such mergers can be thought of as smooth accretion. When two similar mass dark matter halos merge, violent relaxation rapidly transforms the orbital energy of the progenitors into the internal binding energy of the quasi-equilibrium remnant. Any hot gas associated with the progenitors is shock-heated during the merger and settles back into hydrostatic equilibrium in the new halo. If the progenitor halos contained central galaxies, the galaxies also merge as part of the violent relaxation process, producing a new central galaxy in the final system. Such a merger may be accompanied by strong star formation or AGN activity if the merging galaxies contained significant amounts of cold gas. If two merging halos have very different mass, the dynamical processes are less violent. The smaller system orbits within the main halo for an extended period of time during which two processes compete to determine its eventual fate. Dynamical friction transfers energy from its orbit to the main halo, causing it to spiral inwards, while tidal effects remove mass from its outer regions and may eventually dissolve it completely (see Chapter **??**). Dynamical friction is more effective for more massive satellites, but if the mass ratio of the initial halos is large enough, the smaller object (and any galaxy associated with it) can maintain its identity for a long time. This is the process for the build-up of clusters of galaxies: a cluster may be considered as a massive dark matter halo hosting a relatively massive galaxy near its center and many satellites that have not yet dissolved or merged with the central galaxy.

As we will see in Chapters **??** and **??**, numerical simulations show that the merger of two

galaxies of roughly equal mass produces an object reminiscent of an elliptical galaxy, and the result is largely independent of whether the progenitors are spirals or ellipticals. Indeed, current hierarchical models of galaxy formation assume that most, if not all, elliptical galaxies are merger remnants. If gas cools onto this merger remnant with significant angular momentum, a new disk may form, producing a disk-bulge system like that in an early-type spiral galaxy.

It should be obvious from the above discussion that mergers play a crucial role in galaxy formation. Detailed descriptions of halo mergers and galaxy mergers are presented in Chapter **??** and Chapter **??**, respectively.

1.2.8 Dynamical Evolution

When satellite galaxies orbit within dark matter halos, they experience tidal forces due to the central galaxy, due to other satellite galaxies, and due to the potential of the halo itself. These tidal interactions can remove dark matter, gas and stars from the galaxy, a process called tidal stripping (see §**??**), and may also perturb its structure. In addition, if the halo contains a hot gas component, any gas associated with the satellite galaxy will experience a drag force due to the relative motion of the two fluids. If the drag force exceeds the restoring force due to the satellite's own gravity, its gas will be ablated, a process called ram-pressure stripping. These dynamical processes are thought to play an important role in driving galaxy evolution within clusters and groups of galaxies. In particular, they are thought to be partially responsible for the observed environmental dependence of galaxy morphology (see Chapter **??**).

Internal dynamical effects can also reshape galaxies. For example, a galaxy may form in a configuration which becomes unstable at some later time. Large-scale instabilities may then redistribute mass and angular momentum within the galaxy, thereby changing its morphology. A well-known and important example is the bar-instability within disk galaxies. As we shall see in §**??**, a thin disk with too high a surface density is susceptible to a non-axisymmetric instability, which produces a bar-like structure similar to that seen in barred spiral galaxies. These bars may then buckle out of the disk to produce a central ellipsoidal component, a so-called 'pseudo-bulge'. Instabilities may also be triggered in otherwise stable galaxies by interactions. Thus, an important question is whether the sizes and morphologies of galaxies were set at formation, or are the result of later dynamical process ('secular evolution', as it is termed). Bulges are particularly interesting in this context. They may be a remnant of the first stage of galaxy formation, or as mentioned in §1.2.7, may reflect an early merger which has grown a new disk, or may result from buckling of a bar. It is likely that all these processes are important for at least some bulges.

1.2.9 Chemical Evolution

In astronomy, all chemical elements heavier than helium are collectively termed 'metals'. The mass fraction of a baryonic component (e.g. hot gas, cold gas, stars) in metals is then referred to as its metallicity. As we will see in §**??**, the nuclear reactions during the first three minutes of the Universe (the epoch of primordial nucleosynthesis) produced primarily hydrogen ($\sim 75\%$) and helium ($\sim 25\%$), with a very small admixture of metals dominated by lithium. All other metals in the Universe were formed at later times as a consequence of nuclear reactions in stars. When stars expel mass in stellar winds, or in supernova explosions, they enrich the interstellar medium (ISM) with newly synthesized metals.

Evolution of the chemical composition of the gas and stars in galaxies is important for several reasons. First of all, the luminosity and color of a stellar population depend not only on its age and IMF, but also on the metallicity of the stars (see Chapter **??**). Secondly, the cooling efficiency of gas depends strongly on its metallicity, in the sense that more metal-enriched gas cools faster (see §**??**). Thirdly, small particles of heavy elements known as dust grains, which

are mixed with the interstellar gas in galaxies, can absorb significant amounts of the starlight and re-radiate it in infrared wavelengths. Depending on the amount of the dust in the ISM, which scales roughly linearly with its metallicity (see §**??**), this interstellar extinction can significantly reduce the brightness of a galaxy.

As we will see in Chapter **??**, the mass and detailed chemical composition of the material ejected by a stellar population as it evolves depend both on the IMF and on its initial metallicity. In principle, observations of the metallicity and abundance ratios of a galaxy can therefore be used to constrain its star formation history and IMF. In practice, however, the interpretation of the observations is complicated by the fact that galaxies can accrete new material of different metallicity, that feedback processes can blow out gas, perhaps preferentially metals, and that mergers can mix the chemical compositions of different systems.

1.2.10 Stellar Population Synthesis

The light we receive from a given galaxy is emitted by a large number of stars that may have different masses, ages, and metallicities. In order to interpret the observed spectral energy distribution, we need to predict how each of these stars contributes to the total spectrum. Unlike many of the ingredients in galaxy formation, the theory of stellar evolution, to be discussed in Chapter **??**, is reasonably well understood. This allows us to compute not only the evolution of the luminosity, color and spectrum of a star of given initial mass and chemical composition, but also the rates at which it ejects mass, energy and metals into the interstellar medium. If we know the star formation history (i.e., the star formation rate as a function of time) and IMF of a galaxy, we can then synthesize its spectrum at any given time by adding together the spectra of all the stars, after evolving each to the time under consideration. In addition, this also yields the rates at which mass, energy and metals are ejected into the interstellar medium, providing important ingredients for modeling the chemical evolution of galaxies.

Most of the energy of a stellar population is emitted in the optical, or, if the stellar population is very young ($\lesssim 10\,$Myr), in the ultraviolet (see §**??**). However, if the galaxy contains a lot of dust, a significant fraction of this optical and UV light may get absorbed and re-emitted in the infrared. Unfortunately, predicting the final emergent spectrum is extremely complicated. Not only does it depend on the amount of the radiation absorbed, it also depends strongly on the properties of the dust, such as its geometry, its chemical composition, and (the distribution of) the sizes of the dust grains (see §**??**).

Finally, to complete the spectral energy distribution emitted by a galaxy, we also need to add the contribution from a possible AGN. Chapter **??** discusses various emission mechanisms associated with accreting SMBHs. Unfortunately, as we will see, we are still far from being able to predict the detailed spectra for AGN.

1.2.11 The Intergalactic Medium

The intergalactic medium (IGM) is the baryonic material lying between galaxies. This is and has always been the dominant baryonic component of the Universe and it is the material from which galaxies form. Detailed studies of the IGM can therefore give insight into the properties of the pregalactic matter before it condensed into galaxies. As illustrated in Fig. 1.2, galaxies do not evolve as closed boxes, but can affect the properties of the IGM through exchanges of mass, energy and heavy elements. The study of the IGM is thus an integral part of understanding how galaxies form and evolve. As we will see in Chapter **??**, the properties of the IGM can be probed most effectively through the absorption it produces in the spectra of distant quasars (a certain class of active galaxies, see Chapter **??**). Since quasars are now observed out to redshifts beyond

6, their absorption line spectra can be used to study the properties of the IGM back to a time when the Universe was only a few percent of its present age.

1.3 Time Scales

As discussed above, and as illustrated in Fig. 1.1, the formation of an individual galaxy in the standard, hierarchical formation scenario involves the following processes: the collapse and virialization of dark matter halos, the cooling and condensation of gas within the halo, and the conversion of cold gas into stars and a central supermassive black hole. Evolving stars and active AGN eject energy, mass and heavy elements into the interstellar medium, thereby determining its structure and chemical composition and perhaps driving winds into the intergalactic medium. Finally, galaxies can merge and interact, re-shaping their morphology and triggering further starbursts and AGN activity. In general, the properties of galaxies are determined by the competition among all these processes, and a simple way to characterize the relative importance of these processes is to use the time scales associated with them. Here we give a brief summary of the most important time scales in this context.

- **Hubble time:** This is an estimate of the time scale on which the Universe as a whole evolves. It is defined as the inverse of the Hubble constant (see §**??**), which specifies the current cosmic expansion rate. It would be equal to the time since the Big Bang if the Universe had always expanded at its current rate. Roughly speaking, this is the timescale on which substantial evolution of the galaxy population is expected.
- **Dynamical time:** This is the time required to orbit across an equilibrium dynamical system. For a system with mass M and radius R, we define it as $t_{\rm dyn} = \sqrt{3\pi/16G\overline{\rho}}$, where $\overline{\rho} = 3M/4\pi R^3$. This is related to the free-fall time, defined as the time required for a uniform, pressure-free sphere to collapse to a point, as $t_{\rm ff} = t_{\rm dyn}/\sqrt{2}$.
- **Cooling time:** This time scale is the ratio between the thermal energy content and the energy loss rate (through radiative or conductive cooling) for a gas component.
- **Star-formation time:** This time scale is the ratio of the cold gas content of a galaxy to its star-formation rate. It is thus an indication of how long it would take for the galaxy to run out of gas if the fuel for star formation is not replenished.
- **Chemical enrichment time:** This is a measure for the time scale on which the gas is enriched in heavy elements. This enrichment time is generally different for different elements, depending on the lifetimes of the stars responsible for the bulk of the production of each element (see §**??**).
- **Merging time:** This is the typical time that a halo or galaxy must wait before experiencing a merger with an object of similar mass, and is directly related to the major merger frequency.
- **Dynamical friction time:** This is the time scale on which a satellite object in a large halo loses its orbital energy and spirals to the center. As we will see in §**??**, this time scale is proportional to $M_{\rm sat}/M_{\rm main}$, where $M_{\rm sat}$ is the mass of the satellite object and $M_{\rm main}$ is that of the main halo. Thus, more massive galaxies will merge with the central galaxy in a halo more quickly than smaller ones.

These time scales can provide guidelines for incorporating the underlying physical processes in models of galaxy formation and evolution, as we describe in later chapters. In particular, comparing time scales can give useful insights. As an illustration, consider the following examples:

- Processes whose time scale is longer than the Hubble time can usually be ignored. For example, satellite galaxies with mass less than a few percent of their parent halo normally have dynamical friction times exceeding the Hubble time (see §**??**). Consequently, their orbits do

not decay significantly. This explains why clusters of galaxies have so many 'satellite' galaxies – the main halos are so much more massive than a typical galaxy that dynamical friction is ineffective.

- If the cooling time is longer than the dynamical time, hot gas will typically be in hydrostatic equilibrium. In the opposite case, however, the gas cools rapidly, losing pressure support, and collapsing to the halo center on a free-fall time without establishing any hydrostatic equilibrium.
- If the star formation time is comparable to the dynamical time, gas will turn into stars during its initial collapse, a situation which may lead to the formation of something resembling an elliptical galaxy. On the other hand, if the star formation time is much longer than the cooling and dynamical times, the gas will settle into a centrifugally supported disk before forming stars, thus producing a disk galaxy (see §1.4.5).
- If the relevant chemical evolution time is longer than the star formation time, little metal enrichment will occur during star formation and all stars will end up with the same, initial metallicity. In the opposite case, the star-forming gas is continuously enriched, so that stars formed at different times will have different metallicities and abundance patterns (see §**??**).

So far we have avoided one obvious question, namely, what is the time scale for galaxy formation itself? Unfortunately, there is no single useful definition for such a time scale. Galaxy formation is a process, not an event, and as we have seen, this process is an amalgam of many different elements, each with its own time scale. If, for example, we are concerned with its stellar population, we might define the formation time of a galaxy as the epoch when a fixed fraction (e.g. 1% or 50%) of its stars had formed. If, on the other hand, we are concerned with its structure, we might want to define the galaxy's formation time as the epoch when a fixed fraction (e.g. 50% or 90%) of its mass was first assembled into a single object. These two 'formation' times can differ greatly for a given galaxy, and even their ordering can change from one galaxy to another. Thus it is important to be precise about definition when talking about the formation times of galaxies.

1.4 A Brief History of Galaxy Formation

The picture of galaxy formation sketched above is largely based on the hierarchical cold dark matter model for structure formation, which has been the standard paradigm since the beginning of the 1980s. In the following, we give an historical overview of the development of ideas and concepts about galaxy formation up to the present time. This is not intended as a complete historical account, but rather as a summary for young researchers of how our current ideas about galaxy formation were developed. Readers interested in a more extensive historical review can find some relevant material in the book 'The Cosmic Century: A History of Astrophysics and Cosmology' by Malcolm Longair.

1.4.1 Galaxies as Extragalactic Objects

By the end of the 19th century, astronomers had discovered a large number of astronomical objects that differ from stars in that they are fuzzy rather than point-like. These objects were collectively referred to as 'nebulae'. During the period 1771 to 1784 the French astronomer Charles Messier cataloged more than 100 of these objects in order to avoid confusing them with the comets he was searching for. Today the Messier numbers are still used to designate a number of bright galaxies. For example, the Andromeda galaxy is also known as M31, because it is the 31st nebula in Messier's catalog. A more systematic search for nebulae was carried

out by the Herschels, and in 1864 John Herschel published his *General Catalogue of Galaxies* which contains 5079 nebular objects. In 1888, Dreyer published an expanded version as his *New General Catalogue of Nebulae and Clusters of Stars*. Together with its two supplementary *Index Catalogues*, Dreyer's catalogue contained about 15,000 objects. Today, NGC and IC numbers are still widely used to refer to galaxies.

For many years after their discovery, the nature of the nebular objects was controversial. There were two competing ideas, one assumed that all nebulae are objects within our Milky Way, the other that some might be extragalactic objects, individual 'island universes' like the Milky Way. In 1920 the National Academy of Sciences in Washington invited two leading astronomers, Harlow Shapley and Heber Curtis, to debate this issue, an event which has passed into astronomical folklore as 'The Great Debate'. The controversy remained unresolved until 1925, when Edwin Hubble used distances estimated from Cepheid variables to demonstrate conclusively that some nebulae are extragalactic, individual galaxies comparable to our Milky Way in size and luminosity. Hubble's discovery marked the beginning of extragalactic astronomy. During the 1930s, high-quality photographic images of galaxies enabled him to classify galaxies into a broad sequence according to their morphology. Today Hubble's sequence is still widely adopted to classify galaxies.

Since Hubble's time, astronomers have made tremendous progress in systematically searching the skies for galaxies. At present deep CCD imaging and high-quality spectroscopy are available for about a million galaxies.

1.4.2 Cosmology

Only four years after his discovery that galaxies truly are extragalactic, Hubble made his second fundamental breakthrough: he showed that the recession velocities of galaxies are linearly related to their distances (Hubble, 1929, see also Hubble & Humason 1931), thus demonstrating that our Universe is expanding. This is undoubtedly the greatest single discovery in the history of cosmology. It revolutionized our picture of the Universe we live in.

The construction of mathematical models for the Universe actually started somewhat earlier. As soon as Albert Einstein completed his theory of General Relativity in 1916, it was realized that this theory allowed, for the first time, the construction of self-consistent models for the Universe as a whole. Einstein himself was among the first to explore such solutions of his field equations. To his dismay, he found that all solutions require the Universe either to expand or to contract, in contrast with his belief at that time that the Universe should be static. In order to obtain a static solution, he introduced a cosmological constant into his field equations. This additional constant of gravity can oppose the standard gravitational attraction and so make possible a static (though unstable) solution. In 1922 Alexander Friedmann published two papers exploring both static and expanding solutions. These models are today known as Friedmann models, although this work drew little attention until Georges Lemaitre independently rediscovered the same solutions in 1927.

An expanding universe is a natural consequence of General Relativity, so it is not surprising that Einstein considered his introduction of a cosmological constant as 'the biggest blunder of my life' once he learned of Hubble's discovery. History has many ironies, however. As we will see later, the cosmological constant is now back with us. In 1998 two teams independently used the distance-redshift relation of Type Ia supernovae to show that the expansion of the Universe is accelerating at the present time. Within General Relativity this requires an additional mass/energy component with properties very similar to those of Einstein's cosmological constant. Rather than just counterbalancing the attractive effects of 'normal' gravity, the cosmological constant today overwhelms them to drive an ever more rapid expansion.

Since the Universe is expanding, it must have been denser and perhaps also hotter at earlier

times. In the late 1940's this prompted George Gamow to suggest that the chemical elements may have been created by thermonuclear reactions in the early Universe, a process known as primordial nucleosynthesis. Gamow's model was not considered a success, because it was unable to explain the existence of elements heavier than lithium due to the lack of stable elements with atomic mass numbers 5 and 8. We now know that this was not a failure at all; all heavier elements are a result of nucleosynthesis within stars, as first shown convincingly by Fred Hoyle and collaborators in the 1950s. For Gamow's model to be correct, the Universe would have to be hot as well as dense at early times, and Gamow realized that the residual heat should still be visible in today's Universe as a background of thermal radiation with a temperature of a few degrees Kelvin, thus with a peak at microwave wavelengths. This was a remarkable prediction of the cosmic microwave background radiation (CMB), which was finally discovered in 1965. The thermal history suggested by Gamow, in which the Universe expands from a dense and hot initial state, was derisively referred to as the Hot Big Bang by Fred Hoyle, who preferred an unchanging Steady State Cosmology. Hoyle's cosmological theory was wrong, but his name for the correct model has stuck.

The Hot Big Bang model developed gradually during the 1950s and 1960s. By 1964, it had been noticed that the abundance of helium by mass is everywhere about one third that of hydrogen, a result which is difficult to explain by nucleosynthesis in stars. In 1964, Hoyle and Tayler published calculations that demonstrated how the observed helium abundance could emerge from the Hot Big Bang. Three years later, Wagoner et al. (1967) made detailed calculations of a complete network of nuclear reactions, confirming the earlier result and suggesting that the abundances of other light isotopes, such as helium-3, deuterium and lithium could also be explained by primordial nucleosynthesis. This success provided strong support for the Hot Big Bang. The 1965 discovery of the cosmic microwave background showed it to be isotropic and to have a temperature (2.7K) exactly in the range expected in the Hot Big Bang model (Penzias & Wilson, 1965; Dicke et al., 1965). This firmly established the Hot Big Bang as the standard model of cosmology, a status which it has kept up to the present day. Although there have been changes over the years, these have affected only the exact matter/energy content of the model and the exact values of its characteristic parameters.

Despite its success, during the 1960s and 1970s it was realized that the standard cosmology had several serious shortcomings. Its structure implies that the different parts of the Universe we see today were never in causal contact at early times (e.g., Misner, 1968). How then can these regions have contrived to be so similar, as required by the isotropy of the CMB? A second shortcoming is connected with the spatial flatness of the Universe (e.g. Dicke & Peebles, 1979). It was known by the 1960s that the matter density in the Universe is not very different from the critical density for closure, i.e., the density for which the spatial geometry of the Universe is flat. However, in the standard model any tiny deviation from flatness in the early Universe is amplified enormously by later evolution. Thus, extreme fine tuning of the initial curvature is required to explain why so little curvature is observed today. A closely related formulation is to ask how our Universe has managed to survive and to evolve for billions of years, when the timescales of all physical processes in its earliest phases were measured in tiny fractions of a nanosecond. The standard cosmology provides no explanations for these puzzles.

A conceptual breakthrough came in 1981 when Alan Guth proposed that the Universe may have gone through an early period of exponential expansion (inflation) driven by the vacuum energy of some quantum field. His original model had some problems and was revised in 1982 by Linde and by Albrecht & Steinhardt. In this scenario, the different parts of the Universe we see today were indeed in causal contact *before* inflation took place, thereby allowing physical processes to establish homogeneity and isotropy. Inflation also solves the flatness/timescale problem, because the Universe expanded so much during inflation that its curvature radius grew

to be much larger than the presently observable Universe. Thus, a generic prediction of the inflation scenario is that today's Universe should appear flat.

1.4.3 Structure Formation

(a) Gravitational Instability In the standard model of cosmology, structures form from small initial perturbations in an otherwise homogeneous and isotropic universe. The idea that structures can form via gravitational instability in this way originates from Jeans (1902), who showed that the stability of a perturbation depends on the competition between gravity and pressure. Density perturbations grow only if they are larger (heavier) than a characteristic length (mass) scale [now referred to as the Jeans' length (mass)] beyond which gravity is able to overcome the pressure gradients. The application of this Jeans criterion to an expanding background was worked out by, among others, Gamow & Teller (1939) and Lifshitz (1946), with the result that perturbation growth is power-law in time, rather than exponential as for a static background.

(b) Initial Perturbations Most of the early models of structure formation assumed the Universe to contain two energy components, ordinary baryonic matter and radiation (CMB photons and relativistic neutrinos). In the absence of any theory for the origin of perturbations, two distinct models were considered, usually referred to as adiabatic and isothermal initial conditions. In adiabatic initial conditions all matter and radiation fields are perturbed in the same way, so that the total density (or local curvature) varies, but the ratio of photons to baryons, for example, is spatially invariant. Isothermal initial conditions, on the other hand, correspond to initial perturbations in the ratio of components, but with no associated spatial variation in the total density or curvature.†

In the adiabatic case, the perturbations can be considered as applying to a single fluid with a constant specific entropy as long as the radiation and matter remain tightly coupled. At such times, the Jeans' mass is very large and small-scale perturbations execute acoustic oscillations driven by the pressure gradients associated with the density fluctuations. Silk (1968) showed that towards the end of recombination, as radiation decouples from matter, small-scale oscillations are damped by photon diffusion, a process now called Silk damping. Depending on the matter density and the expansion rate of the Universe, the characteristic scale of Silk damping falls in the range of $10^{12} - 10^{14}\,\mathrm{M}_{\odot}$. After radiation/matter decoupling the Jeans' mass drops precipitously to $\simeq 10^{6}\,\mathrm{M}_{\odot}$ and perturbations above this mass scale can start to grow,‡ but there are no perturbations left on the scale of galaxies at this time. Consequently, galaxies must form 'top-down', via the collapse and fragmentation of perturbations larger than the damping scale, an idea championed by Zel'dovich and colleagues.

In the case of isothermal initial conditions, the spatial variation in the ratio of baryons to photons remains fixed before recombination because of the tight coupling between the two fluids. The pressure is spatially uniform, so that there is no acoustic oscillation, and perturbations are not influenced by Silk damping. If the initial perturbations include small-scale structure, this survives until after the recombination epoch, when baryon fluctuations are no longer supported by photon pressure and so can collapse. Structure can then form 'bottom-up' through hierarchical clustering. This scenario of structure formation was originally proposed by Peebles (1965).

By the beginning of the 1970s, the linear evolution of both adiabatic and isothermal perturbations had been worked out in great detail (e.g., Lifshitz, 1946; Silk, 1968; Peebles & Yu, 1970; Sato, 1971; Weinberg, 1971). At that time, it was generally accepted that observed structures must have formed from finite amplitude perturbations which were somehow part of the initial

† Note that the nomenclature 'isothermal', which is largely historical, is somewhat confusing; the term 'isocurvature' would be more appropriate.

‡ Actually, as we will see in Chapter **??**, depending on the gauge adopted, perturbations can also grow before they enter the horizon.

conditions set up at the Big Bang. Harrison (1970) and Zeldovich (1972) independently argued that only one scaling of the amplitude of initial fluctuations with their wavelength could be consistent with the formation of galaxies from fluctuations imposed at very early times. Their suggestion, now known as the Harrison-Zel'dovich initial fluctuation spectrum, has the property that structure on every scale has the same dimensionless amplitude, corresponding to fluctuations in the equivalent Newtonian gravitational potential, $\delta\Phi/c^2 \sim 10^{-4}$.

In the early 1980s, immediately after the inflationary scenario was proposed, a number of authors realized almost simultaneously that quantum fluctuations of the scalar field (called the inflaton) that drives inflation can generate density perturbations with a spectrum that is close to the Harrison-Zeldovich form (Hawking, 1982; Guth & Pi, 1982; Starobinsky, 1982; Bardeen et al., 1983). In the simplest models, inflation also predicts that the perturbations are adiabatic and that the initial density field is Gaussian. When parameters take their natural values, however, these models generically predict fluctuation amplitudes that are much too large, of order unity. This apparent fine-tuning problem is still unresolved.

In 1992 anisotropy in the cosmic microwave background was detected convincingly for the first time by the Cosmic Background Explorer (COBE) (Smoot et al., 1992). These anisotropies provide an image of the structure present at the time of radiation/matter decoupling, $\sim$400,000 years after the Big Bang. The resolved structures are all of very low amplitude and so can be used to probe the properties of the initial density perturbations. In agreement with the inflationary paradigm, the COBE maps were consistent with Gaussian initial perturbations with the Harrison-Zel'dovich spectrum. The fluctuation amplitudes are comparable to those inferred by Harrison and Zel'dovich. The COBE results have since been confirmed and dramatically refined by subsequent observations, most notably by the Wilkinson Microwave Anisotropy Probe (WMAP) (Bennett et al., 2003; Hinshaw et al., 2007). The agreement with simple inflationary predictions remains excellent.

(c) Non-Linear Evolution In order to connect the initial perturbations to the non-linear structures we see today, one has to understand the outcome of non-linear evolution. In 1970 Zel'dovich published an analytical approximation (now referred to as the Zel'dovich approximation) which describes the initial non-linear collapse of a coherent perturbation of the cosmic density field. This model shows that the collapse generically occurs first along one direction, producing a sheet-like structure, often referred to as a 'pancake'. Zeldovich imagined further evolution to take place via fragmentation of such pancakes. At about the same time, Gunn & Gott (1972) developed a simple spherically symmetric model to describe the growth, turn-around (from the general expansion), collapse and virialization of a perturbation. In particular, they showed that dissipationless collapse results in a quasi-equilibrium system with a characteristic radius that is about half the radius at turn-around. Although the non-linear collapse described by the Zel'dovich approximation is more realistic, since it does not assume any symmetry, the spherical collapse model of Gunn & Gott has the virtue that it links the initial perturbation directly to the final quasi-equilibrium state. By applying this model to a Gaussian initial density field, Press & Schechter (1974) developed a very useful formalism (now referred to as Press-Schechter theory) that allows one to estimate the mass function of collapsed objects (i.e., their abundance as a function of mass) produced by hierarchical clustering.

Hoyle (1949) was the first to suggest that perturbations (and the associated proto-galaxies) might gain angular momentum through the tidal torques from their neighbors. A linear perturbation analysis of this process was first carried out correctly and in full generality by Doroshkevich (1970), and was later tested with the help of numerical simulations (Peebles, 1971; Efstathiou & Jones, 1979). The study of Efstathiou and Jones showed that clumps formed through gravitational collapse in a cosmological context typically acquire about 15% of the angular momentum needed for full rotational support. Better simulations in more recent years have shown that the

correct value is closer to 10%. In the case of 'top-down' models, it was suggested that objects could acquire angular momentum not only through gravitational torques as pancakes fragment, but also via oblique shocks generated by their collapse (Doroshkevich, 1973).

1.4.4 The Emergence of the Cold Dark Matter Paradigm

The first evidence that the Universe may contain dark matter (undetected through electromagnetic emission or absorption) can be traced back to 1933, when Zwicky studied the velocities of galaxies in the Coma cluster and concluded that the total mass required to hold the cluster together is about 400 times larger than the luminous mass in stars. In 1937 he reinforced this analysis and noted that galaxies associated with such large amounts of mass should be detectable as gravitational lenses producing multiple images of background galaxies. These conclusions were substantially correct, but remarkably it took more than 40 years for the existence of dark matter to be generally accepted. The tide turned in the mid-1970s with papers by Ostriker et al. (1974) and Einasto et al. (1974) extending Zwicky's analysis and noting that massive halos are required around our Milky Way and other nearby galaxies in order to explain the motions of their satellites. These arguments were supported by continually improving 21cm and optical measurements of spiral galaxy rotation curves which showed no sign of the fall-off at large radius expected if the visible stars and gas were the only mass in the system (Roberts & Rots, 1973; Rubin et al., 1978, 1980). During the same period, numerous suggestions were made regarding the possible nature of this dark matter component, ranging from baryonic objects such as brown-dwarfs, white dwarfs and black holes (e.g., White & Rees, 1978; Carr et al., 1984), to more exotic, elementary particles such as massive neutrinos (Gershtein & Zel'Dovich, 1966; Cowsik & McClelland, 1972).

The suggestion that neutrinos might be the unseen mass was partly motivated by particle physics. In the 1960s and 1970s, it was noticed that Grand Unified Theories (GUTs) permit the existence of massive neutrinos, and various attempts to measure neutrino masses in laboratory experiments were initiated. In the late 1970s, Lyubimov et al. (1980) and Reines et al. (1980) announced the detection of a mass for the electron neutrino at a level of cosmological interest (about 30 eV). Although the results were not conclusive, they caused a surge in studies investigating neutrinos as dark matter candidates (e.g., Bond et al., 1980; Sato & Takahara, 1980; Schramm & Steigman, 1981; Klinkhamer & Norman, 1981), and structure formation in a neutrino-dominated universe was soon worked out in detail. Since neutrinos decouple from other matter and radiation fields while still relativistic, their abundance is very similar to that of CMB photons. Thus, they must have become nonrelativistic at the time the Universe became matter-dominated, implying thermal motions sufficient to smooth out all structure on scales smaller than a few tens of Mpc. The first non-linear structures are then Zel'dovich pancakes of this scale, which must fragment to make smaller structures such as galaxies. Such a picture conflicts directly with observation, however. An argument by Tremaine & Gunn (1979), based on the Pauli exclusion principle, showed that individual galaxy halos could not be made of neutrinos with masses as small as 30 eV, and simulations of structure formation in neutrino-dominated universes by White et al. (1984) demonstrated that they could not produce galaxies without at the same time producing much stronger galaxy clustering than is observed. Together with the failure to confirm the claimed neutrino mass measurements, these problems caused a precipitous decline in interest in neutrino dark matter by the end of the 1980s.

In the early 1980s, alternative models were suggested, in which dark matter is a different kind of weakly interacting massive particle. There were several motivations for this. The amount of baryonic matter allowed by cosmic nucleosynthesis calculations is far too little to provide the flat universe preferred by inflationary models, suggesting that non-baryonic dark matter may be present. In addition, strengthening upper limits on temperature anisotropies in the CMB made it

increasingly difficult to construct self-consistent, purely baryonic models for structure formation; there is simply not enough time between the recombination epoch and the present day to grow the structures we see in the nearby Universe from those present in the high-redshift photon-baryon fluid. Finally, by the early 1980s, particle physics models based on the idea of supersymmetry had provided a plethora of dark matter candidates, such as neutralinos, photinos and gravitinos, that could dominate the mass density of the Universe. Because of their much larger mass, such particles would initially have much smaller velocities than a 30 eV neutrino, and so they were generically referred to as Warm or Cold Dark Matter (WDM or CDM, the former corresponding to a particle mass of order 1 keV, the latter to much more massive particles) in contrast to neutrino-like Hot Dark Matter (HDM). The shortcomings of HDM motivated consideration of a variety of such scenarios (e.g., Peebles, 1982; Blumenthal et al., 1982; Bond et al., 1982; Bond & Szalay, 1983).

Lower thermal velocities result in the survival of fluctuations of galactic scale (for WDM and CDM) or below (for CDM). The particles decouple from the radiation field long before recombination, so perturbations in their density can grow at early times to be substantially larger than the fluctuations visible in the CMB. After the baryons decouple from the radiation, they quickly fall in these dark matter potential wells, causing structure formation to occur sufficiently fast to be consistent with observed structure in today's Universe. Davis et al. (1985) used simulations of the CDM model to show that it could provide a good match to the observed clustering of galaxies provided either the mass density of dark matter is well below the critical value, or (their preferred model) that galaxies are biased tracers of the CDM density field, as expected if they form at the centers of the deepest dark matter potential wells (e.g. Kaiser, 1984). By the mid 1980s, the 'standard' CDM model, in which dark matter provides the critical density, Hubble's constant has a value $\sim 50\,\mathrm{km\,s^{-1}Mpc^{-1}}$, and the initial density field was Gaussian with a Harrison-Zel'dovich spectrum, had established itself as the 'best bet' model for structure formation.

In the early 1990s, measurements of galaxy clustering, notably from the APM galaxy survey (Maddox et al., 1990a; Efstathiou et al., 1990) showed that the standard CDM model predicts less clustering on large scales than is observed. Several alternatives were proposed to remedy this. One was a mixed dark matter (MDM) model, in which the universe is flat, with $\sim 30\%$ of the cosmic mass density in HDM and $\sim 70\%$ in CDM and baryons. Another flat model assumed all dark matter to be CDM, but adopted an enhanced radiation background in relativistic neutrinos (τCDM). A third possibility was an open model, in which today's Universe is dominated by CDM and baryons, but has only about 30% of the critical density (OCDM). A final model assumed the same amounts of CDM and baryons as OCDM but added a cosmological constant in order to make the universe flat (ΛCDM).

Although all these models match observed galaxy clustering on large scales, it was soon realized that galaxy formation occurs too late in the MDM and τCDM models, and that the open model has problems in matching the perturbation amplitudes measured by COBE. ΛCDM then became the default 'concordance' model, although it was not generally accepted until Garnavich et al. (1998) and Perlmutter et al. (1999) used the distance-redshift relation of Type Ia supernovae to show that the cosmic expansion is accelerating, and measurements of small-scale CMB fluctuations showed that our Universe is flat (de Bernardis et al., 2000). It seems that the present-day Universe is dominated by a dark energy component with properties very similar to those of Einstein's cosmological constant.

At the beginning of this century, a number of ground-based and balloon-borne experiments measured CMB anisotropies, notably Boomerang (de Bernardis et al., 2000), MAXIMA (Hanany et al., 2000), DASI (Halverson et al., 2002) and CBI (Sievers et al., 2003). They successfully detected features, known as acoustic peaks, in the CMB power spectrum, and showed their wavelengths and amplitudes to be in perfect agreement with expectations for a ΛCDM cosmology. In 2003, the first year data from WMAP not only confirmed these results, but also allowed much

more precise determinations of cosmological parameters. The values obtained were in remarkably good agreement with independent measurements; the baryon density matched that estimated from cosmic nucleosynthesis, the Hubble constant matched that found by direct measurement, the dark-energy density matched that inferred from Type Ia supernovae, and the implied large-scale clustering in today's Universe matched that measured using large galaxy surveys and weak gravitational lensing (see Spergel et al., 2003, and references therein). Consequently, the ΛCDM model has now established itself firmly as the standard paradigm for structure formation. With further data from WMAP and from other sources, the parameters of this new paradigm are now well constrained (Spergel et al., 2007; Komatsu et al., 2009).

1.4.5 Galaxy Formation

(a) Monolithic Collapse and Merging Although it was well established in the 1930s that there are two basic types of galaxies, ellipticals and spirals, it would take some 30 years before detailed models for their formation were proposed. In 1962, Eggen, Lynden-Bell & Sandage considered a model in which galaxies form from the collapse of gas clouds, and suggested that the difference between ellipticals and spirals reflects the rapidity of star formation during the collapse. If most of the gas turns into stars as it falls in, the collapse is effectively dissipationless and infall motions are converted into the random motion of stars, resulting in a system which might resemble an elliptical galaxy. If, on the other hand, the cloud remains gaseous during collapse, the gravitational energy can be effectively dissipated via shocks and radiative cooling. In this case, the cloud will shrink until it is supported by angular momentum, leading to the formation of a rotationally-supported disk. Gott & Thuan (1976) took this picture one step further and suggested that the amount of dissipation during collapse depends on the amplitude of the initial perturbation. Based on the empirical fact that star formation efficiency appears to scale as ρ^2 (Schmidt, 1959), they argued that protogalaxies associated with the highest initial density perturbations would complete star formation more rapidly as they collapse, and so might produce an elliptical. On the other hand, protogalaxies associated with lower initial density perturbations would form stars more slowly and so might make spirals.

Larson (1974a,b, 1975, 1976) carried out the first numerical simulations of galaxy formation, showing how these ideas might work in detail. Starting from near-spherical rotating gas clouds, he found that it is indeed the ratio of the star-formation time to the dissipation/cooling time which determines whether the system turns into an elliptical or a spiral. He also noted the importance of feedback effects during galaxy formation, arguing that in low mass galaxies, supernovae would drive winds that could remove most of the gas and heavy elements from a system before they could turn into stars. He argued that this mechanism might explain the low surface brightnesses and low metallicities of dwarf galaxies. However, he was unable to obtain the high observed surface brightnesses of bright elliptical galaxies without requiring his gas clouds to be much more slowly rotating than predicted by the tidal torque theory; otherwise they would spin up and make a disk long before they became as compact as the observed galaxies. The absence of highly flattened ellipticals and the fact that many bright ellipticals show little or no rotation (Bertola & Capaccioli, 1975; Illingworth, 1977) therefore posed a serious problem for this scenario. As we now know, its main defect was that it left out the effects of the dark matter.

In a famous 1972 paper, Toomre & Toomre used simple numerical simulations to demonstrate convincingly that some of the extraordinary structures seen in peculiar galaxies, such as long tails, could be produced by tidal interactions between two normal spirals. Based on the observed frequency of galaxies with such signatures of interactions, and on their estimate of the time scale over which tidal tails might be visible, Toomre & Toomre (1972) argued that most elliptical galaxies could be merger remnants. In an extreme version of this picture, all galaxies initially form as disks, while all ellipticals are produced by mergers between pre-existing galaxies. A

virtue of this idea was that almost all known star formation occurs in disk gas. Early simulations showed that the merging of two spheroids produces remnants with density profiles that agree with observed ellipticals (e.g., White, 1978). The more relevant (but also the more difficult) simulations of mergers between disk galaxies were not carried out until the early 1980s (Gerhard, 1981; Farouki & Shapiro, 1982; Negroponte & White, 1983; Barnes, 1988). These again showed merger remnants to have properties similar to those of observed ellipticals.

Although the merging scenario fits nicely into a hierarchical formation scheme, where larger structures grow by mergers of smaller ones, the extreme picture outlined above has some problems. Ostriker (1980) pointed out that observed giant ellipticals, which are dense and can have velocity dispersions as high as $\sim 300\,\mathrm{km\,s^{-1}}$, could not be formed by mergers of present-day spirals, which are more diffuse and almost never have rotation velocities higher than $300\,\mathrm{km\,s^{-1}}$. As we will see below, this problem may be resolved by considering the dark halos of the galaxies, and by recognizing that the high redshift progenitors of ellipticals were more compact than present-day spirals. The merging scenario remains a popular scenario for the formation of (bright) elliptical galaxies.

(b) The Role of Radiative Cooling An important question for galaxy formation theory is why galaxies with stellar masses larger $\sim 10^{12}\,\mathrm{M_\odot}$ are absent or extremely rare. In the adiabatic model, this mass scale is close to the Silk damping scale and could plausibly set a *lower* limit to galaxy masses. However, in the presence of dark matter Silk damping leaves no imprint on the properties of galaxies, simply because the dark matter perturbations are not damped. Press & Schechter (1974) showed that there is a characteristic mass also in the hierarchical model, corresponding to the mass scale of the typical non-linear object at the present time. However, this mass scale is relatively large, and many objects with mass above $10^{12}\,\mathrm{M_\odot}$ are predicted, and indeed are observed as virialized groups and clusters of galaxies. Apparently, the mass scale of galaxies is not set by gravitational physics alone.

In the late 1970s, Silk (1977), Rees & Ostriker (1977) and Binney (1977) suggested that radiative cooling might play an important role in limiting the mass of galaxies. They argued that galaxies can form effectively only in systems where the cooling time is comparable to or shorter than the collapse time, which leads to a characteristic scale of $\sim 10^{12}\,\mathrm{M_\odot}$, similar to the mass scale of massive galaxies. They did not explain why a typical galaxy should form with a mass near this limit, nor did they explicitly consider the effects of dark matter. Although radiative cooling plays an important role in all current galaxy formation theories, it is still unclear if it alone can explain the characteristic mass scale of galaxies, or whether various feedback processes must also be invoked.

(c) Galaxy Formation in Dark Matter Halos By the end of the 1970s, several lines of argument had led to the conclusion that dark matter must play an important role in galaxy formation. In particular, observations of rotation curves of spiral galaxies indicated that these galaxies are embedded in dark halos which are much more extended than the galaxies themselves. This motivated White & Rees (1978) to propose a two-stage theory for galaxy formation; dark halos form first through hierarchical clustering, the luminous content of galaxies then results from cooling and condensation of gas within the potential wells provided by these dark halos. The mass function of galaxies was calculated by applying these ideas within the Press & Schechter model for the growth of non-linear structure. The model of White and Rees contains many of the basic ideas of the modern theory of galaxy formation. They noticed that feedback is required to explain the low overall efficiency of galaxy formation, and invoked Larson's (1974a) model for supernova feedback in dwarf galaxies to explain this. They also noted, but did not emphasize, that even with strong feedback, their hierarchical model predicts a galaxy luminosity function with far too many faint galaxies. This problem is alleviated but not solved by adopting CDM initial conditions rather than the simple power-law initial conditions they adopted. In 1980, Fall

& Efstathiou developed a model of disk formation in dark matter halos, incorporating the angular momentum expected from tidal torques, and showed that many properties of observed disk galaxies can be understood in this way.

Many of the basic elements of galaxy formation in the CDM scenario were already in place in the early 1980s, and were summarized nicely by Efstathiou & Silk (1983) and in Blumenthal et al. (1984). Blumenthal et al. invoked the idea of biased galaxy formation, suggesting that disk galaxies may be associated with density peaks of typical heights in the CDM density field, while giant ellipticals may be associated with higher density peaks. Efstathiou & Silk (1983) discussed in some detail how the two-stage theory of White & Rees (1978) can solve some of the problems in earlier models based on the collapse of gas clouds. In particular, they argued that, within an extended halo, cooled gas can settle into a rotation-supported disk of the observed scale in a fraction of the Hubble time, whereas without a dark matter halo it would take too long for a perturbation to turn around and shrink to form a disk (see Chapter **??** for details). They also argued that extended dark matter halos around galaxies make mergers of galaxies more likely, a precondition for Toomre & Toomre's merger scenario of elliptical galaxy formation to be viable.

Since the early 1990s many studies have investigated the properties of CDM halos using both analytical and N-body methods. Properties studied include the progenitor mass distributions (Bond et al., 1991), merger histories (Lacey & Cole, 1993), spatial clustering (Mo & White, 1996), density profiles (Navarro et al., 1997), halo shapes (e.g., Jing & Suto, 2002), substructure (e.g., Moore et al., 1998; Klypin et al., 1999), and angular-momentum distributions (e.g., Warren et al., 1992; Bullock et al., 2001). These results have paved the way for more detailed models for galaxy formation within the CDM paradigm. In particular, two complementary approaches have been developed: semi-analytical models and hydrodynamical simulations. The semi-analytical approach, originally developed by White & Frenk (1991) and subsequently refined in a number of studies (e.g., Kauffmann et al., 1993; Cole et al., 1994; Dalcanton et al., 1997; Mo et al., 1998; Somerville & Primack, 1999), uses knowledge about the structure and assembly history of CDM halos to model the gravitational potential wells within which galaxies form and evolve, treating all the relevant physical processes (cooling, star formation, feedback, dynamical friction, etc.) in a semi-analytical fashion. The first three-dimensional, hydrodynamical simulations of galaxy formation including dark matter were carried out by Katz in the beginning of the 1990s (Katz & Gunn, 1991; Katz, 1992) and focused on the collapse of a homogeneous, uniformly rotating sphere. The first simulation of galaxy formation by hierarchical clustering from proper cosmological initial conditions was that of Navarro & Benz (1991), while the first simulation of galaxy formation from CDM initial conditions was that of Navarro & White (1994). Since then, numerical simulations of galaxy formation with increasing numerical resolution have been carried out by many authors.

It is clear that the CDM scenario has become the preferred scenario for galaxy formation, and we have made a great deal of progress in our quest towards understanding the structure and formation of galaxies within it. However, as we will see later in this book, there are still many important unsolved problems. It is precisely the existence of these outstanding problems that makes galaxy formation such an interesting subject. It is our hope that this book will help you to equip yourself for your own explorations in this area.

2
Observational Facts

Observational astronomy has developed at an extremely rapid pace. Until the end of the 1940s observational astronomy was limited to optical wavebands. Today we can observe the Universe at virtually all wavelengths covering the electromagnetic spectrum, either from the ground or from space. Together with the revolutionary growth in computer technology and with a dramatic increase in the number of professional astronomers, this has led to a flood of new data. Clearly it is impossible to provide a complete overview of all this information in a single chapter (or even in a single book). Here we focus on a number of selected topics relevant to our forthcoming discussion, and limit ourselves to a simple description of some of the available data. Discussion regarding the interpretation and/or implication of the data is postponed to chapters **??** - **??**, where we use the physical ingredients described in chapters **??** - **??** to interpret the observational results presented here. After a brief introduction of observational techniques, we present an overview of some of the observational properties of stars, galaxies, clusters and groups, large scale structure, the intergalactic medium, and the cosmic microwave background. We end with a brief discussion of cosmological parameters and the matter/energy content of the Universe.

2.1 Astronomical Observations

Almost all information we can obtain about an astronomical object is derived from the radiation we receive from it, or by the absorption it causes in the light of a background object. The radiation from a source may be characterized by its spectral energy distribution (SED), $f_\lambda \, \mathrm{d}\lambda$, which is the total energy of emitted photons with wavelengths in the range λ to $\lambda + \mathrm{d}\lambda$. Technology is now available to detect electromagnetic radiation over an enormous energy range, from low frequency radio waves to high energy gamma rays. However, from the Earth's surface our ability to detect celestial objects is seriously limited by the transparency of our atmosphere. Fig. 2.1 shows the optical depth for photon transmission through the Earth's atmosphere as a function of photon wavelength, along with the wavelength ranges of some commonly used wavebands. Only a few relatively clear windows exist in the optical, near-infrared and radio bands. In other parts of the spectrum, in particular the far-infrared, ultraviolet, X-ray and gamma-ray regions, observations can only be carried out by satellites or balloon-borne detectors.

Although only a very restricted range of frequencies penetrate our atmosphere, celestial objects actually emit over the full range accessible to our instruments. This is illustrated in Fig. 2.2, a schematic representation of the average brightness of the sky as a function of wavelength as seen from a vantage point well outside our own galaxy. With the very important exception of the Cosmic Microwave Background (CMB), which dominates the overall photon energy content of the Universe, the dominant sources of radiation at all energies below the hard gamma-ray regime are related to galaxies, their evolution, their clustering and their nuclei. At radio, far-UV, X-ray and soft gamma-ray wavelengths the emission comes primarily from active galactic nuclei.

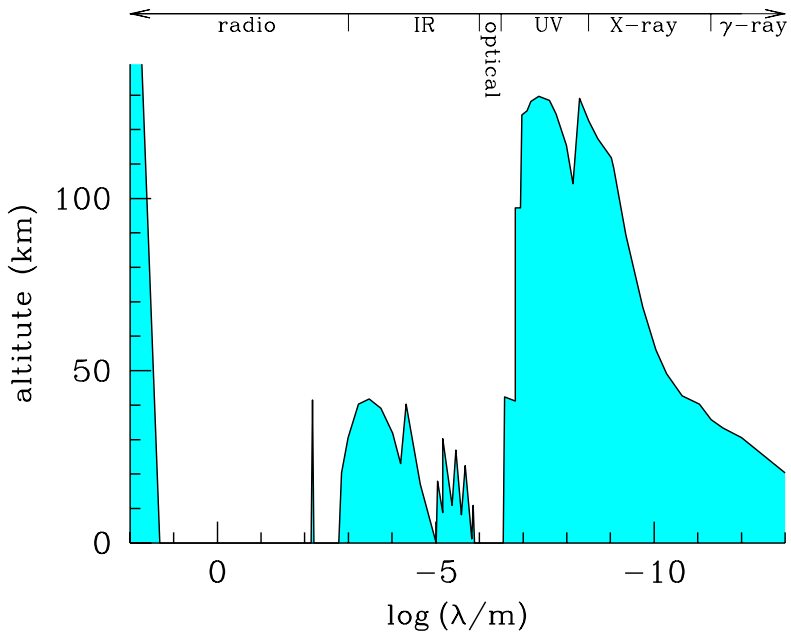


Fig. 2.1. The altitude above sea level at which a typical photon is absorbed as a function of the photon's wavelength. Only radio waves, optical light, the hardest γ-ray, and infrared radiation in a few wavelength windows can penetrate the atmosphere to reach sea level. Observations at all other wavebands have to be carried out above the atmosphere.

Galactic starlight dominates in the near-UV, optical and near-infrared, while dust emission from star-forming galaxies is responsible for most of the far-infrared emission. The hot gas in galaxy clusters emits a significant but non-dominant fraction of the total X-ray background and is the only major source of emission from scales larger than an individual galaxy. Such large structures can, however, be seen in absorption, for example in the light of distant quasars.

2.1.1 Fluxes and Magnitudes

The image of an astronomical object reflects its surface brightness distribution. The surface brightness is defined as the photon energy received by a unit area at the observer per unit time from a unit solid angle in a specific direction. Thus if we denote the surface brightness by I, its units are $[I] = \mathrm{erg\,s^{-1}\,cm^{-2}\,sr^{-1}}$. If we integrate the surface brightness over the entire image, we obtain the flux of the object, f, which has units $[f] = \mathrm{erg\,s^{-1}\,cm^{-2}}$. Integrating the flux over a sphere centered on the object and with radius equal to the distance r from the object to the observer, we obtain the bolometric luminosity of the object:

$$L = 4\pi r^2 f\,, \tag{2.1}$$

with $[L] = \mathrm{erg\,s^{-1}}$. For the Sun, $L = 3.846 \times 10^{33}\,\mathrm{erg\,s^{-1}}$.

The image size of an extended astronomical object is usually defined on the basis of its isophotal contours (curves of constant surface brightness), and the characteristic radius of an isophotal contour at some chosen surface brightness level is usually referred to as an isophotal radius of the object. A well known example is the Holmberg radius defined as the length of the semi-major axis of the isophote corresponding to a surface brightness of $26.5\,\mathrm{mag\,arcsec^{-2}}$ in the B-band. Two other commonly used size measures in optical astronomy are the core radius, defined as the

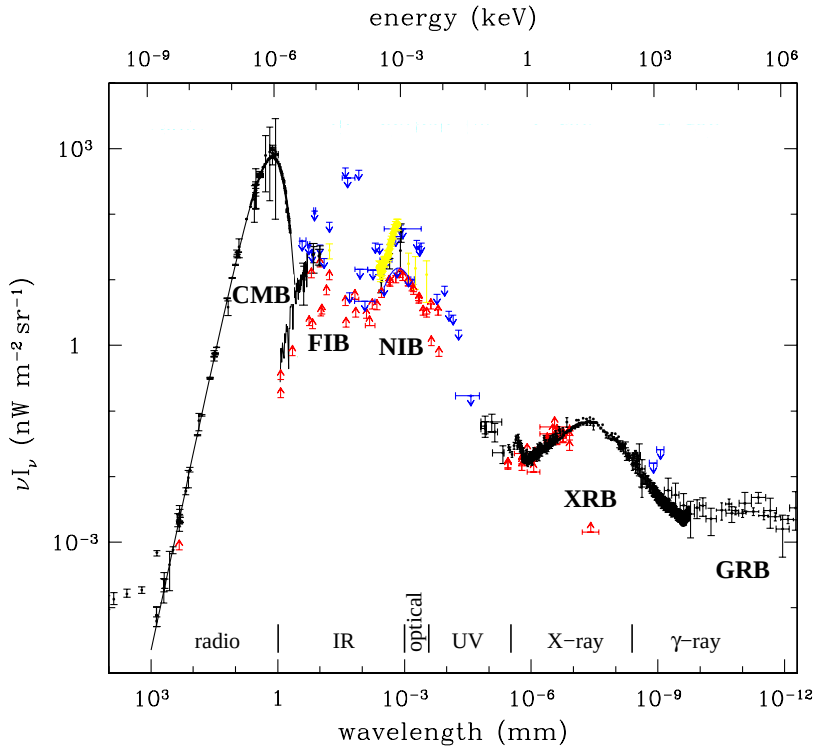


Fig. 2.2. The energy density spectrum of cosmological background radiation as a function of wavelength. The value of νI_ν measures the radiation power per decade of wavelength. This makes it clear that the cosmic microwave background (CMB) contributes most to the overall background radiation, followed by the far- (FIB) and near-infrared (NIB) backgrounds, the X-ray background (XRB) and the γ-ray background (GRB). [Courtesy of D. Scott, see Scott (2000)]

radius where the surface brightness is half of the central surface brightness, and the half-light radius (also called the effective radius), defined as the characteristic radius that encloses half of the total observed flux. For an object at a distance r, its physical size, D, is related to its angular size, θ, by

$$D = r\theta\,. \tag{2.2}$$

Note, though, that relations (2.1) and (2.2) are only valid for relatively small distances. As we will see in Chapter **??**, for objects at cosmological distances, r in Eqs. (2.1) and (2.2) has to be replaced by the luminosity distance and angular diameter distance, respectively.

(a) Wavebands and Bandwidths Photometric observations are generally carried out in some chosen waveband. Thus, the observed flux from an object is related to its SED, f_λ, by

$$f_X = \int f_\lambda F_X(\lambda) R(\lambda) T(\lambda)\,\mathrm{d}\lambda\,. \tag{2.3}$$

Here $F_X(\lambda)$ is the transmission of the filter that defines the waveband (denoted by X), $T(\lambda)$ represents the atmospheric transmission, and $R(\lambda)$ represents the efficiency with which the telescope plus instrument detects photons. In the following we will assume that f_X has been corrected for atmospheric absorption and telescope efficiency (the correction is normally done by calibrating the data using standard objects with known f_λ). In this case, the observed flux depends only

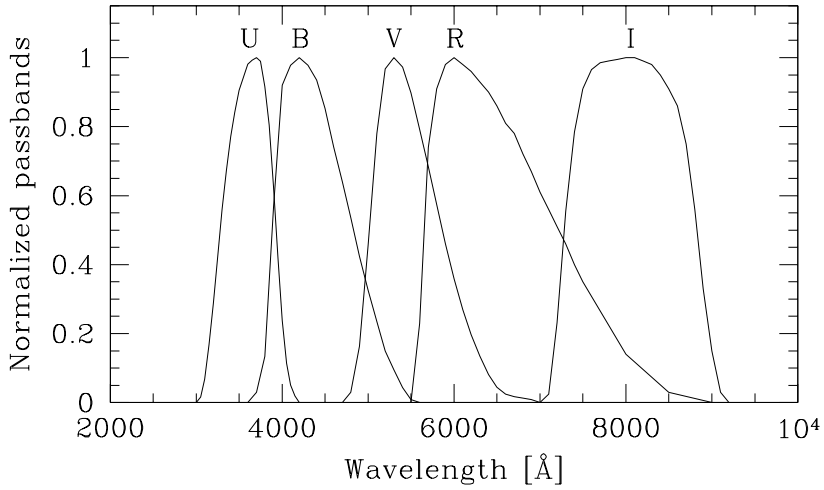


Fig. 2.3. The transmission characteristics of Johnson *UBV* and Kron Cousins *RI* filter systems. [Based on data published in Bessell (1990)]

Table 2.1. *Filter Characteristics of the UBVRI Photometric System.*

Band:	*U*	*B*	*V*	*R*	*I*	*J*	*H*	*K*	*L*	*M*
$\lambda_{\rm eff}$ (nm):	365	445	551	658	806	1220	1630	2190	3450	4750
FWHM (nm):	66	94	88	138	149	213	307	390	472	460
$\mathscr{M}_\odot$:	5.61	5.48	4.83	4.42	4.08	3.64	3.32	3.28	3.25	–
$L_\odot(10^{32}\,{\rm erg/s})$:	1.86	4.67	4.64	6.94	4.71	2.49	1.81	0.82	0.17	–

on the spectral energy distribution and the chosen filter. Astronomers have constructed a variety of photometric systems. A well known example is the standard *UBV* system originally introduced by Johnston. The filter functions for this system are shown in Fig. 2.3. In general, a filter function can be characterized by an effective wavelength, $\lambda_{\rm eff}$, and a characteristic bandwidth, usually quoted as a full width at half maximum (FWHM). The FWHM is defined as $|\lambda_1 - \lambda_2|$, with $F_X(\lambda_1) = F_X(\lambda_2) =$ half the peak value of $F_X(\lambda)$. Table 2.1 lists $\lambda_{\rm eff}$ and the FWHM for the filters of the standard *UBVRI* photometric system. In this system, the FWHM are all of order 10% or larger of the corresponding $\lambda_{\rm eff}$. Such 'broad-band photometry' can be used to characterize the overall shape of the spectral energy distribution of an object with high efficiency. Alternatively, one can use 'narrow-band photometry' with much narrower filters to image objects in a particular emission line or to study its detailed SED properties.

(b) Magnitude and Color For historical reasons, the flux of an astronomical object in the optical band (and also in the near infrared and near ultraviolet bands) is usually quoted in terms of apparent magnitude:

$$m_X = -2.5\log(f_X/f_{X,0})\,, \tag{2.4}$$

where the flux zero-point $f_{X,0}$ has traditionally been taken as the flux in the X band of the bright star Vega. In recent years it has become more common to use 'AB-magnitudes', for which

$$f_{X,0} = 3.6308\times10^{-20}{\rm erg\,s^{-1}cm^{-2}Hz^{-1}}\int F_X(c/\nu){\rm d}\nu\,. \tag{2.5}$$

Here ν is the frequency and c is the speed of light. Similarly, the luminosities of objects (in waveband X) are often quoted as an absolute magnitude: $\mathscr{M}_X = -2.5\log(L_X) + C_X$, where C_X is a zero point. It is usually convenient to write L_X in units of the solar luminosity in the same band, $L_{\odot X}$. The values of $L_{\odot X}$ in the standard *UBVRI* photometric system are listed in Table 2.1. It then follows that

$$\mathscr{M}_X = -2.5\log\left(\frac{L_X}{L_{\odot X}}\right) + \mathscr{M}_{\odot X}\,, \tag{2.6}$$

where $\mathscr{M}_{\odot X}$ is the absolute magnitude of the Sun in the waveband in consideration. Using Eq. (2.1), we have

$$m_X - \mathscr{M}_X = 5\log(r/r_0)\,, \tag{2.7}$$

where r_0 is a fiducial distance at which m_X and $\mathscr{M}_X$ are defined to have the same value. Conventionally, r_0 is chosen to be 10 pc (1 pc = 1 parsec = 3.0856×10^{18}cm; see §2.1.3 for a definition). According to this convention, the Vega absolute magnitudes of the Sun in the *UBVRI* photometric system have the values listed in Table 2.1.

The quantity $(m_X - \mathscr{M}_X)$ for an astronomical object is called its distance modulus. If we know both m_X and $\mathscr{M}_X$ for an object, then Eq. (2.7) can be used to obtain its distance. Conversely, if we know the distance to an object, a measurement of its apparent magnitude (or flux) can be used to obtain its absolute magnitude (or luminosity).

Optical astronomers usually express surface brightness in terms of magnitudes per square arcsecond. In such "units", the surface brightness in a band X is denoted by μ_X, and is related to the surface brightness in physical units, I_X, according to

$$\mu_X = -2.5\log\left(\frac{I_X}{\mathrm{L}_\odot\,\mathrm{pc}^{-2}}\right) + 21.572 + \mathscr{M}_{\odot,X}\,. \tag{2.8}$$

Note that it is the flux, not the magnitude, that is additive. Thus in order to obtain the total (apparent) magnitude from an image, one must first convert magnitude per unit area into flux per unit area, integrate the flux over the entire image, and then convert the total flux back to a total magnitude.

If observations are made for an object in more than one waveband, then the difference between the magnitudes in any two different bands defines a color index (which corresponds to the slope of the SED between the two wavebands). For example,

$$(B-V) \equiv m_B - m_V = \mathscr{M}_B - \mathscr{M}_V \tag{2.9}$$

is called the $(B-V)$ color of the object.

2.1.2 Spectroscopy

From spectroscopic observations one obtains spectra for objects, i.e. their SEDs f_λ or f_ν defined so that $f_\lambda\,\mathrm{d}\lambda$ and $f_\nu\,\mathrm{d}\nu$ are the fluxes received in the elemental wavelength and frequency ranges $\mathrm{d}\lambda$ at λ and $\mathrm{d}\nu$ at ν. From the relation between wavelength and frequency, $\lambda = c/\nu$, we then have that

$$f_\nu = \lambda^2 f_\lambda / c \quad \text{and} \quad f_\lambda = \nu^2 f_\nu / c\,. \tag{2.10}$$

At optical wavelengths, spectroscopy is typically performed by guiding the light from an object to a spectrograph where it is dispersed according to wavelength. For example, in multi-object fiber spectroscopy, individual objects are imaged onto the ends of optical fibers which take the light to prism or optical grating where it is dispersed. The resulting spectra for each individual fiber are then imaged on a detector. Such spectroscopy loses all information about the distribution of each object's light within the circular aperture represented by the end of the fiber. In long-slit

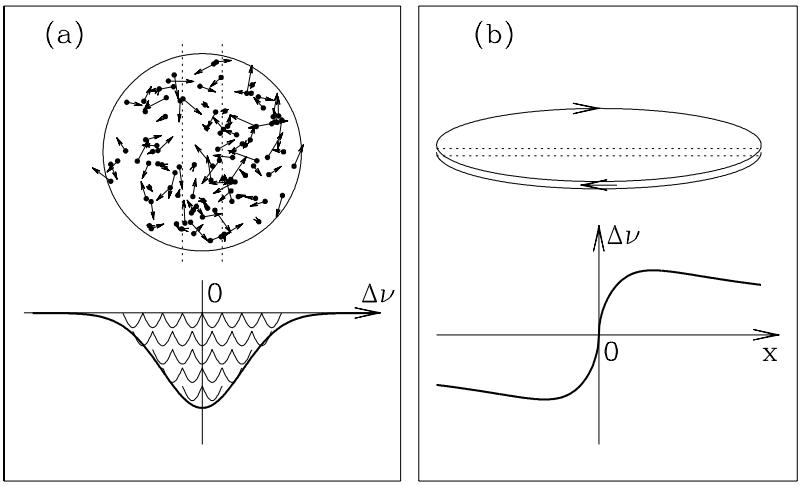


Fig. 2.4. (a) An illustration of the broadening of a spectral line by the velocity dispersion of stars in a stellar system. A telescope collects light from all stars within a cylinder through the stellar system. Each star contributes a narrow spectral line with rest frequency ν_{12}, which is Doppler shifted to a different frequency $\nu = \nu_{12} + \Delta\nu$ due to its motion along the line of sight. The superposition of many such line profiles produces a broadened line, with the profile given by the convolution of the original stellar spectral line and the velocity distribution of the stars in the cylinder. (b) An illustration of long-slit spectroscopy of a thin rotating disk along the major axis of the image. In the plot, the rotation speed is assumed to depend on the distance from the center as $V_{\rm rot}(x) \propto \sqrt{x/(1+x^2)}$.

spectroscopy, on the other hand, the object of interest is imaged directly onto the spectrograph slit, resulting in a separate spectrum from each point of the object falling on the slit. Finally, in an integral field unit (or IFU) the light from each point within the image of an extended object is led to a different point on the slit (for example, by optical fibers) resulting in a three-dimensional data cube with two spatial dimensions and one dimension for the wavelength.

At other wavelengths quite different techniques can be used to obtain spectral information. For example, at infrared and radio wavelengths the incoming signal from a source may be Fourier analyzed in time in order to obtain the power at each frequency, while at X-ray wavelengths the energy of each incoming photon can be recorded and the energies of different photons can be binned to obtain the spectrum.

Spectroscopic observations can give us a lot of information which photometric observations cannot. A galaxy spectrum usually contains a slowly-varying component called the continuum, with localized features produced by emission and absorption lines (see Fig. 2.12 for some examples). It is a superposition of the spectra of all the individual stars in the galaxy, modified by emission and absorption from the gas and dust lying between the stars. From the ultraviolet through the near-infrared the continuum is due primarily to bound-free transitions in the photospheres of the stars, in the mid- and far-infrared it is dominated by thermal emission from dust grains, in the radio it is produced by diffuse relativistic and thermal electrons within the galaxy, and in the X-ray it comes mainly from accretion of gas onto compact stellar remnants or a central black hole. Emission and absorption lines are produced by bound-bound transitions within atoms, ions and molecules, both in the outer photospheres of stars and in the interstellar gas. By analyzing a spectrum, we may infer the relative importance of these various processes, thereby understanding the physical properties of the galaxy. For example, the strength of a particular emission line depends on the abundance of the excited state that produces it, which in turn de-

pends not only on the abundance of the corresponding element but also on the temperature and ionization state of the gas. Thus emission line strengths can be used to measure the temperature, density and chemical composition of interstellar gas. Absorption lines, on the other hand, mainly arise in the atmospheres of stars, and their relative strengths contain useful information regarding the age and metallicity of the galaxy's stellar population. Finally, interstellar dust gives rise to continuum absorption with broad characteristic features. In addition, since dust extinction is typically more efficient at shorter wavelengths, it also causes reddening, a change of the overall slope of the continuum emission.

Spectroscopic observations have another important application. The intrinsic frequency of photons produced by electron transitions between two energy levels E_1 and E_2 is $\nu_{12} = (E_2 - E_1)/h_{\rm P}$, where $h_{\rm P}$ is Planck's constant, and we have assumed $E_2 > E_1$. Now suppose that these photons are produced by atoms moving with velocity $\mathbf{v}$ relative to the observer. Because of the Doppler effect, the observed photon frequency will be (assuming $v \ll c$),

$$\nu_{\rm obs} = \left(1 - \frac{\mathbf{v}\cdot\hat{\mathbf{r}}}{c}\right)\nu_{12}\,, \tag{2.11}$$

where $\hat{\mathbf{r}}$ is the unit vector of the emitting source relative to the observer. Thus, if the source is receding from the observer, the observed frequency is redshifted, $\nu_{\rm obs} < \nu_{12}$; conversely, if the source is approaching the observer, the observed frequency is blueshifted, $\nu_{\rm obs} > \nu_{12}$. It is convenient to define a redshift parameter to characterize the change in frequency,

$$z \equiv \frac{\nu_{12}}{\nu_{\rm obs}} - 1\,. \tag{2.12}$$

For the Doppler effect considered here, we have $z = \mathbf{v}\cdot\hat{\mathbf{r}}/c$. Clearly, by studying the properties of spectral lines from an object, one may infer the kinematics of the emitting (or absorbing) material.

As an example, suppose that the emitting gas atoms in an object have random motions along the line of sight drawn from a velocity distribution $f(v)\,{\rm d}v$. The observed photons will then have the following frequency distribution:

$$F(\nu_{\rm obs})\,{\rm d}\nu_{\rm obs} = f(v)(c/\nu_{12})\,{\rm d}\nu_{\rm obs}\,, \tag{2.13}$$

where v is related to $\nu_{\rm obs}$ by $v = c(1-\nu_{\rm obs}/\nu_{12})$, and we have neglected the natural width of atomic spectral lines. Thus, by observing $F(\nu_{\rm obs})$ (the emission line profile in frequency space), we can infer $f(v)$. If the random motion is caused by thermal effects, we can infer the temperature of the gas from the observed line profile. For a stellar system (e.g. an elliptical galaxy) the observed spectral line is the convolution of the original stellar line profile $S(\nu)$ (which is a luminosity weighted sum of the spectra of all different stellar types that contribute to the flux) with the line-of-sight velocity distribution of all the stars in the observational aperture,

$$F(\nu_{\rm obs}) = \int S[\nu_{\rm obs}(1+v/c)]\,f(v){\rm d}v\,. \tag{2.14}$$

Thus, each narrow, stellar spectral line is broadened by the line-of-sight velocity dispersion of the stars that contribute to that line (see Fig. 2.4a). If we know the type of stars that dominate the spectral lines in consideration, we can estimate $S(\nu)$ and use the above relation to infer the properties of $f(v)$, such as the mean velocity, $\overline{v} = \int v f(v)\,{\rm d}v$, and the velocity dispersion, $\sigma = [\int (v-\overline{v})^2 f(v)\,{\rm d}v]^{1/2}$.

Similarly, long-slit and IFU spectroscopy of extended objects can be used not only to study random motions along each line-of-sight through the source, but also to study large-scale flows in the source. An important example here is the rotation of galaxy disks. Suppose that the rotation of a disk around its axis is specified by a rotation curve, $V_{\rm rot}(R)$, which gives the rotation velocity as a function of distance to the disk center. Suppose further that the inclination angle between

the rotation axis and the line-of-sight is i. If we put a long slit along the major axis of the image of the disk, it is easy to show that the frequency shift along the slit is

$$\nu_{\rm obs}(R) - \nu_{12} = \pm \frac{V_{\rm rot}(R)\sin i}{c}\nu_{12}\,, \tag{2.15}$$

where the $+$ and $-$ signs correspond to points on opposite sides of the disk center (see Fig. 2.4b). Thus the rotation curve of the disk can be measured from its long slit spectrum and from its apparent shape (which allows the inclination angle to be estimated under the assumption that the disk is intrinsically round).

2.1.3 Distance Measurements

A fundamental task in astronomy is the determination of the distances to astronomical objects. As we have seen above, the direct observables from an astronomical object are its angular size on the sky and its energy flux at the position of the observer. Distance is therefore required in order to convert these observables into physical quantities. In this subsection we describe the principles behind some of the most important methods for estimating astronomical distances.

(a) Trigonometric Parallax The principle on which this distance measure is based is very simple. We are all familiar with the following: when walking along one direction, nearby and distant objects appear to change their orientation with respect to each other. If the walked distance b is much smaller than the distance to an object d (assumed to be perpendicular to the direction of motion), then the change of the orientation of the object relative to an object at infinity is $\theta = b/d$. Thus, by measuring b and θ we can obtain the distance d. This is called the trigonometric parallax method, and can be used to measure distances to some relatively nearby stars. In principle, this can be done by measuring the change of the position of a star relative to one or more background objects (assumed to be at infinity) at two different locations. Unfortunately, the baseline provided by the Earth's diameter is so short that even the closest stars do not have a measurable trigonometric parallax. Therefore, real measurements of stellar trigonometric parallax have to make use of the baseline provided by the diameter of the Earth's orbit around the Sun. By measuring the trigonometric parallax, π_t, which is half of the angular change in the position of a star relative to the background as measured over a six month interval, we can obtain the distance to the star as

$$d = \frac{A}{\tan(\pi_t)}\,, \tag{2.16}$$

where $A = 1\,{\rm AU} = 1.49597870 \times 10^{13}{\rm cm}$ is the length of the semi-major axis of the Earth's orbit around the Sun. The distance corresponding to a trigonometric parallax of 1 arcsec is defined as 1 parsec (or 1 pc). From the Earth the accuracy with which π_t can be measured is restricted by atmospheric seeing, which causes a blurring of the images. This problem is circumvented when using satellites. With the Hipparcos Satellite reliable distances have been measured for nearby stars with $\pi_t \gtrsim 10^{-3}$ arcsec, or with distances $d \lesssim 1\,{\rm kpc}$. The GAIA satellite, which is currently scheduled for launch in 2012, will be able to measure parallaxes for stars with an accuracy of $\sim 2\times10^{-4}$ arcsec, which will allow distance measurements to 10 percent accuracy for $\sim 2\times10^{8}$ stars.

(b) Motion-Based Methods The principle of this distance measurement is also very simple. We all know that the angle subtended by an object of diameter l at a distance d is $\theta = l/d$ (assuming $l \ll d$). If we measure the angular diameters of the same object from two distances, d_1 and d_2, then the difference between them is $\Delta\theta = l\Delta d/d^2 = \theta\,\Delta d/d$, where $\Delta d = |d_1 - d_2|$ is assumed to be much smaller than both d_1 and d_2, and $d = (d_1 d_2)^{1/2}$ can be considered the distance to the object. Thus, we can estimate d by measuring $\Delta\theta$ and Δd. For a star cluster

consisting of many stars, the change of its distance over a time interval Δt is given by $\Delta d = v_r \Delta t$, where v_r is the mean radial velocity of the cluster and can be measured from the shift of its spectrum. If we can measure the change of the angular size of the cluster during the same time interval, $\Delta\theta$, then the distance to the cluster can be estimated from $d = \theta\, v_r \Delta t / \Delta\theta$. This is called the moving-cluster method.

Another distance measure is based on the angular motion of cluster stars caused by their velocity with respect to the Sun. If all stars in a star cluster had the same velocity, the extensions of their proper motion vectors would converge to a single point on the celestial sphere (just like the two parallel rails of a railway track appear to converge to a point at large distance). By measuring the proper motions of the stars in a star cluster, this convergent point can be determined. Because of the geometry, the line-of-sight from the observer to the convergent point is parallel to the velocity vector of the star cluster. Hence, the angle, ϕ, between the star cluster and its convergent point, which can be measured, is the same as that between the proper motion vector and its component along the line-of-sight between the observer and the star cluster. By measuring the cluster's radial velocity v_r, one can thus obtain the transverse velocity $v_t = v_r \tan\phi$. Comparing v_t to the proper motion of the star cluster then yields its distance. This is called the convergent-point method and can be used to estimate accurate distances of star clusters up to a few hundred parsec.

(c) Standard Candles and Standard Rulers As shown by Eqs. (2.1) and (2.2), the luminosity and physical size of an object are related through the distance to its flux and angular size, respectively. Since the flux and angular size are directly observable, we can estimate the distance to an object if its luminosity or its physical size can be obtained in a distance-independent way. Objects whose luminosities and physical sizes can be obtained in such a way are called standard candles and standard rulers, respectively. These objects play an important role in astronomy, not only because their distances can be determined, but more importantly, because they can serve as distance indicators to calibrate the relation between distance and redshift, allowing the distances to other objects to be determined from their redshifts, as we will see below.

One important class of objects in cosmic distance measurements is the Cepheid variable stars (or Cepheids for short). These objects are observed to change their apparent magnitudes regularly, with periods ranging from 2 to 150 days. The period is tightly correlated with the star's luminosity, such that

$$\mathscr{M} = -a - b \log P\,, \tag{2.17}$$

where P is the period of light variation in days, and a and b are two constants which can be determined using nearby Cepheids whose distances have been measured using another method. For example, using the trigonometric parallaxes of Cepheids measured with the Hipparcos Satellite, Feast & Catchpole (1997) obtained the following relation between P and the absolute magnitude in the V band: $\mathscr{M}_V = -1.43 - 2.81 \log P$, with a standard error in the zero point of about 0.10 magnitudes (see Madore & Freedman, 1991, for more examples of such calibrations). Once the luminosity-period relation is calibrated, and if it is universally valid, it can be applied to distant Cepheids (whose distances cannot be obtained from trigonometric parallax or proper motion) to obtain their distances from measurements of their variation periods. Since Cepheids are relatively bright, with absolute magnitudes $\mathscr{M}_V \sim -3$, telescopes with sufficiently high spatial resolution, such as the Hubble Space Telescope (HST), allow Cepheid distances to be determined for objects out to $\sim 10\,\mathrm{Mpc}$.

Another important class of objects for distance measurements are Type Ia supernovae (SNIa), which are exploding stars with well-calibrated light profiles. Since these objects can reach peak luminosities up to $\sim 10^{10}\,\mathrm{L}_\odot$ (so that they can outshine an entire galaxy), they can be observed out to cosmological distances of several thousand megaparsecs. Empirically it has been found

that the peak luminosities of SNIa are remarkably similar (e.g., Branch & Tammann, 1992). In fact, there is a small dispersion in peak luminosities, but this has been found to be correlated with the rate at which the luminosity decays and so can be corrected (e.g., Phillips et al., 1999). Thus, one can obtain the *relative* distances to Type Ia supernovae by measuring their light curves. The absolute distances can then be obtained once the absolute values of the light curves of some nearby Type Ia supernovae are calibrated using other (e.g. Cepheid) distances. As we will see in §2.10.1, SNIa play an important role in constraining the large-scale geometry of the Universe.

(d) Redshifts as Distances One of the most important discoveries in modern science was Hubble's (1929) observation that almost all galaxies appear to move away from us, and that their recession velocities increase in direct proportion to their distances from us, $v_r \propto r$. This relation, called the Hubble law, is explained most naturally if the Universe as a whole is assumed to be expanding. If the expansion is homogeneous and isotropic, then the distance between any two objects comoving with the expanding background can be written as $r(t) = a(t)r(t')/a(t')$, where $a(t)$ is a time-dependent scale-factor of the Universe, describing the expansion. It then follows that the relative separation velocity of the objects is

$$v_r = \dot{r} = H(t)r\,, \quad \text{where} \quad H(t) \equiv \dot{a}(t)/a(t)\,. \tag{2.18}$$

This relation applied at the present time gives $v_r = H_0 r$, as observed by Hubble. Since the recession velocity of an object can be measured from its redshift z, the distance to the object simply follows from $r = cz/H_0$ (assuming $v_r \ll c$). In practice, the object under consideration may move relative to the background with some (gravitationally induced) peculiar velocity, $v_{\rm pec}$, so that its observed velocity is the sum of this peculiar velocity along the line-of-sight, $v_{{\rm pec},r}$, and the velocity due to the Hubble expansion:

$$v_r = H_0 r + v_{{\rm pec},r}\,. \tag{2.19}$$

In this case, the redshift is no longer a precise measurement of the distance, unless $v_{{\rm pec},r} \ll H_0 r$. Since for galaxies the typical value for $v_{\rm pec}$ is a few hundred kilometers per second, redshifts can be used to approximate distances for $cz \gg 1000\,{\rm km\,s^{-1}}$.

In order to convert redshifts into distances, we need a value for the Hubble constant, H_0. This can be obtained if the distances to some sufficiently distant objects can be measured independently of their redshifts. As mentioned above, such objects are called distance indicators. For many years, the value of the Hubble constant was very uncertain, with estimates ranging from $\sim 50\,{\rm km\,s^{-1}\,Mpc^{-1}}$ to $\sim 100\,{\rm km\,s^{-1}\,Mpc^{-1}}$ (current constraints on H_0 are discussed in §2.10.1). To parameterize this uncertainty in H_0 it has become customary to write

$$H_0 = 100h\,{\rm km\,s^{-1}\,Mpc^{-1}}\,, \tag{2.20}$$

and to express all quantities that depend on redshift-based distances in terms of the reduced Hubble constant h. For example, distance determinations based on redshifts often contain a factor of h^{-1}, while luminosities based on these distances contain a factor h^{-2}, etc. If these factors are not present, it means that a specific value for the Hubble constant has been assumed, or that the distances were not based on measured redshifts.

2.2 Stars

As we will see in §2.3, the primary visible constituent of most galaxies is the combined light from their stellar population. Clearly, in order to understand galaxy formation and evolution it is important to know the main properties of stars. In Table 2.1 we list some of the photometric properties of the Sun. These, as well as the Sun's mass and radius, $\mathrm{M}_\odot = 2 \times 10^{33}\,{\rm g}$ and $\mathrm{R}_\odot =$

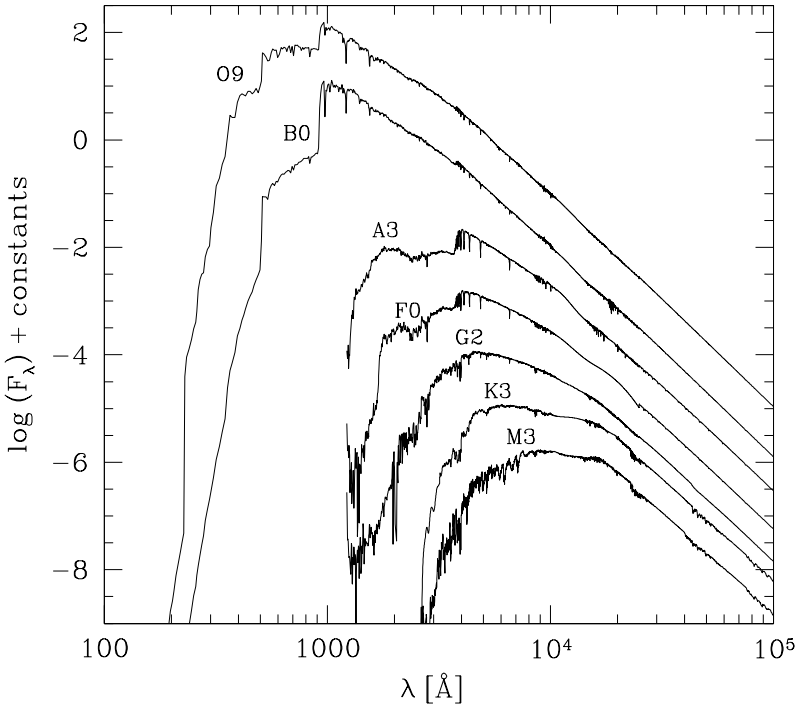


Fig. 2.5. Spectra for stars of different spectral types. f_λ is the flux per angstrom, and an arbitrary constant is added to each spectrum to avoid confusion. [Based on data kindly provided by S. Charlot]

Table 2.2. *Solar Abundances in Number Relative to Hydrogen*

Element:	H	He	C	N	O	Ne	Mg	Si	Fe
$(N/N_{\rm H}) \times 10^5$:	10^5	9800	36.3	11.2	85.1	12.3	3.80	3.55	4.68

Table 2.3. *MK Spectral Classes.*

Class	Temperature	Spectral characteristics
O	28.000-50.000 K	Hot stars with He II absorption; strong UV continuum
B	10.000-28.000 K	He I absorption; H developing in later classes
A	7.500-10.000 K	Strong H lines for A0, decreasing thereafter; Ca II increasing
F	6.000- 7.500 K	Ca II stronger; H lines weaker, metal lines developing
G	5.000- 6.000 K	Ca II strong; metal lines strong; H lines weaker
K	3.500- 5.000 K	Strong metal lines, CH and CN developing; weak blue continuum
M	2.500- 3.500 K	Very red; TiO bands developing strongly

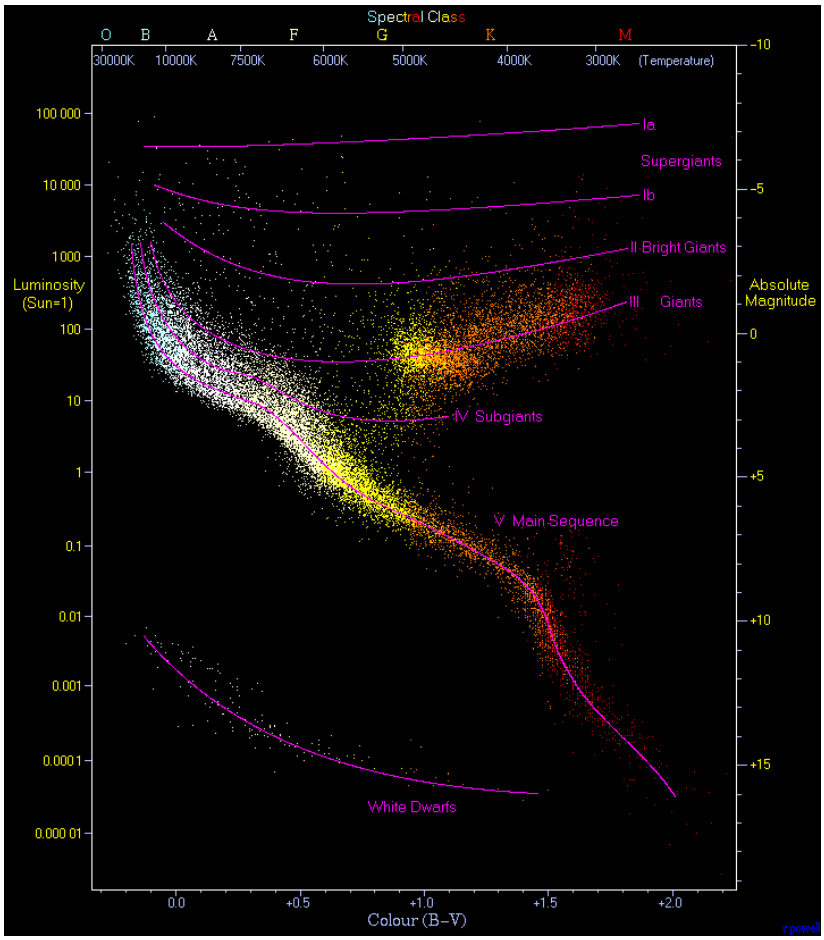


Fig. 2.6. The color-magnitude diagram (i.e. the H-R diagram) of 22000 stars from the Hipparcos Catalogue together with 1000 low-luminosity stars (red and white dwarfs) from the Gliese Catalogue of Nearby Stars. The MK spectral and luminosity classes are also indicated, as are the luminosities in solar units. [Diagram from R. Powell, taken from Wikipedia]

Table 2.4. *MK Luminosity Classes.*

I	Supergiants
II	Bright giants
III	Normal giants
IV	Subgiants
V	Dwarfs (Main Sequence stars)

7×10^{10} cm, are usually used as fiducial values when describing other stars. The abundance by number of some of the chemical elements in the solar system is given in Table 2.2. The fraction in mass of elements heavier than helium is referred to as the metallicity and is denoted by Z, and our Sun has $Z_\odot \approx 0.02$. The relative abundances in a star are usually specified relative to those in the Sun:

$$[A/B] \equiv \log \left[\frac{(n_A/n_B)_\star}{(n_A/n_B)_\odot} \right] , \tag{2.21}$$

where $(n_A/n_B)_\star$ is the number density ratio between element A and element B in the star, and $(n_A/n_B)_\odot$ is the corresponding ratio for the Sun.

Since all stars, except a few nearby ones, are unresolved (i.e., they appear as point sources), the only intrinsic properties that are directly observable are their luminosities, colors and spectra. These vary widely (some examples of stellar spectra are shown in Fig. 2.5) and form the basis for their classification. The most often used classification scheme is the Morgan-Keenan (MK) system, summarized in Tables 2.3 and 2.4. These spectral classes are further divided into decimal subclasses [e.g. from B0 (early) to B9 (late)], while luminosity classes are divided into subclasses such as Ia, Ib etc. The importance of this classification is that, although entirely based on observable properties, it is closely related to the basic physical properties of stars. For example, the luminosity classes are related to surface gravities, while the spectral classes are related to surface temperatures (see e.g. Cox, 2000).

Fig. 2.6 shows the color-magnitude relation of a large number of stars for which accurate distances are available (so that their absolute magnitudes can be determined). Such a diagram is called a Hertzsprung-Russell diagram (abbreviated as H-R diagram), and features predominantly in studies of stellar astrophysics. The MK spectral and luminosity classes are also indicated. Clearly, stars are not uniformly distributed in the color-magnitude space, but lie in several well-defined sequences. Most of the stars lie in the 'main sequence' (MS) which runs from the lower-right to the upper-left. Such stars are called main-sequence stars and have MK luminosity class V. The position of a star in this sequence is mainly determined by its mass. Above the main sequence one finds the much rarer but brighter giants, making up the MK luminosity classes I to IV, while the lower-left part of the H-R diagram is occupied by white dwarfs. The Sun, whose MK type is G2V, lies in the main sequence with V-band absolute magnitude 4.8 and (atmospheric) temperature 5780K.

As a star ages it moves off the MS and starts to traverse the H-R diagram. The location of a star in the H-R diagram as function of time is called its evolutionary track which, again, is determined mainly by its mass. An important property of a stellar population is therefore its initial mass function (IMF), which specifies the abundance of stars as function of their initial mass (i.e., the mass they have at the time when reach the MS shortly after their formation). For a given IMF, and a given star formation history, one can use the evolutionary tracks to predict the abundance of stars in the H-R diagram. Since the spectrum of a star is directly related to its position in the H-R diagram, this can be used to predict the spectrum of an entire galaxy, a procedure which is called spectral synthesis modeling. Detailed calculations of stellar evolution models (see Chapter **??**) show that a star like our Sun has a MS lifetime of about 10 Gyr, and that the MS lifetime scales with mass roughly as M^{-3}, i.e., more massive (brighter) stars spend less time on the MS. This strong dependence of MS lifetime on mass has important observational consequences, because it implies that the spectrum of a stellar system (a galaxy) depends on its star formation history. For a system where the current star formation rate is high, so that many young massive stars are still on the main sequence, the stellar spectrum is expected to have a strong blue continuum produced by O and B stars. On the other hand, for a system where star formation has been terminated a long time ago, so that all massive stars have already evolved off

Table 2.5. *Galaxy Morphological Types.*

Hubble	E	E-SO	SO	SO-Sa	Sa	Sa-b	Sb	Sb-c	Sc	Sc-Irr	Irr
deV	E	SO^-	SO^0	SO^+	Sa	Sab	Sb	Sbc	Scd	Sdm	Im
T	-5	-3	-2	0	1	2	3	4	6	8	10

the MS, the spectrum (now dominated by red giants and the low-mass MS stars) is expected to be red.

2.3 Galaxies

Galaxies, whose formation and evolution is the main topic of this book, are the building blocks of the Universe. They not only are the cradles for the formation of stars and metals, but also serve as beacons that allow us to probe the geometry of space-time. Yet, it is easy to forget that it was not until the 1920's, with Hubble's identification of Cepheid variable stars in the Andromeda nebula, that most astronomers became convinced that the many 'nebulous' objects cataloged by John Dreyer in his 1888 *New General Catalogue of Nebulae and Clusters of Stars* and the two supplementary *Index Catalogues* are indeed galaxies. Hence, extra-galactic astronomy is a relatively new science. Nevertheless, as we will see, we have made tremendous progress: we have surveyed the local population of galaxies in exquisite detail covering the entire range of wavelengths, we have constructed redshift surveys with hundreds of thousands of galaxies to probe the large scale structure of the Universe, and we have started to unveil the population of galaxies at high redshifts, when the Universe was only a small fraction of its current age.

2.3.1 The Classification of Galaxies

Fig. 2.7 shows a collage of images of different kinds of galaxies. Upon inspection, one finds that some galaxies have smooth light profiles with elliptical isophotes, others have spiral arms together with an elliptical-like central bulge, and still others have irregular or peculiar morphologies. Based on such features, Hubble ordered galaxies in a morphological sequence, which is now referred to as the Hubble sequence or Hubble tuning-fork diagram (see Fig. 2.8). Hubble's scheme classifies galaxies into four broad classes:

(i) Elliptical galaxies: These have smooth, almost elliptical isophotes and are divided into sub-types E0, E1, $\cdots$, E7, where the integer is the one closest to $10\,(1-b/a)$, with a and b the lengths of the semi-major and semi-minor axes.

(ii) Spiral galaxies: These have thin disks with spiral arm structures. They are divided into two branches, barred spirals and normal spirals, according to whether or not a recognizable bar-like structure is present in the central part of the galaxy. On each branch, galaxies are further divided into three classes, a, b and c, according to the following three criteria:

- the fraction of the light in the central bulge;
- the tightness with which the spiral arms are wound;
- the degree to which the spiral arms are resolved into stars, HII regions and ordered dust lanes.

These three criteria are correlated: spirals with a pronounced bulge component usually also have tightly wound spiral arms with relatively faint HII regions, and are classified

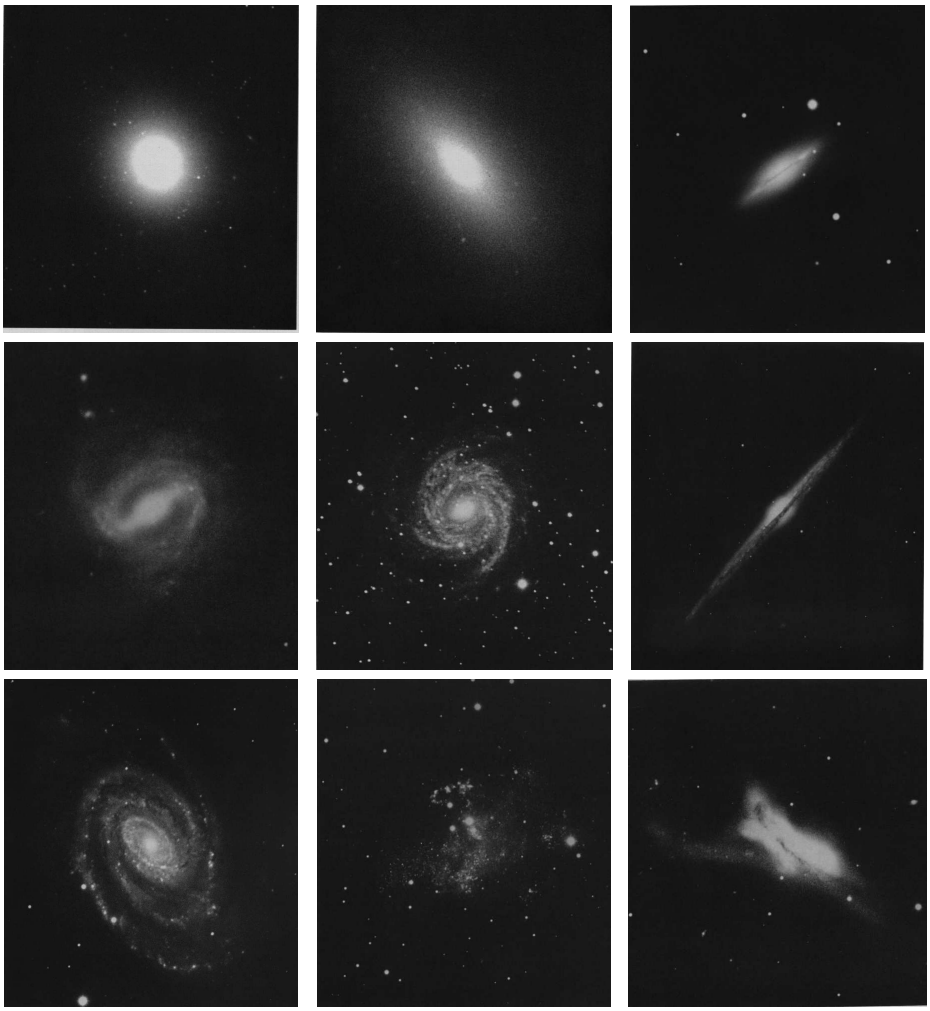


Fig. 2.7. Examples of different types of galaxies. From left to right and top to bottom, NGC 4278 (E1), NGC 3377 (E6), NGC 5866 (SO), NGC 175 (SBa), NGC 6814 (Sb), NGC 4565 (Sb, edge on), NGC 5364 (Sc), Ho II (Irr I), NGC 520 (Irr II). [All images are obtained from the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration]

as Sa's. On the other hand, spirals with weak or absent bulges usually have open arms and bright HII regions and are classified as Sc's. When the three criteria give conflicting indications, Hubble put most emphasis on the openness of the spiral arms.

(iii) Lenticular or S0 galaxies: This class is intermediate between ellipticals and spirals. Like ellipticals, lenticulars have a smooth light distribution with no spiral arms or HII regions. Like spirals they have a thin disk and a bulge, but the bulge is more dominant than that in a spiral galaxy. They may also have a central bar, in which case they are classified as SB0.

(iv) Irregular galaxies: These objects have neither a dominating bulge nor a rotationally symmetric disk and lack any obvious symmetry. Rather, their appearance is generally patchy, dominated by a few HII regions. Hubble did not include this class in his original sequence

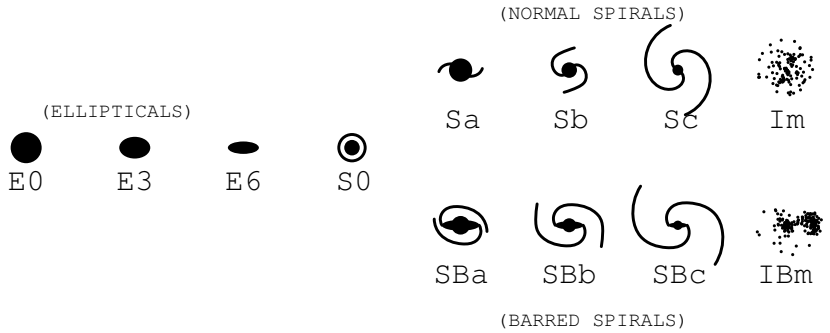


Fig. 2.8. A schematic representation of the Hubble sequence of galaxy morphologies. [Courtesy of R. Abraham, see Abraham (1998)]

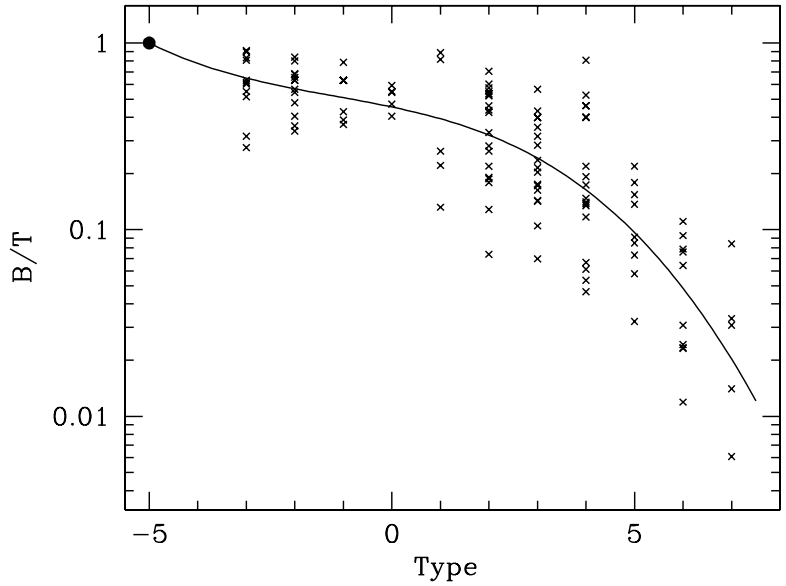


Fig. 2.9. Fractional luminosity of the spheroidal bulge component in a galaxy as a function of morphological type (based on the classification of de Vaucouleurs). Data points correspond to individual galaxies, and the curve is a fit to the mean. Elliptical galaxies (Type $= -5$) are considered to be pure bulges. [Based on data presented in Simien & de Vaucouleurs (1986)]

because he was uncertain whether it should be considered an extension of any of the other classes. Nowadays irregulars are usually included as an extension to the spiral galaxies.

Ellipticals and lenticulars together are often referred to as early-type galaxies, while the spirals and irregulars make up the class of late-type galaxies. Indeed, traversing the Hubble sequence from the left to the right the morphologies are said to change from early- to late-type. Although somewhat confusing, one often uses the terms ‘early-type spirals’ and ‘late-type spirals’ to refer to galaxies at the left or right of the spiral sequence. We caution, though, that this historical nomenclature has no direct physical basis: the reference to ‘early’ or ‘late’ should not be interpreted as reflecting a property of the galaxy’s evolutionary state. Another largely historical

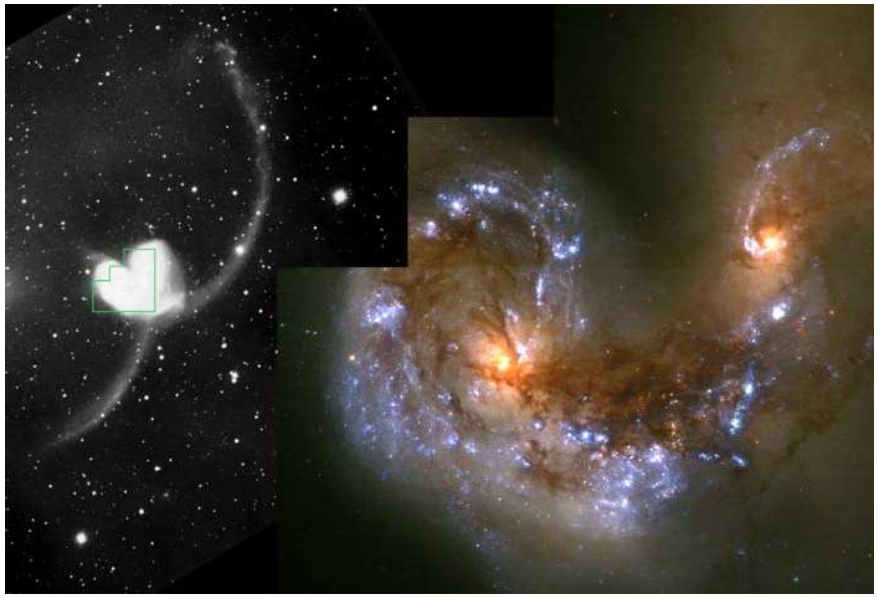


Fig. 2.10. The peculiar galaxy known as the Antennae, a system exhibiting prominent tidal tails (the left inlet), a signature of a recent merger of two spiral galaxies. The close-up of the center reveals the presence of large amounts of dust and many clusters of newly formed stars. [Courtesy of B. Whitmore, NASA, and Space Telescope Science Institute]

nomenclature, which can be confusing at times, is to refer to faint galaxies with $\mathscr{M}_B \gtrsim -18$ as 'dwarf galaxies'. In particular, early-type dwarfs are often split into dwarf ellipticals (dE) and dwarf spheroidals (dSph), although there is no clear distinction between these types – often the term dwarf spheroidals is simply used to refer to early-type galaxies with $\mathscr{M}_B \gtrsim -14$.

Since Hubble, a variety of other classification schemes have been introduced. A commonly used one is due to de Vaucouleurs (1974). He put spirals in the Hubble sequence into a finer gradation by adding new types such as SOa, Sab, Sbc (and the corresponding barred types). After finding that many of Hubble's irregular galaxies in fact had weak spiral arms, de Vaucouleurs also extended the spiral sequence to irregulars, adding types Scd, Sd, Sdm, Sm, Im and I0, in order of decreasing regularity. (The m stands for 'Magellanic' since the Magellanic Clouds are the prototypes of this kind of irregulars). Furthermore, de Vaucouleurs used numbers between -6 and 10 to represent morphological types (the de Vaucouleurs' T types). Table 2.5 shows the correspondence between de Vaucouleurs' notations and Hubble's notations – note that the numerical T-types do not distinguish between barred and unbarred galaxies. As shown in Fig. 2.9, the morphology sequence according to de Vaucouleurs' classification is primarily a sequence in the importance of the bulge.

The Hubble classification and its revisions encompass the morphologies of the majority of the observed galaxies in the local Universe. However, there are also galaxies with strange appearances which defy Hubble's classification. From their morphologies, these "peculiar" galaxies all appear to have been strongly perturbed in the recent past and to be far from dynamical equilibrium, indicating that they are undergoing a transformation. A good example is the Antennae (Fig. 2.10) where the tails are produced by the interaction of the two spiral galaxies, NGC 4038 and NGC 4039, in the process of merging.

The classifications discussed so far are based only on morphology. Galaxies can also be classified according to other properties. For instance, they can be classified into *bright* and *faint*

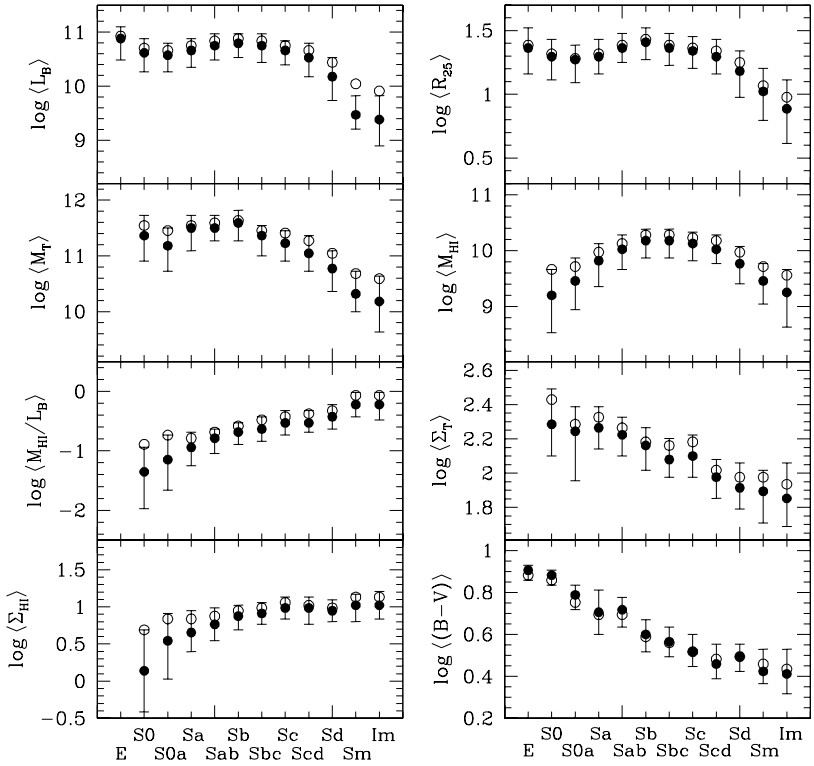


Fig. 2.11. Galaxy properties along the Hubble morphological sequence based on the RC3-UGC sample. Filled circles are medians, open ones are mean values. The bars bracket the 25 and 75 percentiles. Properties plotted are L_B (blue luminosity in erg s^{-1}), R_{25} (the radius in kpc of the 25mag arcsec^{-2} isophote in the B-band), $M_{\rm T}$ (total mass in solar units within a radius $R_{25}/2$), $M_{\rm HI}$ (HI mass in solar units), $M_{\rm HI}/L_B$, $\Sigma_{\rm T}$ (total mass surface density), $\Sigma_{\rm HI}$ (HI mass surface density), and the $B-V$ color. [Based on data presented in Roberts & Haynes (1994)]

according to luminosity, into *high* and *low surface brightness* according to surface brightness, into *red* and *blue* according to color, into *gas-rich* and *gas-poor* according to gas content, into *quiescent* and *starburst* according to their current level of star formation, and into *normal* and *active* according to the presence of an active nucleus. All these properties can be measured observationally, although often with some difficulty. An important aspect of the Hubble sequence (and its modifications) is that many of these properties change systematically along the sequence (see Figs. 2.11 and 2.12), indicating that it reflects a sequence in the basic physical properties of galaxies. However, we stress that the classification of galaxies is far less clear cut than that of stars, whose classification has a sound basis in terms of the H-R diagram and the evolutionary tracks.

2.3.2 Elliptical Galaxies

Elliptical galaxies are characterized by smooth, elliptical surface brightness distributions, contain little cold gas or dust, and have red photometric colors, characteristic of an old stellar population. In this section we briefly discuss some of the main, salient observational properties. A more in-depth discussion, including an interpretation within the physical framework of galaxy formation, is presented in Chapter **??**.

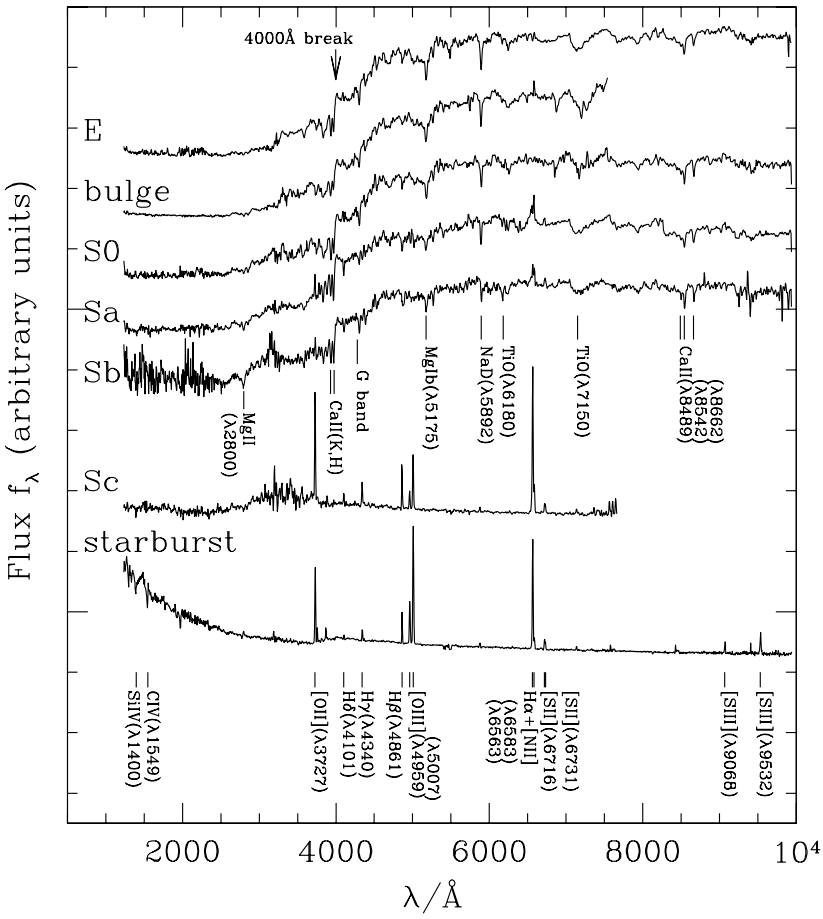


Fig. 2.12. Spectra of different types of galaxies from the ultraviolet to the near-infrared. From ellipticals to late-type spirals, the blue continuum and emission lines become systematically stronger. For early-type galaxies, which lack hot, young stars, most of the light emerges at the longest wavelengths, where one sees absorption lines characteristic of cool K stars. In the blue, the spectrum of early type galaxies show strong H and K absorption lines of calcium and the G band, characteristic of solar type stars. Such galaxies emit little light at wavelengths shorter than 4000 Å and have no emission lines. In contrast, late-type galaxies and starbursts emit most of their light in the blue and near-ultraviolet. This light is produced by hot young stars, which also heat and ionize the interstellar medium giving rise to strong emission lines. [Based on data kindly provided by S. Charlot]

(a) Surface Brightness Profiles The one-dimensional surface brightness profile, $I(R)$, of an elliptical galaxy is usually defined as the surface brightness as a function of the isophotal semi-major axis length R. If the position angle of the semi-major axis changes with radius, a phenomenon called isophote twisting, then $I(R)$ traces the surface brightness along a curve that connects the intersections of each isophote with its own major axis.

The surface brightness profile of spheroidal galaxies is generally well fit by the Sérsic profile

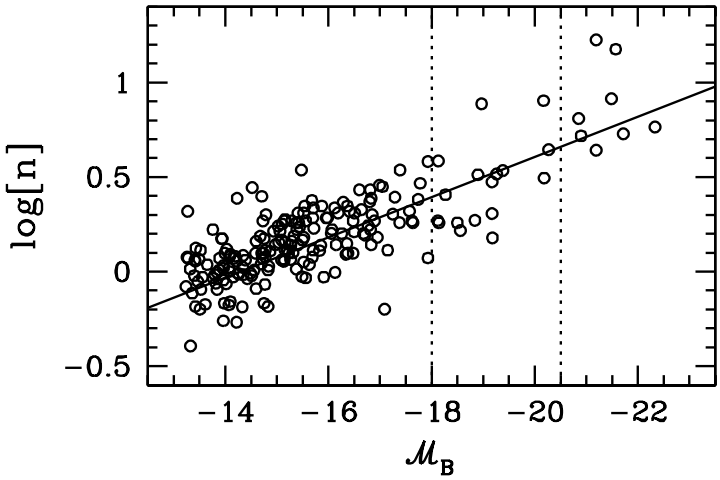


Fig. 2.13. Correlation between the Sérsic index, n, and the absolute magnitude in the B-band for a sample of elliptical galaxies. The vertical dotted lines correspond to $\mathscr{M}_B = -18$ and $\mathscr{M}_B = -20.5$ and are shown to facilitate a comparison with Fig. 2.14. [Data compiled and kindly made available by A. Graham (see Graham & Guzmán, 2003)]

(Sérsic, 1968), or $R^{1/n}$ profile,†

$$I(R) = I_0 \exp\left[-\beta_n \left(\frac{R}{R_\mathrm{e}}\right)^{1/n}\right] = I_\mathrm{e} \exp\left[-\beta_n \left\{\left(\frac{R}{R_\mathrm{e}}\right)^{1/n} - 1\right\}\right], \tag{2.22}$$

where I_0 is the central surface brightness, n is the so-called Sérsic index which sets the concentration of the profile, R_e is the effective radius that encloses half of the total light, and $I_\mathrm{e} = I(R_\mathrm{e})$. Surface brightness profiles are often expressed in terms of $\mu \propto -2.5\log(I)$ (which has the units of mag/arcsec^2), for which the Sérsic profile takes the form

$$\mu(R) = \mu_\mathrm{e} + 1.086\,\beta_n \left[\left(\frac{R}{R_\mathrm{e}}\right)^{1/n} - 1\right]. \tag{2.23}$$

The value for β_n follows from the definition of R_e and is well approximated by $\beta_n = 2n - 0.324$ (but only for $n \gtrsim 1$). Note that Eq. (2.22) reduces to a simple exponential profile for $n = 1$. The total luminosity of a spherical system with a Sérsic profile is

$$L = 2\pi \int_0^\infty I(R)\, R\, \mathrm{d}R = \frac{2\pi n \Gamma(2n)}{(\beta_n)^{2n}} I_0 R_\mathrm{e}^2, \tag{2.24}$$

with $\Gamma(x)$ the gamma function. Early photometry of the surface brightness profiles of normal giant elliptical galaxies was well fit by a de Vaucouleurs profile, which is a Sérsic profile with $n = 4$ (and $\beta_n = 7.67$) and is therefore also called a $R^{1/4}$-profile. With higher accuracy photometry and with measurements of higher and lower luminosity galaxies, it became clear that ellipticals as a class are better fit by the more general Sérsic profile. In fact, the best-fit values for n have been found to be correlated with the luminosity and size of the galaxy: while at the faint end dwarf ellipticals have best-fit values as low as $n \sim 0.5$, the brightest ellipticals can have Sérsic indices $n \gtrsim 10$ (see Fig. 2.13).

Instead of I_0 or I_e, one often characterizes the surface brightness of an elliptical galaxy via the

† A similar formula, but with R denoting 3-D rather than projected radius, was used by Einasto (1965) to describe the stellar halo of the Milky Way.

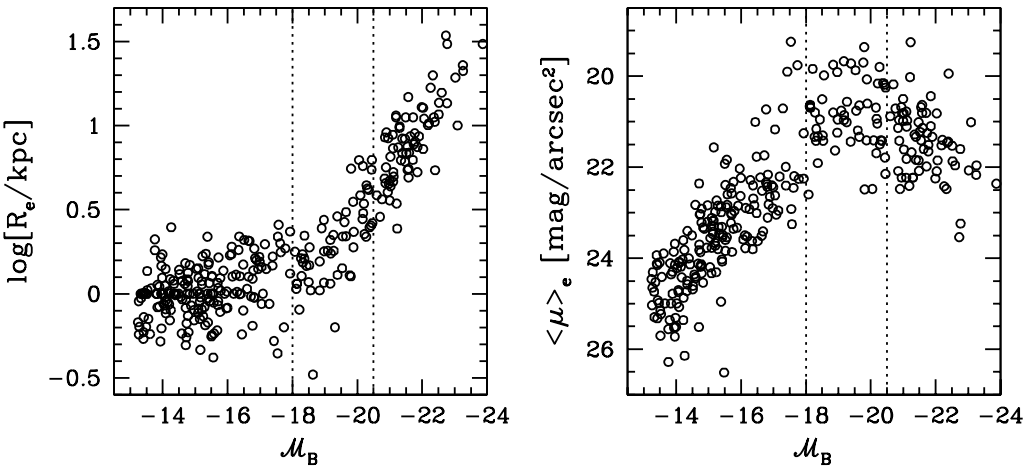


Fig. 2.14. The effective radius (left panel) and the average surface brightness within the effective radius (right panel) of elliptical galaxies plotted against their absolute magnitude in the *B*-band. The vertical dotted lines correspond to $\mathscr{M}_B = -18$ and $\mathscr{M}_B = -20.5$. [Data compiled and kindly made available by A. Graham (see Graham & Guzmán, 2003), combined with data taken from Bender et al. (1992)]

average surface brightness within the effective radius, $\langle I \rangle_{\rm e} = L/(2\pi R_{\rm e}^2)$, or, in magnitudes, $\langle \mu \rangle_{\rm e}$. Fig. 2.14 shows how $R_{\rm e}$ and $\langle \mu \rangle_{\rm e}$ are correlated with luminosity. At the bright end ($\mathscr{M}_B \lesssim -18$), the sizes of elliptical galaxies increase strongly with luminosity. Consequently, the average surface brightness actually decreases with increasing luminosity. At the faint end ($\mathscr{M}_B \gtrsim -18$), however, all ellipticals have roughly the same effective radius ($R_{\rm e} \sim 1\,{\rm kpc}$), so that the average surface brightness *increases* with increasing luminosity. Because of this apparent change-over in properties, ellipticals with $\mathscr{M}_B \gtrsim -18$ are typically called 'dwarf' ellipticals, in order to distinguish them from the 'normal' ellipticals (see §2.3.5). However, this alleged 'dichotomy' between dwarf and normal ellipticals has recently been challenged. A number of studies have argued that there is actually a smooth and continuous sequence of increasing surface brightness with increasing luminosity, except for the very bright end ($\mathscr{M}_B \lesssim -20.5$) where this trend is reversed (e.g., Jerjen & Binggeli, 1997; Graham & Guzmán, 2003).

The fact that the photometric properties of elliptical galaxies undergo a transition around $\mathscr{M}_B \sim -20.5$ is also evident from their central properties (in the inner few hundred parsec). High spatial resolution imaging with the HST has revealed that the central surface brightness profiles of elliptical galaxies are typically not well described by an inward extrapolation of the Sérsic profiles fit to their outer regions. Bright ellipticals with $\mathscr{M}_B \lesssim -20.5$ typically have a deficit in $I(R)$ with respect to the best-fit Sérsic profile, while fainter ellipticals reveal excess surface brightness. Based on the value of the central cusp slope $\gamma \equiv {\rm d}\log I/{\rm d}\log r$ the population of ellipticals has been split into 'core' ($\gamma < 0.3$) and 'power-law' ($\gamma \geq 0.3$) systems. The majority of bright galaxies with $\mathscr{M}_B \lesssim -20.5$ have cores, while power-law galaxies typically have $\mathscr{M}_B > -20.5$ (Ferrarese et al., 1994; Lauer et al., 1995). Early results, based on relatively small samples suggested a bimodal distribution in γ, with virtually no galaxies in the range $0.3 < \gamma < 0.5$. However, subsequent studies have significantly weakened the evidence for a clear dichotomy, finding a population of galaxies with intermediate properties (Rest et al., 2001; Ravindranath et al., 2001). In fact, recent studies, using significantly larger samples, have argued for a smooth transition in nuclear properties, with no evidence for any dichotomy (Ferrarese et al., 2006b; Côté et al., 2007, see also §**??**).

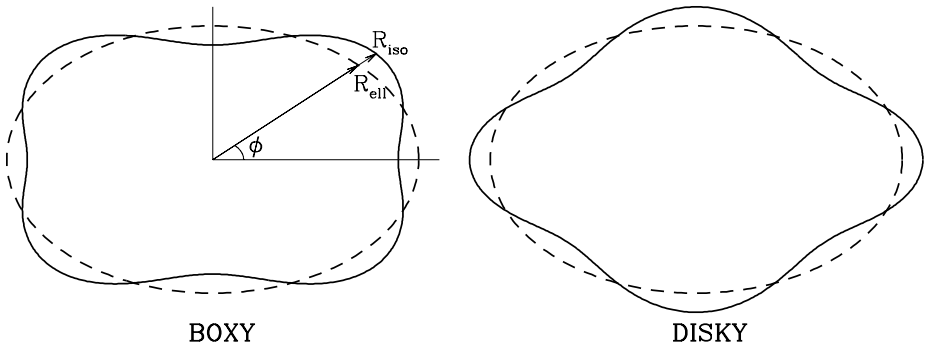


Fig. 2.15. An illustration of boxy and disky isophotes (solid curves). The dashed curves are the corresponding best-fit ellipses.

(b) Isophotal Shapes The isophotes of elliptical galaxies are commonly fitted by ellipses and characterized by their minor-to-major axis ratios b/a (or, equivalently, by their ellipticities $\varepsilon = 1-b/a$) and by their position angles. In general, the ellipticity may change across the system, in which case the overall shape of an elliptical is usually defined by some characteristic ellipticity (e.g. that of the isophote which encloses half the total light). In most cases, however, the variation of ε with radius is not large, so that the exact definition is of little consequence. For normal elliptical galaxies the axis ratio lies in the range $0.3 \lesssim b/a \leq 1$, corresponding to types E0 to E7. In addition to the ellipticity, the position angle of the isophotes may also change with radius, a phenomenon called isophote twisting.

Detailed modeling of the surface brightness of elliptical galaxies shows that their isophotes are generally not exactly elliptical. The deviations from perfect ellipses are conveniently quantified by the Fourier coefficients of the function

$$\Delta(\phi) \equiv R_{\rm iso}(\phi) - R_{\rm ell}(\phi) = a_0 + \sum_{n=1}^{\infty} (a_n \cos n\phi + b_n \sin n\phi) \,, \tag{2.25}$$

where $R_{\rm iso}(\phi)$ is the radius of the isophote at angle ϕ and $R_{\rm ell}(\phi)$ is the radius of an ellipse at the same angle (see Fig. 2.15). Typically one considers the ellipse that best-fits the isophote in question, so that a_0, a_1, a_2, b_1 and b_2 are all consistent with zero within the errors. The deviations from this best-fit isophote are then expressed by the higher-order Fourier coefficients a_n and b_n with $n \geq 3$. Of particular importance are the values of the a_4 coefficients, which indicate whether the isophotes are "disky" ($a_4 > 0$) or "boxy" ($a_4 < 0$), as illustrated in Fig. 2.15. The *diskiness* of an isophote is defined as the dimensionless quantity, a_4/a, where a is the length of the semi-major axis of the isophote's best-fit ellipse. We caution that some authors use an alternative method to specify the deviations of isophotes from pure ellipses. Instead of using isophote deviation from an ellipse, they quantify how the *intensity* fluctuates along the best-fit ellipse:

$$I(\phi) = I_0 + \sum_{n=1}^{\infty} (A_n \cos n\phi + B_n \sin n\phi) \,, \tag{2.26}$$

with I_0 the intensity of the best-fit ellipse. The coefficients A_n and B_n are (approximately) related to a_n and b_n according to

$$A_n = a_n \left| \frac{{\rm d}I}{{\rm d}R} \right| , \qquad B_n = b_n \left| \frac{{\rm d}I}{{\rm d}R} \right| , \tag{2.27}$$

where $R = a\sqrt{1-\varepsilon}$, with ε the ellipticity of the best-fit ellipse.

The importance of the disky/boxy classification is that boxy and disky ellipticals turn out to

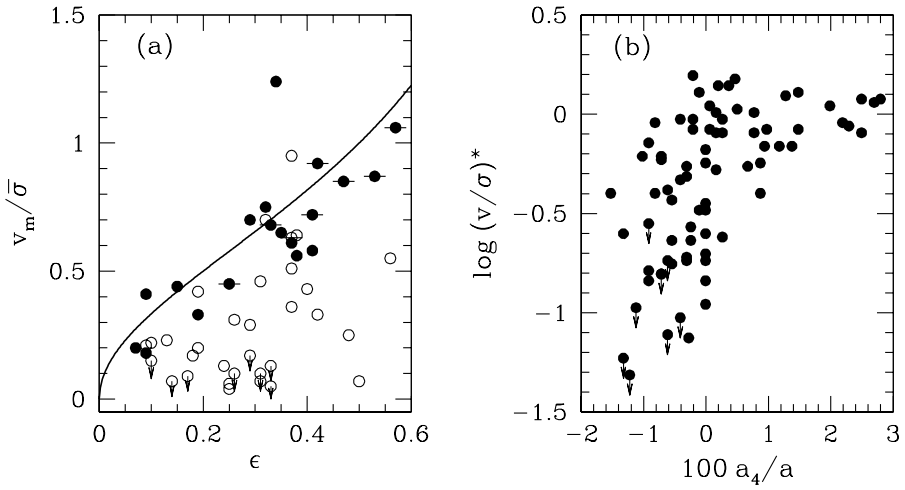


Fig. 2.16. (a) The ratio $v_{\rm m}/\overline{\sigma}$ for ellipticals and bulges (with bulges marked by horizontal bars) versus ellipticity. Open circles are for bright galaxies with $\mathscr{M}_B \leq 20.5$, with upper limits marked by downward arrows; solid circles are for early-types with $-20.5 < \mathscr{M}_B < -18$. The solid curve is the relation expected for an oblate galaxy flattened by rotation. [Based on data published in Davies et al. (1983)] (b) The rotation parameter $(v/\sigma)^*$ (defined as the ratio of $v_{\rm m}/\overline{\sigma}$ to the value expected for an isotropic oblate spheroid flattened purely by rotation) versus the average diskiness of the galaxy. [Based on data published in Kormendy & Bender (1996)]

have systematically different properties. Boxy ellipticals are usually bright, rotate slowly, and show stronger than average radio and X-ray emission, while disky ellipticals are fainter, have significant rotation and show little or no radio and X-ray emission (e.g., Bender et al., 1989; Pasquali et al., 2007). In addition, the diskiness is correlated with the nuclear properties as well; disky ellipticals typically have steep cusps, while boxy ellipticals mainly harbor central cores (e.g., Jaffe et al., 1994; Faber et al., 1997).

(c) Colors Elliptical galaxies in general have red colors, indicating that their stellar contents are dominated by old, metal-rich stars (see §**??**). In addition, the colors are tightly correlated with the luminosity such that brighter ellipticals are redder (Sandage & Visvanathan, 1978). As we will see in §**??**, the slope and (small) scatter of this color-magnitude relation puts tight constraints on the star formation histories of elliptical galaxies. Ellipticals also display color gradient. In general, the outskirt has a bluer color than the central region. Peletier et al. (1990) obtained a mean logarithmic gradient of $\Delta(U-R)/\Delta\log r = -0.20 \pm 0.02$ mag in $U-R$, and of $\Delta(B-R)/\Delta\log r = -0.09 \pm 0.02$ mag in $B-R$, in good agreement with the results obtained by Franx et al. (1989).

(d) Kinematic Properties Giant ellipticals generally have low rotation velocities. Observationally, this may be characterized by the ratio of maximum line-of-sight streaming motion $v_{\rm m}$ (relative to the mean velocity of the galaxy) to $\overline{\sigma}$, the average value of the line-of-sight velocity dispersion interior to $\sim R_{\rm e}/2$. This ratio provides a measure of the relative importance of ordered and random motions within the galaxy. For isotropic, oblate galaxies flattened by the centrifugal force generated by rotation, $v_{\rm m}/\overline{\sigma} \approx \sqrt{\varepsilon/(1-\varepsilon)}$, with ε the ellipticity of the spheroid (see §**??**). As shown in Fig. 2.16a, for bright ellipticals, $v_{\rm m}/\overline{\sigma}$ lies well below this prediction, indicating that their flattening must be due to velocity anisotropy, rather than rotation. In contrast, ellipticals of intermediate luminosities (with absolute magnitude $-20.5 \lesssim \mathscr{M}_B \lesssim -18.0$) and spiral

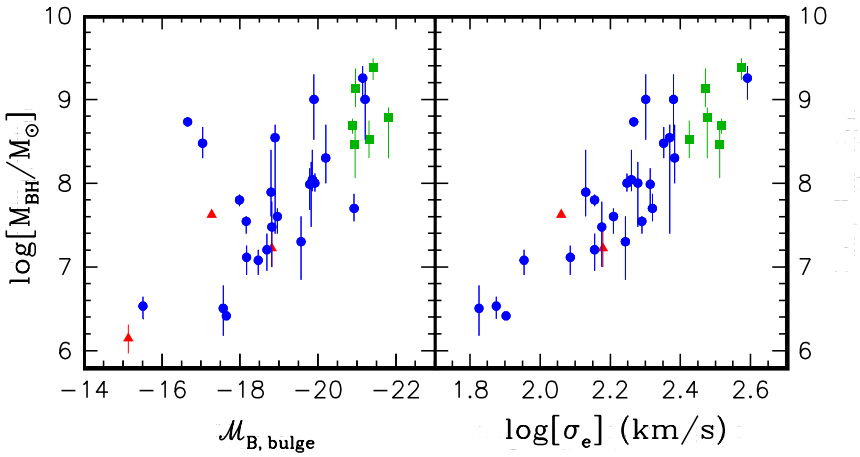


Fig. 2.17. The masses of central black holes in ellipticals and spiral bulges plotted against the absolute magnitude (left) and velocity dispersion (right) of their host spheroids. [Adapted from Kormendy (2001)]

bulges have $v_{\rm m}/\overline{\sigma}$ values consistent with rotational flattening. Fig. 2.16b shows, as noted above, that disky and boxy ellipticals have systematically different kinematics: while disky ellipticals are consistent with rotational flattening, rotation in boxy ellipticals is dynamically unimportant.

When the kinematic structure of elliptical galaxies is examined in more detail a wide range of behavior is found. In most galaxies the line-of-sight velocity dispersion depends only weakly on position and is constant or falls at large radii. Towards the center the dispersion may drop weakly, remain flat, or rise quite sharply. The behavior of the mean line-of-sight streaming velocity is even more diverse. While most galaxies show maximal streaming along the major axis, a substantial minority show more complex behavior. Some have non-zero streaming velocities along the minor axis, and so it is impossible for them to be an oblate body rotating about its symmetry axis. Others have mean motions which change suddenly in size, in axis, or in sign in the inner regions, the so-called kinematically decoupled cores. Such variations point to a variety of formation histories for apparently similar galaxies.

At the very center of most nearby ellipticals (and also spiral and S0 bulges) the velocity dispersion is observed to rise more strongly than can be understood as a result of the gravitational effects of the observed stellar populations alone. It is now generally accepted that this rise signals the presence of a central supermassive black hole. Such a black hole appears to be present in virtually every galaxy with a significant spheroidal component, and to have a mass which is roughly 0.1 percent of the total stellar mass of the spheroid (Fig. 2.17). A more detailed discussion of supermassive black holes is presented in §**??**.

(e) Scaling Relations The kinematic and photometric properties of elliptical galaxies are correlated. In particular, ellipticals with a larger (central) velocity dispersion are both brighter, known as the Faber-Jackson relation, and larger, known as the D_n-σ relation (D_n is the isophotal diameter within which the average, enclosed surface brightness is equal to a fixed value). Furthermore, when plotted in the three-dimensional space spanned by $\log\sigma_0$, $\log R_{\rm e}$ and $\log\langle I\rangle_{\rm e}$, elliptical galaxies are concentrated in a plane (see Fig. 2.18) known as the fundamental plane. In mathematical form, this plane can be written as

$$\log R_{\rm e} = a\log\sigma_0 + b\log\langle I\rangle_{\rm e} + \text{constant}, \qquad (2.28)$$

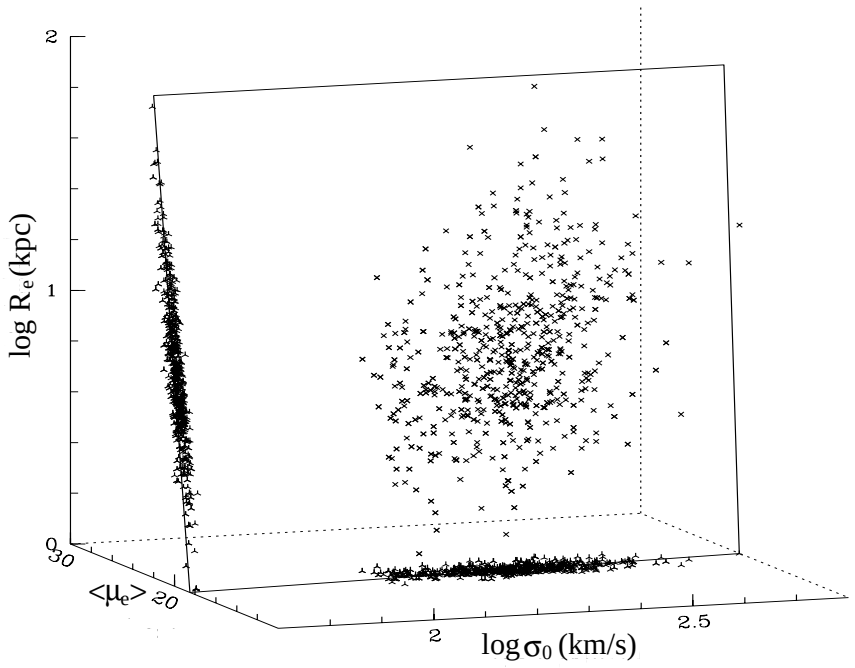


Fig. 2.18. The fundamental plane of elliptical galaxies in the $\log R_{\rm e}$-$\log\sigma_0$-$\langle\mu\rangle_{\rm e}$ space (σ_0 is the central velocity dispersion, and $\langle\mu\rangle_{\rm e}$ is the mean surface brightness within $R_{\rm e}$ expressed in magnitudes per square arcsecond). [Plot kindly provided by R. Saglia, based on data published in Saglia et al. (1997) and Wegner et al. (1999)]

where $\langle I\rangle_{\rm e}$ is the mean surface brightness within $R_{\rm e}$ (not to be confused with $I_{\rm e}$, which is the surface brightness at $R_{\rm e}$). The values of a and b have been estimated in various photometric bands. For example, Jørgensen et al. (1996) obtained $a = 1.24 \pm 0.07$, $b = -0.82 \pm 0.02$ in the optical, while Pahre et al. (1998) obtained $a = 1.53 \pm 0.08$, $b = -0.79 \pm 0.03$ in the near-infrared. More recently, using 9,000 galaxies from the Sloan Digital Sky Survey (SDSS), Bernardi et al. (2003) found the best fitting plane to have $a = 1.49 \pm 0.05$ and $b = -0.75 \pm 0.01$ in the SDSS r-band with a *rms* of only 0.05. The Faber-Jackson and D_n-σ relations are both 2-dimensional projections of this fundamental plane. While the D_n-σ projection is close to edge-on and so has relatively little scatter, the Faber-Jackson projection is significantly tilted resulting in somewhat larger scatter. These relations can not only be used to determine the distances to elliptical galaxies, but are also important for constraining theories for their formation (see §**??**).

(f) Gas Content Although it was once believed that elliptical galaxies contain neither gas nor dust, it has become clear over the years that they actually contain a significant amount of interstellar medium which is quite different in character from that in spiral galaxies (e.g., Roberts et al., 1991; Buson et al., 1993). Hot ($\sim 10^7$ K) X-ray emitting gas usually dominates the interstellar medium (ISM) in luminous ellipticals, where it can contribute up to $\sim 10^{10}\,{\rm M}_\odot$ to the total mass of the system. This hot gas is distributed in extended X-ray emitting atmospheres (Fabbiano, 1989; Mathews & Brighenti, 2003), and serves as an ideal tracer of the gravitational potential in which the galaxy resides (see §**??**).

In addition, many ellipticals also contain small amounts of warm ionized (10^4 K) gas as well as cold (< 100 K) gas and dust. Typical masses are $10^2 - 10^4\,{\rm M}_\odot$ in ionized gas and $10^6 - 10^8\,{\rm M}_\odot$ in the cold component. Contrary to the case for spirals, the amounts of dust and of atomic and

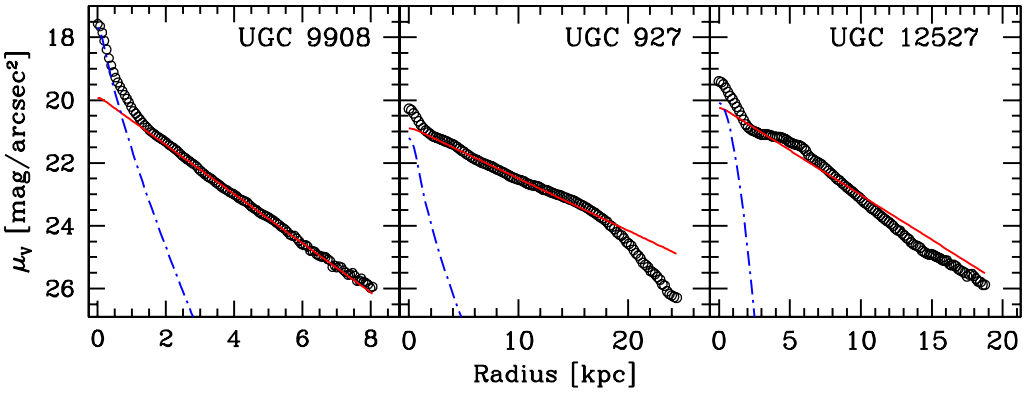


Fig. 2.19. The surface brightness profiles of three disk galaxies plus their decomposition in an exponential disk (solid line) and a Sérsic bulge (dot-dashed line). [Based on data published in MacArthur et al. (2003) and kindly made available by L. MacArthur]

molecular gas are not correlated with the luminosity of the elliptical. In many cases, the dust and/or ionized gas is located in the center of the galaxy in a small disk component, while other ellipticals reveal more complex, filamentary or patchy dust morphologies (e.g., van Dokkum & Franx, 1995; Tran et al., 2001). This gas and dust either results from accumulated mass loss from stars within the galaxy or has been accreted from external systems. The latter is supported by the fact that the dust and gas disks are often found to have kinematics decoupled from that of the stellar body (e.g., Bertola et al., 1992)

2.3.3 Disk Galaxies

Disk galaxies have a far more complex morphology than ellipticals. They typically consist of a thin, rotationally supported disk with spiral arms and often a bar, plus a central bulge component. The latter can dominate the light of the galaxy in the earliest types and may be completely absent in the latest types. The spiral structure is best seen in face-on systems and is defined primarily by young stars, HII regions, molecular gas and dust absorption. Edge-on systems, on the other hand, give a better handle on the vertical structure of the disk, which often reveals two separate components: a thin disk and a thick disk. In addition, there are indications that disk galaxies also contain a spheroidal, stellar halo, extending out to large radii. In this subsection we briefly summarize the most important observational characteristics of disk galaxies. A more in-depth discussion, including models for their formation, is presented in Chapter **??**.

(a) Surface Brightness Profiles Fig. 2.19 shows the surface brightness profiles of three disk galaxies, as measured along their projected, major axes. A characteristic of these profiles is that they typically reveal a range over which $\mu(R)$ can be accurately fitted by a straight line. This corresponds to an exponential surface brightness profile

$$I(R) = I_0 \exp\left(-R/R_{\rm d}\right) , \quad I_0 = \frac{L}{2\pi R_{\rm d}^2} , \tag{2.29}$$

(i.e., a Sérsic profile with $n = 1$). Here R is the cylindrical radius, $R_{\rm d}$ is the exponential scale-length, I_0 is the central luminosity surface density, and L is the total luminosity. The effective radius enclosing half of the total luminosity is $R_{\rm e} \simeq 1.67 R_{\rm d}$. Following Freeman (1970) it has become customary to associate this exponential surface brightness profile with the actual disk

component. The central regions of the majority of disk galaxies show an excess surface brightness with respect to a simple inward extrapolation of this exponential profile. This is interpreted as a contribution from the bulge component, and such interpretation is supported by images of edge-on disk galaxies, which typically reveal a central, roughly spheroidal, component clearly thicker than the disk itself (see e.g., NGC 4565 in Fig. 2.7). At large radii, the surface brightness profiles often break to a much steeper (roughly exponential) profile (an example is UGC 927, shown in Fig. 2.19). These breaks occur at radii $R_{\rm b} = \alpha R_{\rm d}$ with α in the range 2.5 to 4.5 (e.g., Pohlen et al., 2000; de Grijs et al., 2001).

Fig. 2.20 shows $R_{\rm e}$ and $\mu_{\rm e}$ as functions of the absolute magnitude for a large sample of disk dominated galaxies (i.e., with a small or negligible bulge component). Clearly, as expected, more luminous galaxies tend to be larger, although there is large scatter, indicating that galaxies of a given luminosity span a wide range in surface brightnesses. Note that, similar to ellipticals with $\mathscr{M}_B \gtrsim -20.5$, more luminous disk galaxies on average have a higher surface brightness (see Fig. 2.14).

When decomposing the surface brightness profiles of disk galaxies into the contributions of disk and bulge, one typically fits $\mu(R)$ with the sum of an exponential profile for the disk and a Sérsic profile for the bulge. We caution, however, that these bulge-disk decompositions are far from straightforward. Often the surface brightness profiles show clear deviations from a simple sum of an exponential plus Sérsic profile (e.g., UGC 12527 in Fig. 2.19). In addition, seeing tends to blur the central surface brightness distribution, which has to be corrected for, dust can cause significant extinction, and bars and spiral arms represent clear deviations from perfect axisymmetry. In addition, disks are often lop-sided (the centers of different isophotes are offset from each other in one particular direction) and can even be warped (the disk is not planar, but different disk radii are tilted with respect to each other). These difficulties can be partly overcome by using the full two-dimensional information in the image, by using color information to correct for dust, and by using kinematic information. Such studies require much detailed work and even then ambiguities remain.

Despite these uncertainties, bulge-disk decompositions have been presented for large samples of disk galaxies (e.g., de Jong, 1996a; Graham, 2001; MacArthur et al., 2003). These studies have shown that more luminous bulges have a larger best-fit Sérsic index, similar to the relation found for elliptical galaxies (Fig. 2.13): while the relatively massive bulges of early-type spirals have surface brightness profiles with a best-fit Sérsic index $n \sim 4$, the surface brightness profiles of bulges in late-type spirals are better fit with $n \lesssim 1$. In addition, the ratio between the effective radius of the bulge and the disk scale length is found to be roughly independent of Hubble type, with an average of $\langle r_{\rm e,b}/R_{\rm d}\rangle = 0.22 \pm 0.09$. The fact that the bulge-to-disk ratio increases from late-type to early-type, therefore indicates that brighter bulges have a higher surface brightness.

Although the majority of bulges have isophotes that are close to elliptical, a non-negligible fraction of predominantly faint bulges in edge-on, late-type disk galaxies have isophotes that are extremely boxy, or sometimes even have the shape of a peanut. As we will see in §**??**, these peanut-shaped bulges are actually bars that have been thickened out of the disk plane.

(b) Colors In general, disk galaxies are bluer than elliptical galaxies of the same luminosity. As discussed in §**??**, this is mainly owing to the fact that disk galaxies are still actively forming stars (young stellar populations are blue). Similar to elliptical galaxies, more luminous disks are redder, although the scatter in this color-magnitude relation is much larger than that for elliptical galaxies. Part of this scatter is simply due to inclination effects, with more inclined disks being more extincted and hence redder, although the intrinsic scatter (corrected for dust extinction) is still significantly larger than for ellipticals. In general, disk galaxies also reveal color gradients, with the outer regions being bluer than the inner regions (e.g., de Jong, 1996b).

Although it is often considered standard lore that disks are blue and bulges are red, this is

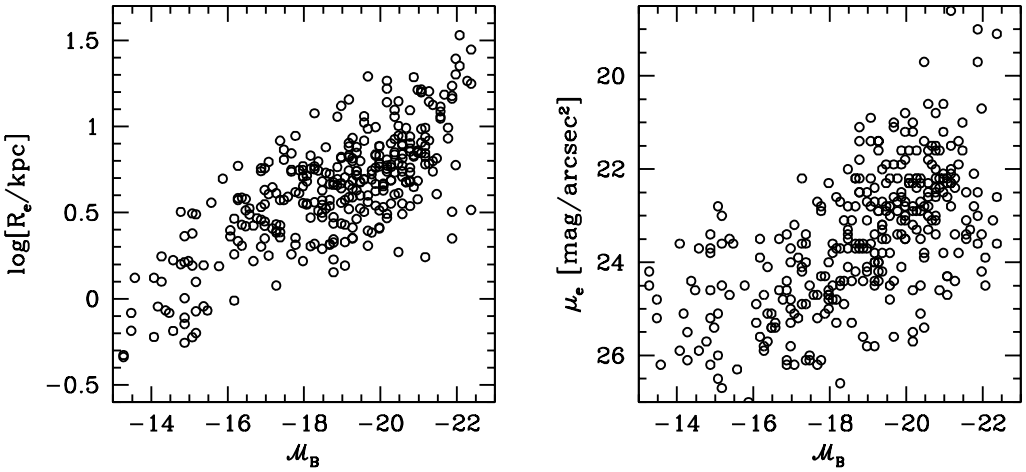


Fig. 2.20. The effective radius (left panel) and the surface brightness at the effective radius (right panel) of disk dominated galaxies plotted against their absolute magnitude in the *B*-band. [Based on data published in Impey et al. (1996)]

not supported by actual data. Rather, the colors of bulges are in general very similar to, or at least strongly correlated with, the central colors of their associated disks (e.g., de Jong, 1996a; Peletier & Balcells, 1996; MacArthur et al., 2004). Consequently, bulges also span a wide range in colors.

(c) Disk Vertical Structure Galaxy disks are not infinitesimally thin. Observations suggest that the surface brightness distribution in the 'vertical' (z-) direction is largely independent of the distance R from the disk center. The three-dimensional luminosity density of the disk is therefore typically written in separable form as

$$\nu(R,z) = \nu_0 \exp(-R/R_d)\, f(z)\,. \tag{2.30}$$

A general fitting function commonly used to describe the luminosity density of disks in the z-direction is

$$f_n(z) = \mathrm{sech}^{2/n}\left(\frac{n|z|}{2z_\mathrm{d}}\right), \tag{2.31}$$

where n is a parameter controlling the shape of the profile near $z = 0$ and z_d is called the scale height of the disk. Note that all these profiles project to face-on surface brightness profiles given by Eq. (2.29) with $I_0 = a_n \nu_0 z_\mathrm{d}$, with a_n a constant. Three values of n have been used extensively in the literature:

$$f_n(z) = \begin{cases} \mathrm{sech}^2(z/2z_\mathrm{d}) & a_n = 4 \quad n = 1 \\ \mathrm{sech}(z/z_\mathrm{d}) & a_n = \pi \quad n = 2 \\ \exp(-|z|/z_\mathrm{d}) & a_n = 2 \quad n = \infty \end{cases}. \tag{2.32}$$

The sech^2-form for $n = 1$ corresponds to a self-gravitating isothermal sheet. Although this model has been used extensively in dynamical modeling of disk galaxies (see §**??**), it is generally recognized that the models with $n = 2$ and $n = \infty$ provide better fits to the observed surface brightness profiles. Note that all $f_n(z)$ decline exponentially at large $|z|$; they only differ near the midplane, where larger values of n result in steeper profiles. Unfortunately, since dust is usually concentrated near the mid-plane, it is difficult to accurately constrain n. The typical value of the ratio between the vertical and radial scale lengths is $z_\mathrm{d}/R_\mathrm{d} \sim 0.1$, albeit with considerable scatter.

Finally, it is found that most (if not all) disks have excess surface brightness, at large distances from the midplane, that cannot be described by Eq. (2.31). This excess light is generally ascribed to a separate 'thick disk' component, whose scale height is typically a factor three larger than for the 'thin disk'. The radial scale lengths of thick disks, however, are remarkably similar to those of their corresponding thin disks, with typical ratios of $R_{\rm d,thick}/R_{\rm d,thin}$ in the range $1.0-1.5$, while the stellar mass ratios $M_{\rm d,thick}/M_{\rm d,thin}$ decrease from ~ 1 for low mass disks with $V_{\rm rot} \lesssim 75\,{\rm km\,s^{-1}}$ to ~ 0.2 for massive disks with $V_{\rm rot} \gtrsim 150\,{\rm km\,s^{-1}}$ (Yoachim & Dalcanton, 2006).

(d) Stellar Halos The Milky Way contains a halo of old, metal poor stars with a density distribution that falls off as a power-law, $\rho \propto r^{-\alpha}$ ($\alpha \sim 3$). In recent years, however, it has become clear that the stellar halo reveals a large amount of substructure in the form of stellar streams (e.g., Helmi et al., 1999; Yanny et al., 2003; Bell et al., 2008). These streams are associated with material that has been tidally stripped from satellite galaxies and globular clusters (see §**??**), and in some cases they can be unambiguously associated with their original stellar structure (e.g., Ibata et al., 1994; Odenkirchen et al., 2002). Similar streams have also been detected in our neighbor galaxy, M31 (Ferguson et al., 2002).

However, the detection of stellar halos in more distant galaxies, where the individual stars cannot be resolved, has proven extremely difficult due to the extremely low surface brightnesses involved (typically much lower than that of the sky). Nevertheless, using extremely deep imaging, Sackett et al. (1994) detected a stellar halo around the edge-on spiral galaxy NGC 5907. Later, and deeper observations of this galaxy suggest that this extraplanar emission is once again associated with a ring-like stream of stars (Zheng et al., 1999). By stacking the images of hundreds of edge-on disk galaxies, Zibetti et al. (2004) were able to obtain statistical evidence for stellar halos around these systems, suggesting that they are in fact rather common. On the other hand, recent observations of the nearby late-type spiral M33 seem to exclude the presence of a significant stellar halo in this galaxy (Ferguson et al., 2007). Currently the jury is still out as to what fraction of (disk) galaxies contain a stellar halo, and as to what fraction of the halo stars are associated with streams versus a smooth, spheroidal component.

(e) Bars and Spiral Arms More than half of all spirals show bar-like structures in their inner regions. This fraction does not seem to depend significantly on the spiral type, and indeed S0 galaxies are also often barred. Bars generally have isophotes which are more squarish than ellipses and can be fit by the 'generalized ellipse' formula, $(|x|/a)^c + (|y|/b)^c = 1$, where a, b and c are constants and c is substantially larger than 2. Bars are, in general, quite elongated, with axis ratios in their equatorial planes ranging from about 2.5 to 5. Since it is difficult to observe bars in edge-on galaxies, their thickness is not well determined. However, since bars are so common, some limits may be obtained from the apparent thickness of the central regions of edge-on spirals. Such limits suggest that most bars are very flat, probably as flat as the disks themselves, but the bulges complicate this line of argument and it is possible that some bulges (for example, the peanut-shaped bulges) are directly related to bars (see §**??**).

Galaxy disks show a variety of spiral structure. 'Grand-design' systems have arms (most frequently two) which can be traced over a wide range of radii and in many, but far from all, cases are clearly related to a strong bar or to an interacting neighbor. 'Flocculent' systems, on the other hand, contain many arm segments and have no obvious large-scale pattern. Spiral arms are classified as leading or trailing according to the sense in which the spiral winds (moving from center to edge) relative to the rotation sense of the disk. Almost all spirals for which an unambiguous determination can be made are trailing.

Spiral structure is less pronounced (though still present) in red light than in blue light. The spiral structure is also clearly present in density maps of atomic and molecular gas and in maps of dust obscuration. Since the blue light is dominated by massive and short-lived stars born in dense molecular clouds, while the red light is dominated by older stars which make up the bulk

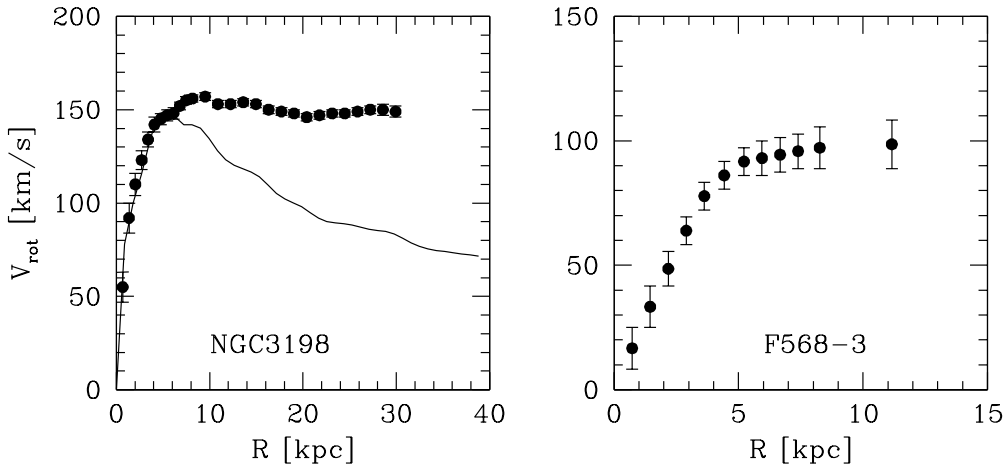


Fig. 2.21. The rotation curves of the Sc galaxy NGC 3198 (left) and the low-surface brightness galaxy F568-3 (right). The curve in the left panel shows the contribution from the disk mass assuming a mass-to-light ratio of $3.8\,M_\odot/L_\odot$. [Based on data published in Begeman (1989) and Swaters et al. (2000)]

of the stellar mass of the disk, this suggests that spiral structure is not related to the star formation process alone, but affects the structure of all components of disks, a conclusion which is more secure for grand-design than for flocculent spirals (see §**??** for details).

(f) Gas Content Unlike elliptical galaxies which contain gas predominantly in a hot and highly ionized state, the gas component in spiral galaxies is mainly in neutral hydrogen (HI) and molecular hydrogen (H_2). Observations in the 21-cm lines of HI and in the mm-lines of CO have produced maps of the distribution of these components in many nearby spirals (e.g., Young & Scoville, 1991). The gas mass fraction increases from about 5% in massive, early-type spirals (Sa/SBa) to as much as 80% in low mass, low surface brightness disk galaxies (McGaugh & de Blok, 1997). In general, while the distribution of molecular gas typically traces that of the stars, the distribution of HI is much more extended and can often be traced to several Holmberg radii. Analysis of emission from HII regions in spirals provides the primary means for determining their metal abundance (in this case the abundance of interstellar gas rather than of stars). Metallicity is found to decrease with radius. As a rule of thumb, the metal abundance decreases by an order of magnitude for a hundred-fold decrease in surface density. The mean metallicity also correlates with luminosity (or stellar mass), with the metal abundance increasing roughly as the square root of stellar mass (see §2.4.4).

(g) Kinematics The stars and cold gas in galaxy disks move in the disk plane on roughly circular orbits. Therefore, the kinematics of a disk are largely specified by its rotation curve $V_{\rm rot}(R)$, which expresses the rotation velocity as a function of galactocentric distance. Disk rotation curves can be measured using a variety of techniques, most commonly optical long-slit or IFU spectroscopy of HII region emission lines, or radio or millimeter interferometry of line emission from the cold gas. Since the HI gas is usually more extended than the ionized gas associated with HII regions, rotation curves can be probed out to larger galactocentric radii using spatially resolved 21-cm observations than using optical emission lines. Fig. 2.21 shows two examples of disk rotation curves. For massive galaxies these typically rise rapidly at small radii and then are almost constant over most of the disk. In dwarf and lower surface brightness systems a slower central rise is common. There is considerable variation from system to system,

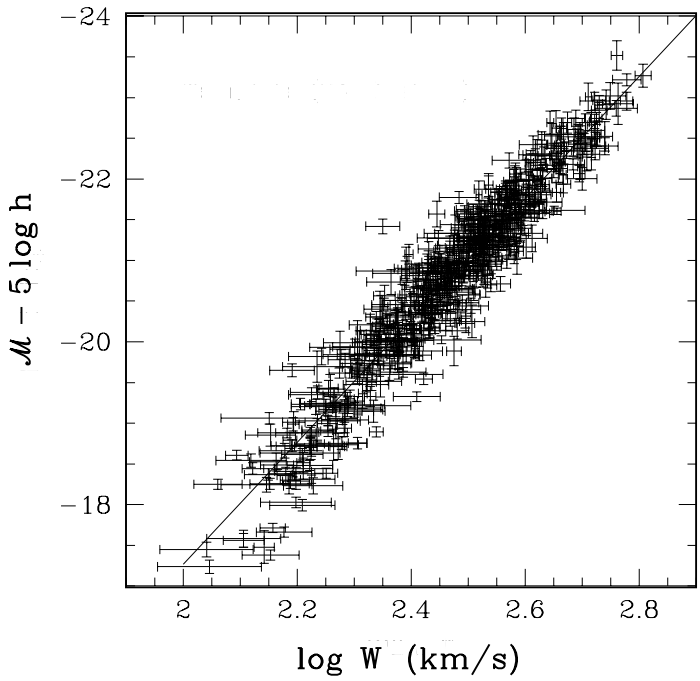


Fig. 2.22. The Tully-Fisher relation in the I-band. Here W is the linewidth of the HI 21 cm line which is roughly equal to twice the maximum rotation velocity, $V_{\rm max}$. [Adapted from Giovanelli et al. (1997) by permission of AAS]

and features in rotation curves are often associated with disk structures such as bars or spiral arms.

The rotation curve is a direct measure of the gravitational force within a disk. Assuming, for simplicity, spherical symmetry, the total enclosed mass within radius r can be estimated from

$$M(r) = rV_{\rm rot}^2(r)/G. \tag{2.33}$$

In the outer region, where $V_{\rm rot}(r)$ is roughly a constant, this implies that $M(r) \propto r$, so that the enclosed mass of the galaxy (unlike its enclosed luminosity) does not appear to be converging. For the rotation curve of NGC 3198 shown in Fig. 2.21, the last measured point corresponds to an enclosed mass of $1.5 \times 10^{11}\,{\rm M}_\odot$, about four times larger than the stellar mass. Clearly, the asymptotic total mass could even be much larger than this. The fact that the observed rotation curves of spiral galaxies are flat at the outskirts of their disks is evidence that they possess massive halos of unseen, dark matter. This is confirmed by studies of the kinematics of satellite galaxies and of gravitational lensing, both suggesting that the enclosed mass continues to increase roughly with radius out to at least ten times the Holmberg radius.

The kinematics of bulges are difficult to measure, mainly because of contamination by disk light. Nevertheless, the existing data suggests that the majority are rotating rapidly (consistent with their flattened shapes being due to the centrifugal forces), and in the same sense as their disk components.

(h) Tully-Fisher Relation Although spiral galaxies show great diversity in luminosity, size, rotation velocity and rotation-curve shape, they obey a well-defined scaling relation between luminosity L and rotation velocity (usually taken as the maximum of the rotation curve well away

from the center, $V_{\rm max}$). This is known as the Tully-Fisher relation, an example of which is shown in Fig. 2.22. The observed Tully-Fisher relation is usually expressed in the form $L = AV_{\rm max}^{\alpha}$, where A is the zero-point and α is the slope. The observed value of α is between 2.5 and 4, and is larger in redder bands (e.g., Pierce & Tully, 1992). For a fixed $V_{\rm max}$, the scatter in luminosity is typically 20 percent. This tight relation can be used to estimate the distances to spiral galaxies, using the principle described in §2.1.3(c). However, as we show in Chapter **??**, the Tully-Fisher relation is also important for our understanding of galaxy formation and evolution, as it defines a relation between dynamical mass (due to stars, gas, and dark matter) and luminosity.

2.3.4 The Milky Way

We know much more about our own Galaxy, the Milky Way, than about most other galaxies, simply because our position within it allows its stellar and gas content to be studied in considerable detail. This 'internal perspective' also brings disadvantages, however. For example, it was not demonstrated until the 1920's and 30's that the relatively uniform brightness of the Milky Way observed around the sky does not imply that we are close to the center of the system, but rather is a consequence of obscuration of distant stars by dust. This complication, combined with the problem of measuring distances, is the main reason why many of the Milky Way's large-scale properties (e.g., its total luminosity, its radial structure, its rotation curve) are still substantially more uncertain than those of some external galaxies.

Nevertheless, we believe that the Milky Way is a relatively normal spiral galaxy. Its main baryonic component is the thin stellar disk, with a mass of $\sim 5\times10^{10}\,{\rm M}_\odot$, a radial scale length of $\sim 3.5\,$kpc, a vertical scale height of $\sim 0.3\,$kpc, and an overall diameter of $\sim 30\,$kpc. The Sun lies close to the midplane of the disk, about 8 kpc from the Galactic Center, and rotates around the center of the Milky Way with a rotation velocity of $\sim 220\,{\rm km\,s^{-1}}$. In addition to this thin disk component, the Milky Way also contains a thick disk whose mass is 10-20 percent of that of the thin disk. The vertical scale height of the thick disk is $\sim 1\,$kpc, but its radial scale length is remarkably similar to that of the thin disk. The thick disk rotates slower than the thin disk, with a rotation velocity at the solar radius of $\sim 175\,{\rm km\,s^{-1}}$.

In addition to the thin and thick disks, the Milky Way also contains a bulge component with a total mass of $\sim 10^{10}\,{\rm M}_\odot$ and a half-light radius of $\sim 1\,$kpc, as well as a stellar halo, whose mass is only about 3 percent of that of the bulge despite its much larger radial extent. The stellar halo has a radial number density distribution $n(r) \propto r^{-\alpha}$, with $2 \lesssim \alpha \lesssim 4$, reaches out to at least 40 kpc, and shows no sign of rotation (i.e., its structure is supported against gravity by random rather than ordered motion). The structure and kinematics of the bulge are more complicated. The near-infrared image of the Milky Way, obtained with the COBE satellite, shows a modest, somewhat boxy bulge. As discussed in §**??**, it is believed that these boxy bulges are actually bars. This bar-like nature of the Milky Way bulge is supported by the kinematics of atomic and molecular gas in the inner few kiloparsecs (Binney et al., 1991), by microlensing measurements of the bulge (Zhao et al., 1995), and by asymmetries in the number densities of various types of stars (Whitelock & Catchpole, 1992; Stanek et al., 1994; Sevenster, 1996). The very center of the Milky Way is also known to harbor a supermassive black hole with a mass approximately $2\times10^{6}\,{\rm M}_\odot$. Its presence is unambiguously inferred from the radial velocities, proper motions and accelerations of stars which pass within 100 astronomical units (1.5×10^{15}cm) of the central object (Genzel et al., 2000; Schödel et al., 2003; Ghez et al., 2005).

During World War II the German astronomer W. Baade was interned at Mount Wilson in California, where he used the unusually dark skies produced by the blackout to study the stellar populations of the Milky Way. He realized that the various components are differentiated not only by their spatial distributions and their kinematics, but also by their age distributions and their chemical compositions. He noted that the disk population (which he called Population

I) contains stars of all ages and with heavy element abundances ranging from about 0.2 to 1 times solar. The spheroidal component (bulge plus halo), which he called Population II, contains predominantly old stars and near the Sun its heavy element abundances are much lower than in the disk. More recent work has shown that younger disk stars are more concentrated to the midplane than older disk stars, that disk stars tend to be more metal-rich near the Galactic center than at large radii, and that young disk stars tend to be somewhat more metal-rich than older ones. In addition, it has become clear that the spheroidal component contains stars with a very wide range of metal abundances. Although the majority are within a factor of 2 or 3 of the solar value, almost the entire metal-rich part of the distribution lies in the bulge. At larger radii the stellar halo is predominantly metal-poor with a metallicity distribution reaching down to very low values: the current record holder has an iron content that is about 200,000 times smaller than that of the Sun! Finally, the relative abundances of specific heavy elements (for example, Mg and Fe) differ systematically between disk and spheroid. As we will see in Chapter **??**, all these differences indicate that the various components of the Milky Way have experienced very different star formation histories (see also §**??**).

The Milky Way also contains about $5 \times 10^9\,M_\odot$ of cold gas, almost all of which is moving on circular orbits close to the plane of the disk. The majority of this gas (~ 80 percent) is neutral, atomic hydrogen (HI), which emits radio emission at 21 cm. The remaining ~ 20 percent of the gas is in molecular form and is most easily traced using millimeter-wave line emission from carbon monoxide (CO). The HI has a scale height of $\sim 150\,$pc and a velocity dispersion of $\sim 9\,\mathrm{km\,s^{-1}}$. Between 4 and 17 kpc its surface density is roughly constant, declining rapidly at both smaller and larger radii. The molecular gas is more centrally concentrated than the atomic gas, and mainly resides in a ring-like distribution at $\sim 4.5\,$kpc from the center, and with a FWHM of $\sim 2\,$kpc. Its scale height is only $\sim 50\,$pc, while its velocity dispersion is $\sim 7\,\mathrm{km\,s^{-1}}$, somewhat smaller than that of the atomic gas. The molecular gas is arranged in molecular cloud complexes with typical masses in the range 10^5 to $10^7\,M_\odot$ and typical densities of order 100 atoms/cc. New stars are born in clusters and associations embedded in the dense, dust-enshrouded cores of these molecular clouds (see Chapter **??**). If a star-forming region contains O and B stars, their UV radiation soon creates an ionized bubble, an "HII region", in the surrounding gas. Such regions produce strong optical line emission which makes them easy to identify and to observe. Because of the (ongoing) star formation, the ISM is enriched with heavy elements. In the solar neighborhood, the metallicity of the ISM is close to that of the Sun, but it decreases by a factor of a few from the center of the disk to its outer edge.

Three other diffuse components of the Milky Way are observed at levels which suggest that they may significantly influence its evolution. Most of the volume of the Galaxy near the Sun is occupied by hot gas at temperatures of about 10^6K and densities around 10^{-4} atoms/cc. This gas is thought to be heated by stellar winds and supernovae and contains much of the energy density of the ISM. A similar energy density resides in relativistic protons and electrons (cosmic rays) which are thought to have been accelerated primarily in supernova shocks. The third component is the Galactic magnetic field which has a strength of a few μG, is ordered on large scales, and is thought to play a significant role in regulating star formation in molecular clouds.

The final and dominant component of the Milky Way appears to be its dark halo. Although the 'dark matter' out of which this halo is made has not been observed directly (except perhaps for a small fraction in the form of compact objects, see §2.10.2), its presence is inferred from the outer rotation curve of the Galaxy, from the high velocities of the most extreme local Population II stars, from the kinematics of globular star clusters and dwarf galaxies in the stellar halo, and from the infall speed of our giant neighbor, the Andromeda nebula. The estimated total mass of this unseen distribution of dark matter is about $10^{12}\,M_\odot$ and it is thought to extend well beyond 100 kpc from the Galactic center.

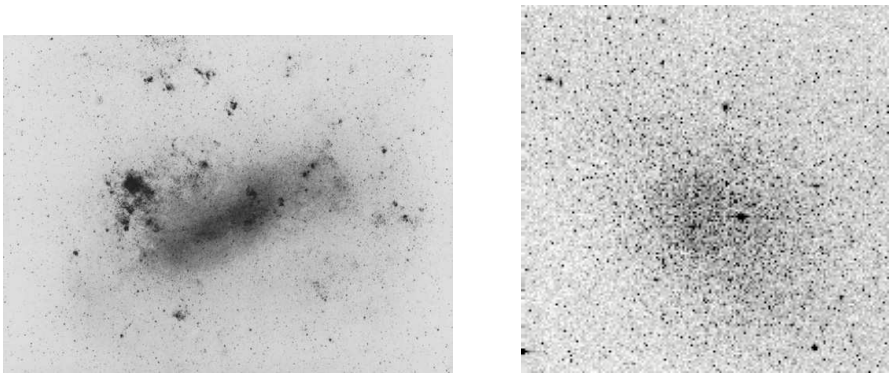


Fig. 2.23. Images of two dwarf galaxies: the Large Magellanic cloud (LMC, left panel), which is a prototypical dwarf irregular, and the dwarf spheroidal Fornax (right panel). [Courtesy of NASA/IPAC Extragalactic Database]

2.3.5 Dwarf Galaxies

For historical reasons, galaxies with $\mathscr{M}_B \gtrsim -18$ are often called dwarf galaxies (Sandage & Binggeli, 1984). These galaxies span roughly six orders of magnitude in luminosity, although the faint end is subject to regular changes as fainter and fainter galaxies are constantly being discovered. The current record holder is Willman I, a dwarf spheroidal galaxy in the local group with an estimated magnitude of $\mathscr{M}_V \simeq -2.6$ (Willman et al., 2005; Martin et al., 2007).

By number, dwarfs are the most abundant galaxies in the Universe, but they contain a relatively small fraction of all stars. Their structure is quite diverse, and they do not fit easily into the Hubble sequence. The clearest separation is between gas-rich systems with ongoing star formation – the dwarf irregulars (dIrr) – and gas-poor systems with no young stars – the dwarf ellipticals (dE) and dwarf spheroidals (dSph). Two examples of them are shown in Fig. 2.23.

Fig. 2.24 sketches the regions in the parameter space of effective radius and absolute magnitude that are occupied by different types of galaxies. Spirals and dwarf irregulars cover roughly four orders of magnitude in luminosity, almost two orders of magnitude in size, and about three orders of magnitude in surface brightness. As their name suggests, dwarf irregulars have highly irregular structures, often being dominated by one or a few bright HII regions. Their gas content increases with decreasing mass and in extreme objects, such as blue compact dwarfs, the so-called 'extragalactic HII regions', the HI extent can be many times larger than the visible galaxy. The larger systems seem to approximate rotationally supported disks, but the smallest systems show quite chaotic kinematics. The systems with regular rotation curves often appear to require substantial amounts of dark matter even within the visible regions of the galaxy.

Dwarf ellipticals are gas-poor systems found primarily in groups and clusters of galaxies. Their structure is regular, with luminosity profiles closer to exponential than to the de Vaucouleurs law (see Fig. 2.13). In addition, they have lower metallicities than normal ellipticals, although they seem to follow the same relation between metallicity and luminosity.

Dwarf spheroidals (dSphs) are faint objects of very low surface brightness, which have so far only been identified unambiguously within the Local Group (see §2.5.2). Their structure is relatively regular and they appear to contain no gas and no, or very few, young stars with ages less than about 1 Gyr. However, several dSphs show unambiguous evidence for several distinct bursts of star formation. Their typical sizes range from a few tens to several hundreds of parsec, while their luminosities span almost five orders of magnitude. Their kinematics indicate dynamical mass-to-light ratios that can be as large as several hundreds times that of the Sun, which is

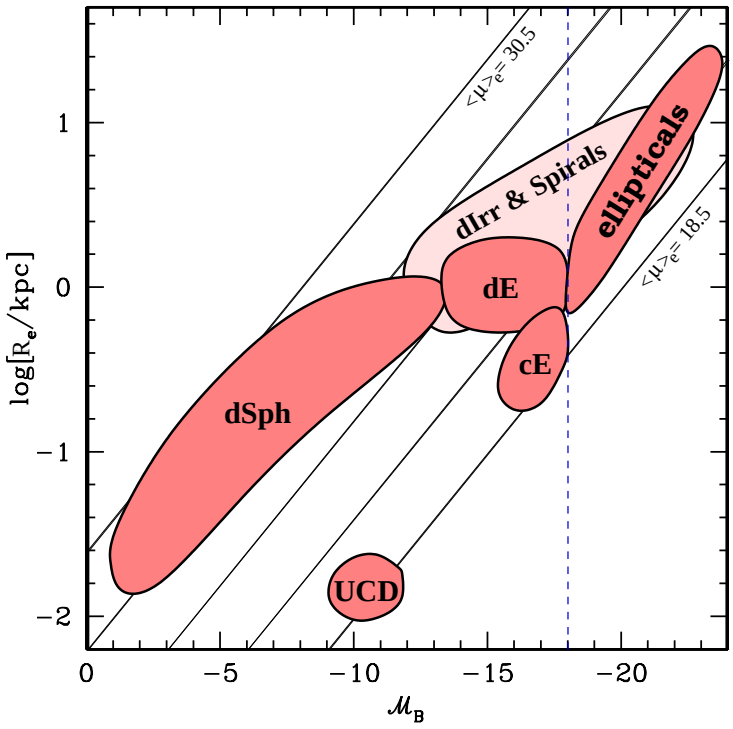


Fig. 2.24. A sketch of the regions in the parameter space of effective radius and absolute magnitude (both in the *B*-band) occupied by different types of galaxies. The spheroidal systems are split in ellipticals, dwarf ellipticals (dE), compact ellipticals (cE), dwarf spheroidals (dSph), and ultra-compact dwarfs (UCD). The dashed, vertical line corresponds to $\mathscr{M}_B = -18$, and reflects the magnitude limit below which galaxies are often classified as dwarfs. The diagonal lines are lines of constant surface brightness; galaxies roughly span 5 orders of magnitude in surface brightness, from $\langle\mu_B\rangle_e \sim -18.5$ to $\langle\mu_B\rangle_e \sim -30.5$.

usually interpreted as implying a large dark matter content (Mateo, 1998; Gilmore et al., 2007). One of the most luminous dSphs, the Sagittarius dwarf, currently lies only about 20 kpc from the center of the Milky Way and is being torn apart by the Milky Way's tidal forces.

The distinction between 'dwarf' and 'regular' galaxies had its origin in the observation that ellipticals with $\mathscr{M}_B \gtrsim -18$ are not well described by the de Vaucouleurs $R^{1/4}$-law. Instead, their surface brightness profiles were found to be closer to exponential (e.g., Faber & Lin, 1983; Binggeli et al., 1984). This distinction was further strengthened by the work of Kormendy (1985) who found that bright ellipticals have their surface brightness decrease with increasing luminosity, while dEs have increasing surface brightness with increasing luminosity (see Fig. 2.14). This gave rise to the concept of a clear dichotomy between dwarf and regular ellipticals. More recently, however, it has been argued that this 'dichotomy', with a characteristic scale at $\mathscr{M}_B \simeq -18$, is an artefact of sample selection and of the fact that the surface brightness profiles were fit with either an $R^{1/4}$-profile or an exponential. Fitting with the more general Sérsic profiles instead indicates clearly that there is a smooth trend between the best-fit Sérsic index and absolute magnitude (see Fig. 2.13) and an equally smooth trend between absolute magnitude and *central* surface brightness (see Graham & Guzmán, 2003, and references therein). Hence, there seems to be no clear distinction between dEs and 'regular' ellipticals. Neither is there a clear distinction between dEs and dSphs; the latter simply make up the low luminosity extreme

of the dEs, typically with $\mathscr{M}_B \gtrsim -14$. Although we will adhere to the 'historical' nomenclature throughout this book, we caution that there is no clear physical motivation for discriminating between dSphs, dEs, and 'regular' ellipticals (but see §**??**).

Fig. 2.24 also sketches the location in size-luminosity space occupied by a special class of (dwarf) galaxies known as compact ellipticals (cEs). These are characterized by unusually high surface brightness for their luminosity, although they do seem to form a smooth continuation of the size-luminosity relation of 'regular' ellipticals. The proto-typical example is M32, a companion of the Andromeda galaxy M31. Compact ellipticals are very rare, and only a handful of these systems are known. Some authors have argued that the bulges of (early-type) disk galaxies occupy the same region in parameter space as the cEs, suggesting that these two types of objects are somehow related (e.g., Bender et al., 1992). Finally, Drinkwater et al. (2003) have recently identified a new class of (potential) galaxies, called ultra-compact dwarfs (UCDs). They typically have $\mathscr{M}_B \sim -11$ and effective radii of 10 to 20 pc, giving them an average surface brightness comparable to that of cEs. Their nature is still very uncertain. In particular, it is still unclear whether they should be classified as galaxies, or whether they merely reflect the bright end of the population of globular clusters. Alternatively, they may also be the remnant nuclei of disrupted low surface brightness galaxies (see below).

2.3.6 Nuclear Star Clusters

In their landmark study of the Virgo cluster, Binggeli et al. (1987) found that $\sim 25\%$ of the dEs contain a massive star cluster at their centers (called the nucleus), which clearly stands out against the low surface brightness of its host galaxy. Following this study it has become customary to split the population of dEs into 'nucleated' and 'non-nucleated'. Binggeli et al. (1987) did not detect any nuclei in the more luminous ellipticals, although they cautioned that these might have been missed in their photographic survey due to the high surface brightness of the underlying galaxy. Indeed, more recent studies, capitalizing on the high spatial resolution afforded by the HST, have found that as much as $\sim 80\%$ of all early-type galaxies with $\mathscr{M}_B \lesssim -15$ are nucleated (e.g., Grant et al., 2005; Côté et al., 2006). In addition, HST imaging of late-type galaxies has revealed that 50-70% of these systems also have compact stellar clusters near their photometric centers (e.g., Phillips et al., 1996; Böker et al., 2002). These show a remarkable similarity in luminosity and size to those detected in early-type galaxies. However, the nuclear star clusters in late-type galaxies seem to have younger stellar ages than their counterparts in early-type galaxies (e.g., Walcher et al., 2005; Côté et al., 2006). Thus a large fraction of all galaxies, independent of their morphology, environment or gas content, contain a nuclear star cluster at their photometric center. The only exception seem to be the brightest ellipticals, with $\mathscr{M}_B \lesssim -20.5$, which seem to be devoid of nuclear star clusters. Note that this magnitude corresponds to the transition from disky, power-law ellipticals to boxy, core ellipticals (see §2.3.2), supporting the notion of a fundamental transition at this luminosity scale.

On average, nuclear star clusters are an order of magnitude more luminous than the peak of the globular cluster luminosity function of their host galaxies, have stellar masses in the range $\sim 10^6 - 10^8\,\mathrm{M}_\odot$, and typical radii of $\sim 5\,\mathrm{pc}$. This makes nuclear star clusters the densest stellar systems known (e.g., Geha et al., 2002; Walcher et al., 2005). In fact, they are not that dissimilar to the ultra-compact dwarfs, suggesting a possible relation (e.g., Bekki et al., 2001).

As discussed in §2.3.2 (see also §**??**), the majority of bright spheroids (ellipticals and bulges) seem to contain a supermassive black hole (SMBH) at their nucleus. The majority of spheroids with secure SMBH detections have magnitudes in the range $-22 \lesssim \mathscr{M}_B \lesssim -18$. Although it is unclear whether (the majority of) fainter spheroids also harbor SMBHs, current data seems to support a view in which bright galaxies ($\mathscr{M}_B \lesssim -20$) often, and perhaps always, contain SMBHs but not stellar nuclei, while at the faint end ($\mathscr{M}_B \gtrsim -18$) stellar nuclei become the dominant

feature. Intriguingly, Ferrarese et al. (2006a) have shown that stellar nuclei and SMBHs obey a common scaling relation between their mass and that of their host galaxy, with $M_{\rm CMO}/M_{\rm gal} = 0.018^{+0.034}_{-0.012}$ (where CMO stands for Central Massive Object), suggesting that SMBHs and nuclear clusters share a common origin. This is somewhat clouded, though, by the fact that nuclear star clusters and SMBHs are not mutually exclusive. The two best known cases in which SMBHs and stellar nuclei coexist are M32 (Verolme et al., 2002) and the Milky Way (Ghez et al., 2003; Schödel et al., 2003).

2.3.7 Starbursts

In normal galaxies like the Milky Way, the specific star formation rates are typically of order $0.1\,{\rm Gyr}^{-1}$, which implies star formation time scales (defined as the ratio between the total stellar mass and the current star formation rate) that are comparable to the age of the Universe. There are, however, systems in which the (specific) star formation rates are 10 or even 100 times higher, with implied star formation time scales as short as 10^8 years. These galaxies are referred to as starbursts. The star formation activity in such systems (at least in the most massive ones) is often concentrated in small regions, with sizes typically about 1 kpc, much smaller than the disk sizes in normal spiral galaxies.

Because of the large current star formation rate, a starburst contains a large number of young stars. Indeed, for blue starbursts where the star formation regions are not obscured by dust, their spectra generally have strong blue continuum produced by massive stars, and show strong emission lines from HII regions produced by the UV photons of O and B stars (see Fig. 2.12). Since the formation of stars is, in general, associated with the production of large amounts of dust,† most of the strong starbursts are not observed directly via their strong UV emission. Rather, the UV photons produced by the young stars are absorbed by dust and re-emitted in the far-infrared. In extreme cases these starbursting galaxies emit the great majority of their light in the infrared, giving rise to the population of infrared luminous galaxies (LIRGs) discovered in the 1980s with the Infrared Astronomical Satellite (IRAS). A LIRG is defined as a galaxy with a far-infrared luminosity exceeding $10^{11}\,{\rm L}_\odot$ (Soifer et al., 1984). If its far-infrared luminosity exceeds $10^{12}\,{\rm L}_\odot$ it is called an ultraluminous infrared galaxies (ULIRG).

The fact that starbursts are typically confined to a small region (usually the nucleus) of the starbursting galaxy, combined with their high star formation rates, requires a large amount of cold gas to be accumulated in a small region in a short time. The most efficient way of achieving this is through mergers of gas-rich galaxies, where the interstellar media of the merging systems can be strongly compressed and concentrated by tidal interactions (see §**??**). This scenario is supported by the observation that massive starbursts (in particular ULIRGs) are almost exclusively found in strongly interacting systems with peculiar morphologies.

2.3.8 Active Galactic Nuclei

The centers of many galaxies contain small, dense and luminous components known as active galactic nuclei (AGN). An AGN can be so bright that it outshines its entire host galaxy, and differs from a normal stellar system in its emission properties. While normal stars emit radiation primarily in a relatively narrow wavelength range between the near-infrared and the near-UV, AGN are powerful emitters of non-thermal radiation covering the entire electromagnetic spectrum from the radio to the gamma-ray regime. Furthermore, the spectra of many AGN contain strong emission lines and so contrast with normal stellar spectra which are typically dominated by absorption lines (except for galaxies with high specific star formation rates). According to

† It is believed that dust is formed in the atmospheres of evolved stars and in supernova explosions.

Table 2.6. *Relative Number Densities of Galaxies in the Local Universe*

Type of object	Number density
Spirals	1
Lenticulars	0.1
Ellipticals	0.2
Irregulars	0.05
Dwarf galaxies	10
Peculiar galaxies	0.05
Starbursts	0.1
Seyferts	10^{-2}
Radio galaxies	10^{-4}
QSOs	10^{-5}
Quasars	10^{-7}

their emission properties, AGN are divided into a variety of sub-classes, including radio sources, Seyferts, liners, blazars and quasars (see Chapter **??** for definitions).

Most of the emission from an AGN comes from a very small, typically unresolved region; high-resolution observations of relatively nearby objects with HST or with radio interferometry demonstrate the presence of compact emitting regions with sizes smaller than a few parsecs. These small sizes are consistent with the fact that some AGN reveal strong variability on time scales of only a few days, indicating that the emission must emanate from a region not much larger than a few light-days across. The emission from these nuclei typically reveals a relatively featureless power-law continuum at radio, optical and X-ray wavelengths, as well as broad emission lines in the optical and X-ray bands. On somewhat larger scales, AGN often manifest themselves in radio, optical and even X-ray jets, and in strong but narrow optical emission lines from hot gas. The most natural explanation for the energetics of AGN, combined with their small sizes, is that AGN are powered by the accretion of matter onto a supermassive black hole (SMBH) with a mass of 10^6 to $10^9\,\mathrm{M}_\odot$. Such systems can be extremely efficient in converting gravitational energy into radiation. As mentioned in §2.3.2, virtually all spheroidal galaxy components (i.e., ellipticals and bulges) harbor a SMBH whose mass is tightly correlated with that of the spheroid, suggesting that the formation of SMBHs is tightly coupled to that of their host galaxies. Indeed, the enormous energy output of AGN may have an important feedback effect on the formation and evolution of galaxies. Given their importance for galaxy formation, Chapter **??** is entirely devoted to AGN, including a more detailed overview of their observational properties.

2.4 Statistical Properties of the Galaxy Population

So far our description has focused on the properties of separate classes of galaxies. We now turn our attention to statistics that describe the galaxy population as a whole, i.e., that describe how galaxies are distributed with respect to these properties. As we will see in §§2.5 and 2.7, the galaxy distribution is strongly clustered on scales up to ~ 10 Mpc, which implies that one needs to probe a large volume in order to obtain a sample that is representative of the entire population. Therefore, the statistical properties of the galaxy population are best addressed using large galaxy redshift surveys. Currently the largest redshift surveys available are the two-degree Field Galaxy Redshift Survey (2dFGRS; Colless et al., 2001) and the Sloan Digital Sky Survey (SDSS; York et al., 2000), both of which probe the galaxy distribution at a median redshift $z \sim 0.1$. The 2dFGRS has measured redshifts for $\sim 220,000$ galaxies over ~ 2000 square degrees down to

a limiting magnitude of $b_J < 19.45$. The source catalogue for the survey is the APM galaxy catalogue, which is based on Automated Plate Measuring machine (APM) scans of photographic plates (Maddox et al., 1990b). The SDSS consists of a photometrically and astrometrically calibrated imaging survey covering more than a quarter of the sky in five broad optical bands (u, g, r, i, z) that were specially designed for the survey (Fukugita et al., 1996), plus a spectroscopic survey of $\sim 10^6$ galaxies ($r < 17.77$) and $\sim 10^5$ quasars detected in the imaging survey.

The selection function of these and other surveys plays an important role in the observed sample properties. For example, most surveys select galaxies above a given flux limit (i.e., the survey is complete down to a given apparent magnitude). Since intrinsically brighter galaxies will reach the flux limit at larger distances, a flux limited survey is biased towards brighter galaxies. This is called the Malmquist bias and needs to be corrected for when trying to infer the intrinsic probability distribution of galaxies. There are two ways to do this. One is to construct a volume limited sample, by only selecting galaxies brighter than a given absolute magnitude limit, $M_{\rm lim}$, and below a given redshift, $z_{\rm lim}$, where $z_{\rm lim}$ is the redshift at which a galaxy with absolute magnitude $M_{\rm lim}$ has an apparent magnitude equal to the survey limit. Alternatively, one can weight each galaxy by the inverse of $V_{\rm max}$, defined as the survey volume out to which the specific galaxy in question could have been detected given the flux limit of the survey. The advantage of this method over the construction of volume-limited samples is that one does not have to discard any data. However, the disadvantage is that intrinsically faint galaxies can only be seen over a relatively small volume (i.e., $V_{\rm max}$ is small), so that they get very large weights. This tends to make the measurements extremely noisy at low luminosities.

As a first example of a statistical description of the galaxy population, Table 2.6 lists the number densities of the various classes of galaxies described in the previous section, relative to that of spiral galaxies. Note, however, that these numbers are only intended as a rough description of the galaxy population in the nearby Universe. The real galaxy population is extremely diverse, and an accurate description of the galaxy number density is only possible for a well-defined sample of galaxies.

2.4.1 Luminosity Function

Arguably one of the most fundamental properties of a galaxy is it luminosity (in some waveband). An important statistic of the galaxy distribution is therefore the luminosity function, $\phi(L)\mathrm{d}L$, which describes the number density of galaxies with luminosities in the range $L \pm \mathrm{d}L$. Fig. 2.25 shows the luminosity function in the photometric b_J-band obtained from the 2dFGRS. At the faint end $\phi(L)$ seems to follow a power-law which truncates at the bright end, where the number density falls roughly exponentially. A similar behavior is also seen in other wavebands, so that the galaxy luminosity function is commonly fitted by a Schechter function (Schechter, 1976) of the form

$$\phi(L)\mathrm{d}L = \phi^* \left(\frac{L}{L^*}\right)^{\alpha} \exp\left(-\frac{L}{L^*}\right) \frac{\mathrm{d}L}{L^*} \,. \tag{2.34}$$

Here L^* is a characteristic luminosity, α is the faint-end slope, and ϕ^* is an overall normalization. As shown in Fig. 2.25, this function fits the observed luminosity function over a wide range. From the Schechter function, we can write the mean number density, $n_{\rm g}$, and the mean luminosity density, $\mathscr{L}$, of galaxies in the Universe as

$$n_{\rm g} \equiv \int_0^\infty \phi(L)\,\mathrm{d}L = \phi^* \Gamma(\alpha+1)\,, \tag{2.35}$$

and

$$\mathscr{L} \equiv \int_0^\infty \phi(L)\,L\,\mathrm{d}L = \phi^* L^* \Gamma(\alpha+2)\,, \tag{2.36}$$

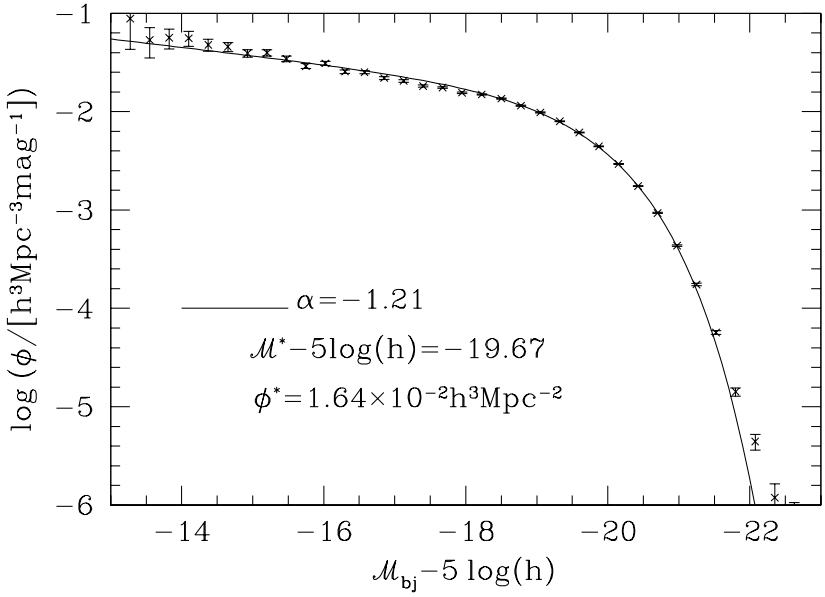


Fig. 2.25. The luminosity function of galaxies in the b_j-band as obtained from the 2-degree Field Galaxy Redshift Survey. [Based on data published in Norberg et al. (2002b)]

where $\Gamma(x)$ is the Gamma function. Note that n_g diverges for $\alpha \leq -1$, while $\mathscr{L}$ diverges for $\alpha \leq -2$. Observations from the near-UV to the near-infrared show that $-2 < \alpha < -1$, indicating that the number density is dominated by faint galaxies while the luminosity density is dominated by bright ones.

As we will see in Chapter **??**, the luminosity function of galaxies depends not only on the waveband, but also on the morphological type, the color, the redshift, and the environment of the galaxy. One of the most challenging problems in galaxy formation is to explain the general shape of the luminosity function and the dependence on other galaxy properties.

2.4.2 Size Distribution

Size is another fundamental property of a galaxy. As shown in Figs. 2.14 and 2.20, galaxies of a given luminosity may have very different sizes (and therefore surface brightnesses). Based on a large sample of galaxies in the SDSS, Shen et al. (2003) found that the size distribution for galaxies of a given luminosity L can roughly be described by a lognormal function,

$$P(R|L)\,\mathrm{d}R = \frac{1}{\sqrt{2\pi}\sigma_{\ln R}} \exp\left[-\frac{\ln^2(R/\overline{R})}{2\sigma_{\ln R}^2}\right] \frac{\mathrm{d}R}{R}\,, \tag{2.37}$$

where $\overline{R}$ is the median and $\sigma_{\ln R}$ the dispersion. Fig. 2.26 shows that $\overline{R}$ increases with galaxy luminosity roughly as a power law for both early-type and late-type galaxies, and that the dependence is stronger for early types. The dispersion $\sigma_{\ln R}$, on the other hand, is similar for both early and late type galaxies, decreasing from ~ 0.5 for galaxies with $M_r \gtrsim -20.5$ to ~ 0.25 for brighter galaxies.

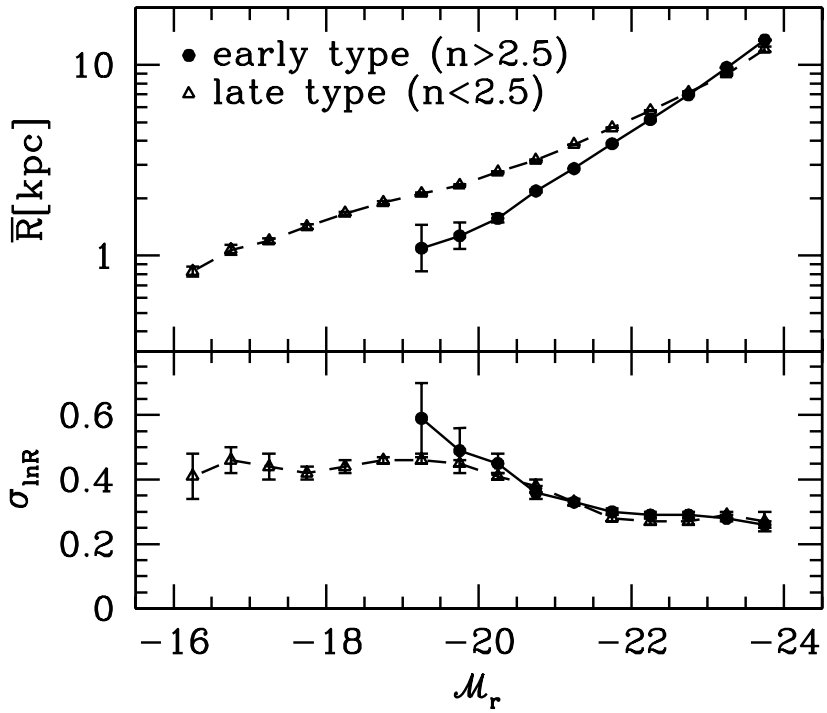


Fig. 2.26. The median (upper panel) and dispersion (lower panel) of the size distribution of galaxies in the SDSS as function of their *r*-band magnitude. Results are shown separately for early-type (solid dots) and late-type (open triangles) galaxies defined according to the Sérsic index *n*. [Kindly provided to us by S. Shen, based on data published in Shen et al. (2003)]

2.4.3 Color Distribution

As shown in Fig. 2.5, massive stars emit a larger fraction of their total light at short wavelengths than low-mass stars. Since more massive stars are in general shorter-lived, the color of a galaxy carries important information about its star formation history. However, the color of a star also depends on its metallicity, in the sense that stars with higher metallicities are redder. In addition, dust extinction is more efficient at bluer wavelengths, so that the color of a galaxy also contains information regarding its chemical composition and dust content.

The left panel of Fig. 2.27 shows the distribution of the $^{0.1}(g-r)$ colors of galaxies in the SDSS, where the superscript indicates that the magnitudes have been converted to the same rest-frame wavebands at $z = 0.1$. The most salient characteristic of this distribution is that it is clearly bimodal, revealing a relatively narrow peak at the red end of the distribution plus a significantly broader distribution at the blue end. To first order, this simply reflects that galaxies come in two different classes: early-type galaxies, which have relatively old stellar populations and are therefore red, and late-type galaxies, which have ongoing star formation in their disks, and are therefore blue. However, it is important to realize that this color-morphology relation is not perfect: a disk galaxy may be red due to extensive dust extinction, while an elliptical may be blue if it had a small amount of star formation in the recent past.

The bimodality of the galaxy population is also evident from the color-magnitude relation, plotted in the right-hand panel of Fig. 2.27. This shows that the galaxy population is divided into a red sequence and a blue sequence (also sometimes called the blue cloud). Two trends are noteworthy. First of all, at the bright end the red sequence dominates, while at the faint end the

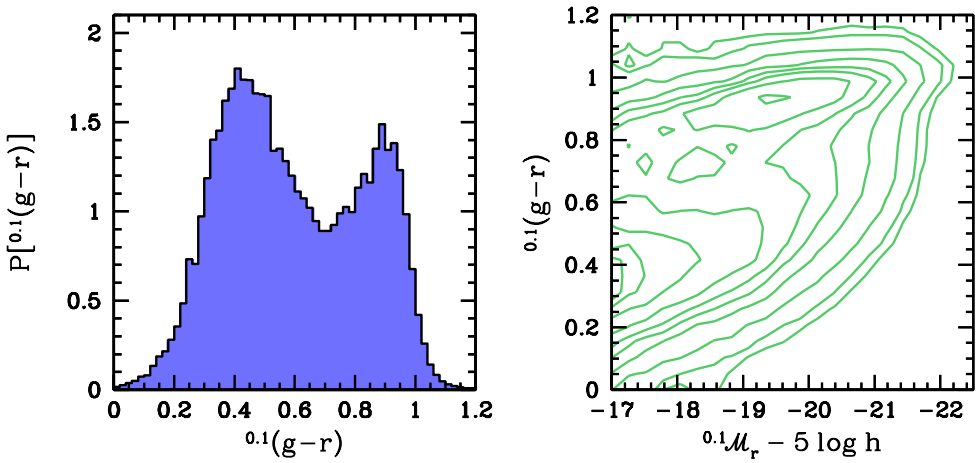


Fig. 2.27. The probability density of galaxy colors (left) and the color-magnitude relation (right) of $\sim 365,000$ galaxies in the SDSS. Each galaxy has been weighted by $1/V_{\rm max}$ to correct for Malmquist bias. Note the pronounced bimodality in the color distribution, and the presence of both a red sequence and a blue sequence in the color-magnitude relation.

majority of the galaxies are blue. As we will see in Chapter **??**, this is consistent with the fact that the bright (faint) end of the galaxy luminosity function is dominated by early-type (late-type) galaxies. Secondly, within each sequence brighter galaxies appear to be redder. As we will see in Chapters **??** and **??** this most likely reflects that the stellar populations in brighter galaxies are both older and more metal rich, although it is still unclear which of these two effects dominates, and to what extent dust plays a role.

2.4.4 The Mass-Metallicity Relation

Another important parameter to characterize a galaxy is its average metallicity, which reflects the amount of gas that has been reprocessed by stars and exchanged with its surroundings. One can distinguish two different metallicities for a given galaxy: the average metallicity of the stars and that of the gas. Depending on the star formation history and the amount of inflow and outflow, these metallicities can be significantly different. Gas-phase metallicities can be measured from the emission lines in a galaxy spectrum, while the metallicity of the stars can be obtained from the absorption lines which originate in the atmospheres of the stars.

Fig. 2.28 shows the relation between the gas-phase oxygen abundance and the stellar mass of SDSS galaxies. The oxygen abundance is expressed as $12 + \log[({\rm O/H})]$, where O/H is the abundance by number of oxygen relative to hydrogen. Since the measurement of gas-phase abundances requires the presence of emission lines in the spectra, all these galaxies are still forming stars, and the sample is therefore strongly biased towards late-type galaxies. Over about three orders of magnitude in stellar mass the average gas-phase metallicity increases by an order of magnitude. The relation is remarkably tight and reveals a clear flattening above a few times $10^{10}\,{\rm M}_\odot$. The average stellar metallicity follows a similar trend with stellar mass but with much larger scatter at the low mass end (Gallazzi et al., 2005). An interpretation of these results in terms of the chemical evolution of galaxies is presented in Chapter **??**.

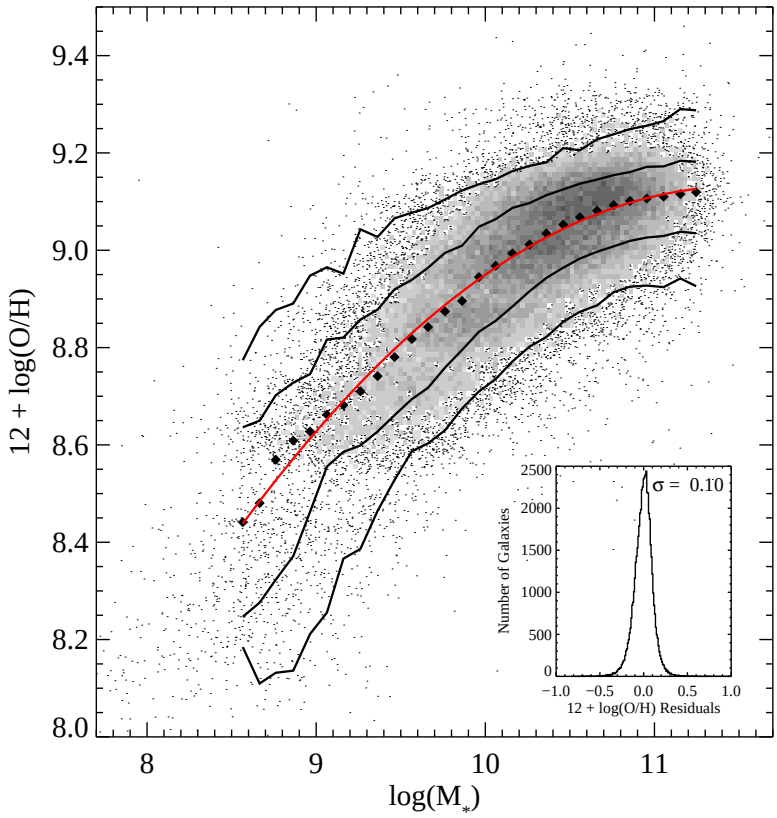


Fig. 2.28. The relation between stellar mass, in units of solar masses, and the gas-phase oxygen abundance for ∼53,400 star-forming galaxies in the SDSS. For comparison, the Sun has $12 + \log[(\mathrm{O/H})] = 8.69$. The large black points represent the median in bins of 0.1 dex in mass. The solid lines are the contours which enclose 68% and 95% of the data. The gray line shows a polynomial fit to the data. The inset shows the residuals of the fit. [Adapted from Tremonti et al. (2004) by permission of AAS]

2.4.5 Environment Dependence

As early as the 1930s it was realized that the morphological mix of galaxies depends on environment, with denser environments (e.g., clusters, see §2.5.1) hosting larger fractions of early-type galaxies (Hubble & Humason, 1931). This morphology-density relation was quantified more accurately in a paper by Dressler (1980b), who studied the morphologies of galaxies in 55 clusters and found that the fraction of spiral galaxies decreases from ∼ 60 percent in the lowest density regions to less than 10 percent in the highest density regions, while the elliptical fraction basically reveals the opposite behavior (see Fig. 2.29). Note that the fraction of S0 galaxies is significantly higher in clusters than in the general field, although there is no strong trend of S0 fraction with density within clusters.

More recently, the availability of large galaxy redshift surveys has paved the way for far more detailed studies into the environment dependence of galaxy properties. It is found that in addition to a larger fraction of early-type morphologies, denser environments host galaxies that are on average more massive, redder, more concentrated, less gas-rich, and have lower specific star formation rates (e.g., Kauffmann et al., 2004; Baldry et al., 2006; Weinmann et al., 2006). Interpreting these findings in terms of galaxy formation processes, however, is complicated by the fact that various galaxy properties are strongly correlated even at a fixed environment. An

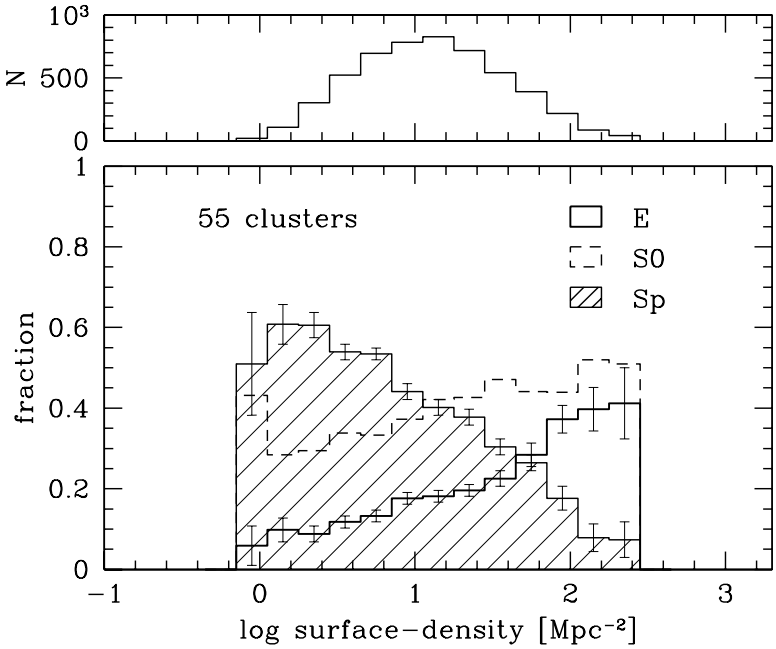


Fig. 2.29. The morphology-density relation, which shows the fractions of galaxies of individual morphological types as functions of galaxy surface number density. The lower panel shows such relations for 55 clusters, while the upper panel shows the number of galaxies in each density bin. [After Dressler (1980a)]

important outstanding question, therefore, is which relationship with environment is truly causal, and which are just reflections of other correlations that are actually independent of environment (see §**??** for a more detailed discussion).

2.5 Clusters and Groups of Galaxies

A significant fraction of the galaxies in the present-day Universe is collected into groups and clusters in which the number density of galaxies is a few tens to a few hundreds times higher than the average. The densest and most populous of these aggregations are called galaxy clusters, which typically contain more than 50 relatively bright galaxies in a volume only a few megaparsecs across. The smaller, less populous aggregations are called 'groups', although there is no well defined distinction. Groups and clusters are the most massive, virialized objects in the Universe, and they are important laboratories to study the evolution of the galaxy population. Because of their high surface densities and large number of very luminous member galaxies, they can be identified out to very large distances, making them also useful as cosmological probes. In this section we summarize some of their most important properties, focusing in particular on their populations of galaxies.

2.5.1 Clusters of Galaxies

In order to select clusters (or groups) of galaxies from the observed galaxy distribution, one needs to adopt some selection criteria. In order for the selected clusters to be dynamically significant,

two selection criteria are usually set. One is that the selected system must have high enough density, and the other is that the system must contain a sufficiently large number of galaxies.

According to these criteria, Abell (1958) selected 1682 galaxy clusters from the Palomar Sky Survey, which are now referred to as the Abell clusters. The two selection criteria set by Abell are

(i) Richness criterion: each cluster must have at least 50 member galaxies with apparent magnitudes $m < m_3 + 2$, where m_3 is the apparent magnitude of the third brightest member. The richness of a cluster is defined to be the number of member galaxies with apparent magnitudes between m_3 and $m_3 + 2$. Rich Abell clusters are those with richness greater than 50, although Abell also listed poor clusters with richness in the range from 30 to 50.

(ii) Compactness criterion: only galaxies with distances to the cluster center smaller than $1.5\,h^{-1}$Mpc (the Abell radius) are selected as members. Given the richness criterion, the compactness criterion is equivalent to a density criterion.

Abell also classified a cluster as regular if its galaxy distribution is more or less circularly symmetric and concentrated, otherwise as irregular. The two most well-studied clusters, because of their proximity, are the Virgo cluster and the Coma cluster. The Virgo cluster, which is the rich cluster nearest to our Galaxy, is a very representative example. It lacks clear symmetry, and reveals significant substructure, indicating that the dynamical relaxation on the largest scales is not yet complete. The Coma cluster, on the other hand, is a fairly rare species. It is extremely massive, and is richer than 95% of all clusters catalogued by Abell. Furthermore, it appears remarkably relaxed, with a highly concentrated and symmetric galaxy distribution with no sign of significant subclustering.

The Abell catalogue was constructed using visual inspections of photographic sky plates. Since its publication, this has been improved upon using special purpose scanning machines (such as the APM at Cambridge and COSMOS at Edinburgh), which resulted in digitized versions of the photographic plates allowing for a more objective identification of clusters (e.g., Lumsden et al., 1992; Dalton et al., 1997). More recently, several cluster catalogues have been constructed from large galaxy redshift surveys such as the 2dFGRS and the SDSS (e.g. Bahcall et al., 2003; Miller et al., 2005; Koester et al., 2007). Based on all these catalogues it is now well established that the number density of rich clusters is of the order of $10^{-5}\ h^3\,\mathrm{Mpc}^{-3}$, about 1000 times smaller than that of L^* galaxies.

(a) Galaxy Populations As we have seen in §2.4.5, clusters are in general rich in early-type galaxies. The fraction of E+SO galaxies is about 80% in regular clusters, and about 50% in irregular clusters, compared to about 30% in the general field. This is generally interpreted as evidence that galaxies undergo morphological transformations in dense (cluster) environments, and various mechanisms have been suggested for such transformations (see §**??**).

The radial number density distribution of galaxies in clusters is well described by $n(r) \propto 1/[r^\gamma(r+r_s)^{3-\gamma}]$, where r_s is a scale radius and γ is the logarithmic slope of the inner profile. The value of γ is typically ~ 1 and the scale radius is typically $\sim 20\%$ of the radius of the cluster (e.g., van der Marel et al., 2000; Lin et al., 2004). As we will see in Chapter **??** this is very similar to the density distribution of dark matter halos, suggesting that within clusters galaxies are a reasonably fair tracer of the mass distribution. There is, however, evidence for some segregation by mass and morphology/color, with more massive, red, early-type galaxies following a more concentrated number density distribution than less massive, blue, late-type galaxies (e.g., Quintana, 1979; Carlberg et al., 1997; Adami et al., 1998; Yang et al., 2005a; van den Bosch et al., 2008).

Often the brightest cluster galaxy (BCG) has an extraordinarily diffuse and extended outer

envelope, in which case it is called a cD galaxy (where the 'D' stands for diffuse). They typically have best-fit Sérsic indices that are much larger than four, and are often located at or near the center of the cluster (because of this, it is useful mnemonic to think of "cD" as meaning "centrally dominant"). cD galaxies are the most massive galaxies known, with stellar masses often exceeding $10^{12}\,\mathrm{M}_\odot$, and their light can make up as much as $\sim 30\%$ of the entire visible light of a rich cluster of galaxies. However, it is unclear whether the galaxy's diffuse envelope should be considered part of the galaxy or as 'intracluster light' (ICL), stars associated with the cluster itself rather than with any particular galaxy. In a few cD galaxies the velocity dispersion appears to rise strongly in the extended envelope, approaching value similar to that of the cluster in which the galaxy is embedded. This supports the idea that these stars are more closely associated with the cluster than with the galaxy (i.e. they are the cluster equivalent of the stellar halo in the Milky Way) . cD galaxies are believed to have grown through the accretion of multiple galaxies in the cluster, a process called galactic cannibalism (see §**??**). Consistent with this, nearby cD's frequently appear to have multiple nuclei (e.g., Schneider et al., 1983)

(b) The Butcher-Oemler Effect When studying the galaxy populations of clusters at intermediate redshifts ($0.3 \lesssim z \lesssim 0.5$), Butcher & Oemler (1978) found a dramatic increase in the fraction of blue galaxies compared to present day clusters, which has become known as the Butcher-Oemler effect. Although originally greeted with some skepticism (see Dressler, 1984, for a review), this effect has been confirmed by numerous studies. In addition, morphological studies, especially those with the HST, have shown that the Butcher-Oemler effect is associated with an increase of the spiral fraction with increasing redshift, and that many of these spirals show disturbed morphologies (e.g., Couch et al., 1994; Wirth et al., 1994).

In addition, spectroscopic data has revealed that a relatively large fraction of galaxies in clusters at intermediate redshifts have strong Balmer lines in absorption and no emission lines (Dressler & Gunn, 1983). This indicates that these galaxies were actively forming stars in the past, but had their star formation quenched in the last 1 to 2 Gyr. Although they were originally named 'E+A' galaxies, currently they are more often referred to as 'k+a' galaxies or as post-starburst galaxies (since their spectra suggest that they must have experienced an elevated amount of star formation prior to the quenching). Dressler et al. (1999) have shown that the fraction of k+a galaxies in clusters at $z \sim 0.5$ is significantly larger than in the field at similar redshifts, and that they have mostly spiral morphologies.

All these data clearly indicate that the population of galaxies in clusters is rapidly evolving with redshift, most likely due to specific processes that operate in dense environments (see §**??**).

(c) Mass Estimates Galaxies are moving fast in clusters. For rich clusters, the typical line-of-sight velocity dispersion, $\sigma_{\rm los}$, of cluster member galaxies is of the order of $1000\,\mathrm{km\,s^{-1}}$. If the cluster has been relaxed to a static dynamical state, which is roughly true for regular clusters, one can infer a dynamical mass estimate from the virial theorem (see §**??**) as

$$M = A\frac{\sigma_{\rm los}^2 R_{\rm cl}}{G}\,, \tag{2.38}$$

where A is a pre-factor (of order unity) that depends on the density profile and on the exact definition of the cluster radius $R_{\rm cl}$. Using this technique one obtains a characteristic mass of $\sim 10^{15} h^{-1}\,\mathrm{M}_\odot$ for rich clusters of galaxies. Together with the typical value of the total luminosity in a cluster, this implies a typical mass-to-light ratio for clusters,

$$(M/L_B)_{\rm cl} \sim 350 h(\mathrm{M}_\odot/\mathrm{L}_\odot)_B\,. \tag{2.39}$$

Hence, only a small fraction of the total gravitational mass of a cluster is associated with galaxies.

Ever since the first detection by the UHURU satellite in the 1970s, it has become clear that clusters are bright X-ray sources, with characteristic luminosities ranging from $L_X \sim 10^{43}$ to

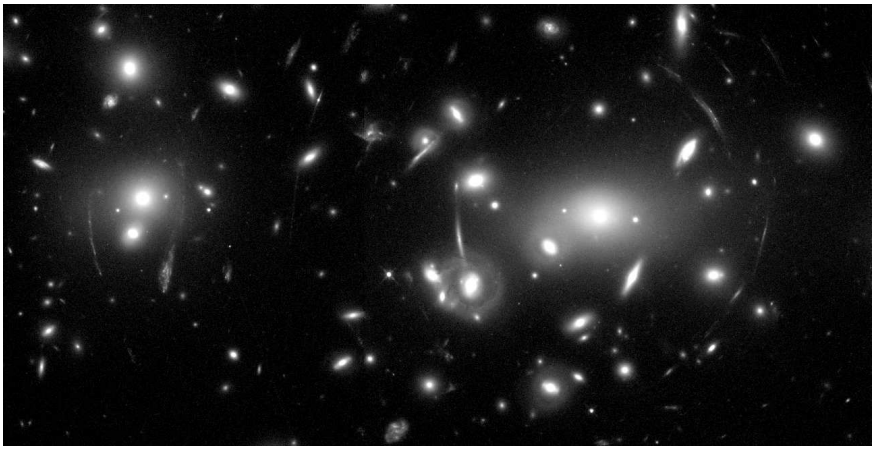


Fig. 2.30. Hubble Space Telescope image of the cluster Abell 2218. The arcs and arclets around the center of the cluster are images of background galaxies that are strongly distorted due to gravitational lensing. [Courtesy of W. Couch, R. Ellis, NASA, and Space Telescope Science Institute]

$\sim 10^{45}\,\mathrm{ergs}^{-1}$. This X-ray emission is spatially extended, with detected sizes of $\sim 1\,\mathrm{Mpc}$, and so it cannot originate from the individual member galaxies. Rather, the spectral energy distribution of the X-ray emission suggests that the emission mechanism is thermal bremsstrahlung (see §**??**) from a hot plasma. The inferred temperatures of this intracluster medium (ICM) are in the range $10^7 - 10^8\,\mathrm{K}$, corresponding to a typical photon energy of $1 - 10\,\mathrm{keV}$, so that the gas is expected to be fully ionized.

For a fully ionized gas, the thermal bremsstrahlung emissivity, i.e. the emission power per unit frequency per unit volume, is related to its density and temperature roughly as

$$\varepsilon_{\mathrm{ff}}(\nu) \propto n^2 T^{-1/2} \exp\left(-\frac{h_{\mathrm{P}}\nu}{k_{\mathrm{B}}T}\right). \tag{2.40}$$

The quantity we observe from a cluster is the X-ray surface brightness, which is the integration of the emissivity along the line of sight:†

$$S_\nu(x,y) \propto \int \varepsilon_{\mathrm{ff}}(\nu;x,y,z)\,\mathrm{d}z\,. \tag{2.41}$$

If S_ν is measured as a function of ν (i.e. photon energy), the temperature at a given projected position (x,y) can be estimated from the shape of the spectrum. Note that this temperature is an emissivity-weighted mean along the line of sight, if the temperature varies with z. Once the temperature is known, the amplitude of the surface brightness can be used to estimate $\int n^2\,\mathrm{d}z$ which, together with a density model, can be used to obtain the gas density distribution. Thus, X-ray observations of clusters can be used to estimate the corresponding masses in hot gas. These are found to fall in the range $(10^{13} - 10^{14})h^{-5/2}\,\mathrm{M}_\odot$, about ten times as large as the total stellar mass in member galaxies. Furthermore, as we will see in §**??**, if the X-ray gas is in hydrostatic equilibrium with the cluster potential, so that the local pressure gradient is balanced by the gravitational force, the observed temperature and density distribution of the gas can also be used to estimate the *total* mass of the cluster.

Another method to measure the total mass of a cluster of galaxies is through gravitational lensing. According to General Relativity, the light from a background source is deflected when

† Here we ignore redshifting and surface brightness dimming due to the expansion of the Universe; see §**??**.

it passes a mass concentration in the foreground, an effect called gravitational lensing. As discussed in more detail in §**??**, gravitational lensing can have a number of effects: it can create multiple images on the sky of the same background source, it can magnify the flux of the source, and it can distort the shape of the background source. In particular, the image of a circular source is distorted into an ellipse if the source is not close to the line-of-sight to the lens so that the lensing effect is weak (weak lensing). Otherwise, if the source is close to the line-of-sight to the lens, the image is stretched into an arc or an arclet (strong lensing).

Both strong and weak lensing can be used to estimate the total gravitational mass of a cluster. In the case of strong lensing, one uses giant arcs and arclets, which are the images of background galaxies lensed by the gravitational field of the cluster (see Fig. 2.30). The location of an arc in a cluster provides a simple way to estimate the projected mass of the cluster within the circle traced by the arc. Such analyses have been carried out for a number of clusters, and the total masses thus obtained are in general consistent with those based on the internal kinematics, the X-ray emission, or weak lensing. Typically the total cluster masses are found to be an order of magnitude larger than the combined masses of stars and hot gas, indicating that clusters are dominated by dark matter, as first pointed out by Fritz Zwicky in the 1930s.

2.5.2 Groups of Galaxies

By definition, groups are systems of galaxies with richness less than that of clusters, although the dividing line between groups and clusters is quite arbitrary. Groups are selected by applying certain richness and compactness criteria to galaxy surveys, similar to what Abell used for selecting clusters. Typically, groups selected from redshift surveys include systems with at least 3 galaxies and with a number density enhancement of the order of 20 (e.g. Geller & Huchra, 1983; Nolthenius & White, 1987; Eke et al., 2004; Yang et al., 2005a; Berlind et al., 2006; Yang et al., 2007). Groups so selected typically contain 3–30 L^* galaxies, have a total B-band luminosity in the range $10^{10.5}$–$10^{12}h^{-2}\,\mathrm{L}_\odot$, have radii in the range $(0.1-1)\,h^{-1}$Mpc, and have typical (line of sight) velocity dispersion of the order of $300\,\mathrm{km\,s^{-1}}$. As for clusters, the total dynamical mass of a group can be estimated from its size and velocity dispersion using the virial theorem (2.38), and masses thus obtained roughly cover the range $10^{12.5}-10^{14}h^{-1}\,\mathrm{M}_\odot$. Therefore, the typical mass-to-light ratio of galaxy groups is $(M/L_B)\sim 100h(\mathrm{M}_\odot/\mathrm{L}_\odot)_B$, significantly lower than that for clusters.

(a) Compact Groups A special class of groups are the so-called compact groups. Each of these systems consists of only a few galaxies but with an extremely high density enhancement. A catalogue of about 100 compact groups was constructed by Hickson (1982) from an analysis of photographic plates. These Hickson Compact Groups (HCGs) typically consist of only 4 or 5 galaxies and have a projected radius of only 50–100 kpc. A large fraction ($\sim 40\%$) of the galaxies in HCGs show evidence for interactions, and based on dynamical arguments, it is expected that the HCGs are each in the process of merging to perhaps form a single bright galaxy.

(b) The Local Group The galaxy group that has been studied in most detail is the Local Group, of which the Milky Way and M31 are the two largest members. The Local Group is a loose association of galaxies which fills an irregular region just over 1 Mpc across. Because we are in it, we can probe the members of the Local Group down to much fainter magnitudes than is possible in any other group. Table 2.7 lists the 30 brightest members of the Local Group, while Fig. 2.31 shows their spatial distribution. Except for a few of the more distant objects, the majority of the Local Group members can be assigned as satellites of either the Milky Way or M31. The largest satellite of the Milky Way is the Large Magellanic Cloud (LMC). Its luminosity is about one tenth of that of its host and it is currently actively forming stars. Together with its smaller companion, the Small Magellanic Cloud (SMC), it follows a high angular momentum

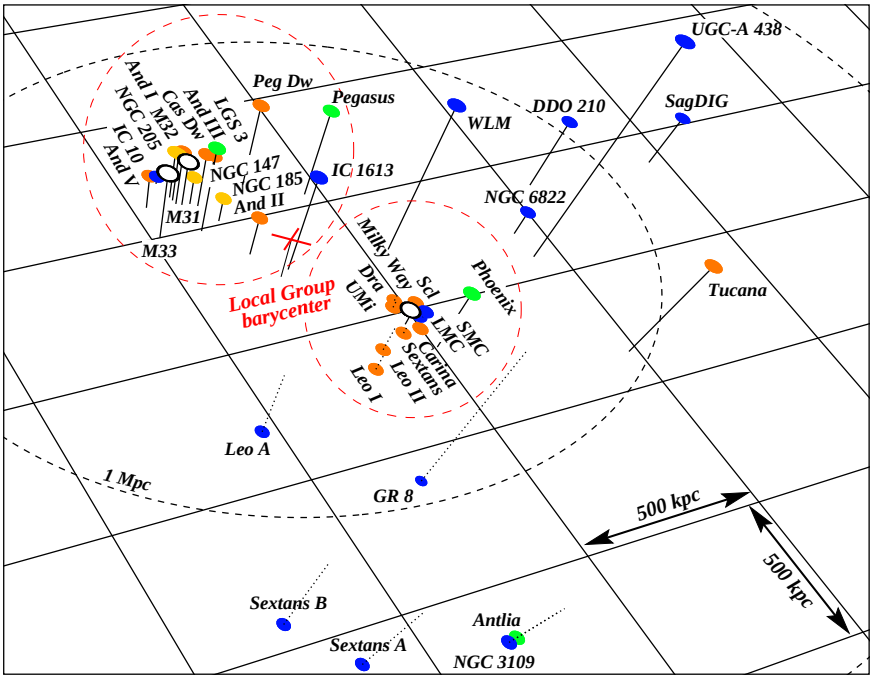


Fig. 2.31. Schematic distribution of galaxies in the local group. [Courtesy of E. Grebel, see Grebel (1999)]

orbit almost perpendicular to the Milky Way's disk and currently lies about 50 kpc from the Galactic center. Both Magellanic Clouds have metallicities significantly lower than that of the Milky Way. All the other satellites of our Galaxy are low mass, gas-free and metal-poor dwarf spheroidals. The most massive of these are the Fornax and Sagittarius systems. The latter lies only about 20 kpc from the Galactic center and is in the process of being disrupted by the tidal effects of its host. Several of the dwarf spheroidals contain stellar populations with a range of ages, some being ten times younger than typical Population II stars.

The Andromeda nebula itself is similar to but more massive than the Milky Way, with a more prominent bulge population and somewhat less active current star formation. Its largest satellite is the bulge-less dwarf spiral M33, which is only slightly brighter than the LMC and is actively forming stars. M31 also has two close dwarf elliptical companions, M32 and NGC 205, and two similar satellites, NGC 147 and NGC 185, at somewhat larger distances. These galaxies are denser and more luminous than dwarf spheroidals, but are also devoid of gas and young stars (NGC 205 actually has a small star-forming region in its nucleus). Finally M31 has its own retinue of dwarf spheroidal satellites.

The more distant members of the Local group are primarily dwarf irregular galaxies with active star formation, similar to but less luminous than the Magellanic Clouds. Throughout the Local Group there is a very marked tendency for galaxies with a smaller stellar mass to have a lower metallicity, with the smallest dwarfs having metallicities about one tenth of the solar value (Mateo, 1998).

Table 2.7. *Local Group members*

Name	Type	$\mathscr{M}_V$	l, b	Distance (kpc)
Milky Way (Galaxy)	Sbc	−20.6	0, 0	8
LMC	Irr	−18.1	280,−33	49
SMC	Irr	−16.2	303,−44	58
Sagittarius	dSph/E7	−14.0	6,−14	24
Fornax	dSph/E3	−13.0	237,−65	131
Leo I (DDO 74)	dSph/E3	−12.0	226,49	270
Sculptor	dSph/E3	−10.7	286,−84	78
Leo II (DDO 93)	dSph/E0	−10.2	220,67	230
Sextans	dSph/E4	−10.0	243,42	90
Carina	dSph/E4	−9.2	260,−22	87
Ursa Minor (DDO 199)	dSph/E5	−8.9	105,45	69
Draco (DDO 208)	dSph/E3	−8.6	86,35	76
M 31 (NGC 224)	Sb	−21.1	121,−22	725
M 33 (NGC 598)	Sc	−18.9	134,−31	795
IC 10	Irr	−17.6	119,−03	1250
NGC 6822 (DDO 209)	Irr	−16.4	25,−18	540
M 32 (NGC 221)	dE2	−16.4	121,−22	725
NGC 205	dE5	−16.3	121,−21	725
NGC 185	dE3	−15.3	121,−14	620
IC 1613 (DDO 8)	Irr	−14.9	130,−60	765
NGC 147 (DDO 3)	dE4	−14.8	120,−14	589
WLM (DDO 221)	Irr	−14.0	76, −74	940
Pegasus (DDO 216)	Irr	−12.7	94, −43	759
Leo A	Irr	−11.7	196, 52	692
And I	dSph/E0	−11.7	122,−25	790
And II	dSph/E3	−11.7	129,−29	587
And III	dSph/E6	−10.2	119,−26	790
Phoenix	Irr	−9.9	272,−68	390
LGC 3	Irr	−9.7	126,−41	760
Tucana	dSph/E5	−9.6	323,−48	900

2.6 Galaxies at High Redshifts

Since galaxies at higher redshifts are younger, a comparison of the (statistical) properties of galaxies at different redshifts provides a direct window on their formation and evolution. However, a galaxy of given luminosity and size is both fainter and of lower surface brightness when located at higher redshifts (see §**??**). Thus, if high-redshift galaxies have similar luminosities and sizes as present-day galaxies, they would be extremely faint and of very low surface brightness, making them very difficult to detect. Indeed, until the mid 1990s, the known high-redshift galaxies with $z \gtrsim 1$ were almost exclusively active galaxies, such as quasars, QSOs and radio galaxies, simply because these were the only galaxies sufficiently bright to be observable with the facilities available then. Thanks to a number of technological advancements in both telescopes and detectors, we have made enormous progress, and today the galaxy population can be probed out to $z \gtrsim 6$.

The search for high-redshift galaxies usually starts with a photometric survey of galaxies in multiple photometric bands down to very faint magnitude limits. Ideally, one would like to have redshifts for all these galaxies and study the entire galaxy population at all different redshifts. In reality, however, it is extremely time-consuming to obtain spectra of faint galaxies even with the 10-meter class telescopes available today. In order to make progress, different techniques have

been used, which basically fall in three categories: (i) forsake the use of spectra and only use photometry either to analyze the number counts of galaxies down to very faint magnitudes or to derive photometric redshifts, (ii) use broad-band color selection to identify target galaxies likely to be at high redshift for follow-up spectroscopy, and (iii) use narrow-band photometry to find objects with a strong emission line in a narrow redshift range. Here we give a brief overview of these different techniques.

2.6.1 Galaxy Counts

In the absence of redshifts, some information about the evolution of the galaxy population can be obtained from galaxy counts, $\mathscr{N}(m)$, defined as the number of galaxies per unit apparent magnitude (in a given waveband) per unit solid angle:

$$\mathrm{d}^2N(m) = \mathscr{N}(m)\,\mathrm{d}m\,\mathrm{d}\omega. \tag{2.42}$$

Although the measurement of $\mathscr{N}(m)$ is relatively straightforward from any galaxy catalogue with uniform photometry, interpreting the counts in terms of galaxy number density as a function of redshift is far from trivial. First of all, the waveband in which the apparent magnitudes are measured corresponds to different rest-frame wavebands at different redshifts. To be able to test for evolution in the galaxy population with redshift, this shift in waveband needs to be corrected for. But such correction is not trivial to make, and can lead to large uncertainties (see §**??**). Furthermore, both cosmology and evolution can affect $\mathscr{N}(m)$. In order to break this degeneracy, and to properly test for evolution, accurate constraints on cosmological parameters are required.

Despite these difficulties, detailed analyses of galaxy counts have resulted in a clear detection of evolution in the galaxy population. Fig. 2.32 shows the galaxy counts in four wavebands obtained from a variety of surveys. The solid dots are obtained from the Hubble Deep Fields (Ferguson et al., 2000) imaged to very faint magnitudes with the HST. The solid lines in Fig. 2.32 show the predictions for a realistic cosmology in which it is assumed that the galaxy population does not evolve with redshift. A comparison with the observed counts shows that this model severely underpredicts the galaxy counts of faint galaxies, especially in the bluer wavebands. The nature of this excess of faint blue galaxies will be discussed in §**??**.

2.6.2 Photometric Redshifts

Since spectroscopy relies on dispersing the light from an object according to wavelength, accurate redshifts, which require sufficient signal-to-noise in individual emission and/or absorption lines, can only be obtained for relatively bright objects. An alternative, although less reliable, technique to measure redshifts relies on broad band photometry. By measuring the flux of an object in a relatively small number of wavebands, one obtains a very crude sampling of the object's SED. As we have seen, the SEDs of galaxies reveal a number of broad spectral features (see Fig. 2.12). An important example is the 4000 Å break, which is due to a sudden change in the opacity at this wavelength in the atmospheres of low mass stars, and therefore features predominantly in galaxies with stellar population ages $\gtrsim 10^8$ yr. Because of this 4000 Å break and other broad spectral features, the colors of a population of galaxies at a given redshift only occupy a relatively small region of the full multi-dimensional color space. Since this region changes as function of redshift, the broad-band colors of a galaxy can be used to estimate its redshift.

In practice one proceeds as follows. For a given template spectrum, either from an observed galaxy or computed using population synthesis models, one can determine the relative fluxes expected in different wavebands for a given redshift. By comparing these expected fluxes with the observed fluxes one can determine the best-fit redshift and the best-fit template spectrum (which

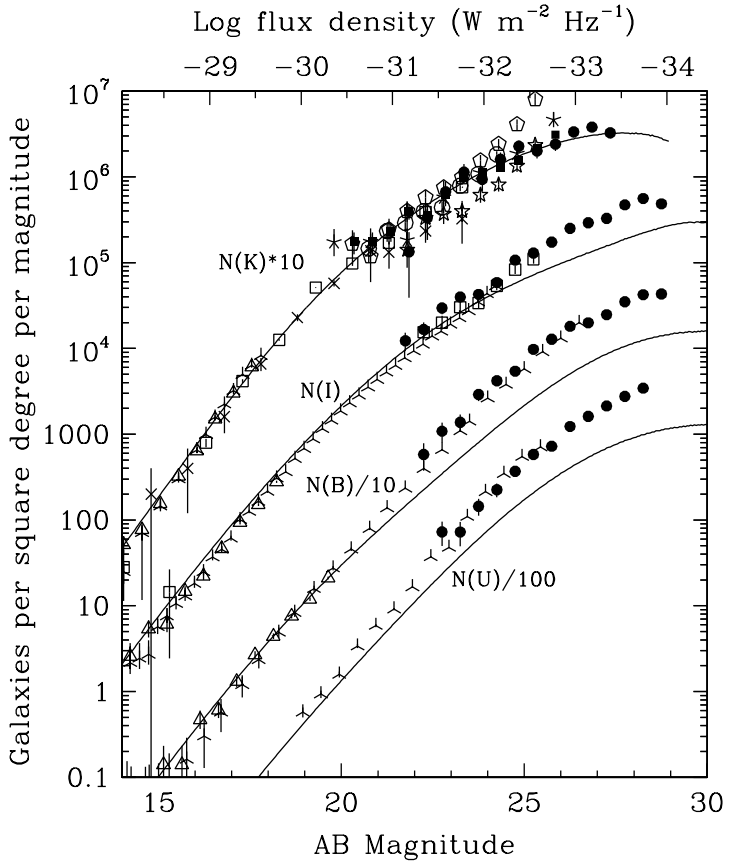


Fig. 2.32. Galaxy counts in the *U*, *B*, *I* and *K* bands obtained from the Hubble deep fields (solid symbols) and a number of other ground-based surveys (other symbols). The solid lines show the predictions for a realistic cosmology in which it is assumed that the galaxy population does not evolve with redshift. [Adapted from Ferguson et al. (2000) by permission of ARAA]

basically reflects the spectral type of the galaxy). The great advantage of this method is that photometric redshifts can be measured much faster than their spectroscopic counterparts, and that it can be extended to much fainter magnitudes. The obvious downside is that photometric redshifts are far less reliable. While a spectroscopic redshift can easily be measured to a relative error of less than 0.1 percent, photometric errors are typically of the order of 3 to 10 percent, depending on which and how many wavebands are used. Furthermore, the error is strongly correlated with the spectral type of the galaxy. It is typically much larger for star forming galaxies, which lack a pronounced 4000 Å break, than for galaxies with an old stellar population.

A prime example of a photometric redshift survey, illustrating the strength of this technique, is the COMBO-17 survey (Wolf et al., 2003), which comprises a sample of $\sim 25,000$ galaxies with photometric redshifts obtained from photometry in 17 relatively narrow optical wavebands. Because of the use of a relatively large number of filters, this survey was able to reach an average redshift accuracy of ~ 3 percent, sufficient to study various statistical properties of the galaxy population as a function of redshift.

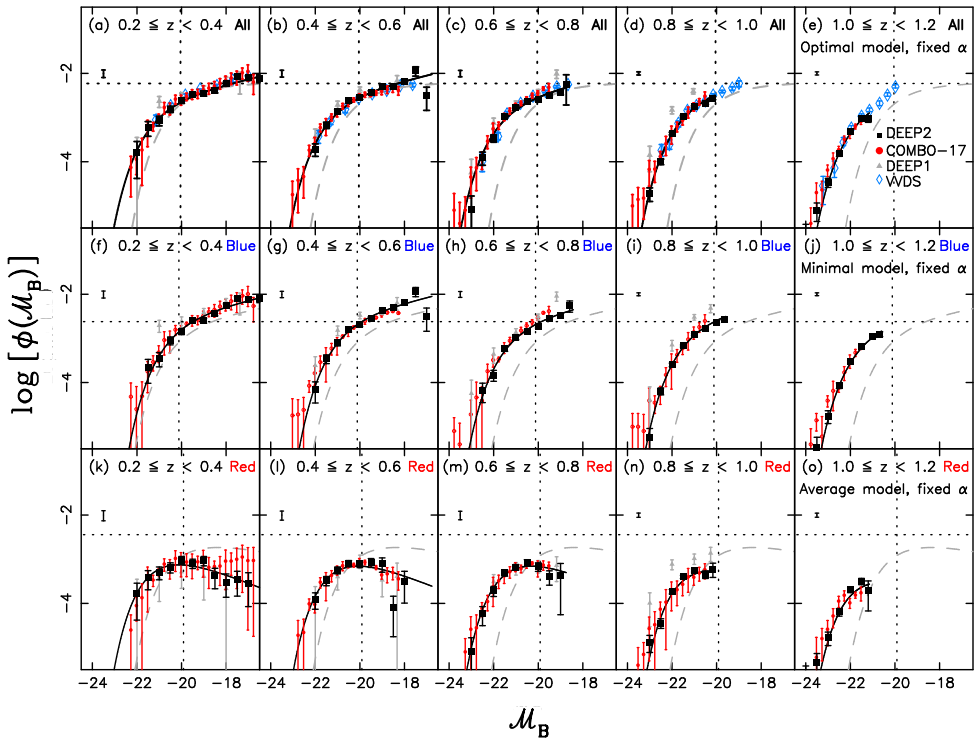


Fig. 2.33. Luminosity functions measured in different redshift bins for 'All' galaxies (top row), 'Blue' galaxies (middle row), and 'Red' galaxies (bottom row). Different symbols correspond to results obtained from different redshift surveys (DEEP1, DEEP2, COMBO-17 and VVDS, as indicated). The solid black lines indicate Schechter functions fitted to the DEEP2 results. For comparison, the dashed grey lines show the Schechter functions for local samples obtained from the SDSS. Overall the agreement between the different surveys is very good. [Adapted from Faber et al. (2007) by permission of AAS]

2.6.3 Galaxy Redshift Surveys at $z \sim 1$

In order to investigate the nature of the excess of faint-blue galaxies detected with galaxy counts, a number of redshift surveys out to $z \sim 1$ were carried out in the mid 1990s using 4m class telescopes, including the Canada-France Redshift Survey (CFRS; Lilly et al., 1995) and the Autofib-LDSS survey (Ellis et al., 1996). These surveys, containing the order of 1000 galaxies, allowed a determination of galaxy luminosity functions (LFs) covering the entire redshift range $0 < z \lesssim 1$. The results, although limited by small number statistics, confirmed that the galaxy population is evolving with redshift, in agreement with the results obtained from the galaxy counts.

With the completion of a new class of 10-meter telescopes, such as the KECK and the VLT, it became possible to construct much larger redshift samples at intermediate to high redshifts. Currently the largest redshift survey at $z \sim 1$ is the DEEP2 Redshift Survey (Davis et al., 2003), which contains about 50,000 galaxies brighter than $R_{\rm AB} \approx 24.1$ in a total of ~ 3 square degrees in the sky. The adopted color criteria ensure that the bulk of the galaxies selected for spectroscopy have redshifts in the range $0.7 \lesssim z \lesssim 1.4$. Results from DEEP2 show, among others, that the color bimodality observed in the local Universe (see §2.4.3) is already present at $z \sim 1$ (Bell et al., 2004; Willmer et al., 2006; Cooper et al., 2007). Together with COMBO-17, the DEEP2 survey has provided accurate measurements of the galaxy luminosity function, split according to color, out to $z \sim 1.2$. As shown in Fig. 2.33, the different surveys yield results in excellent

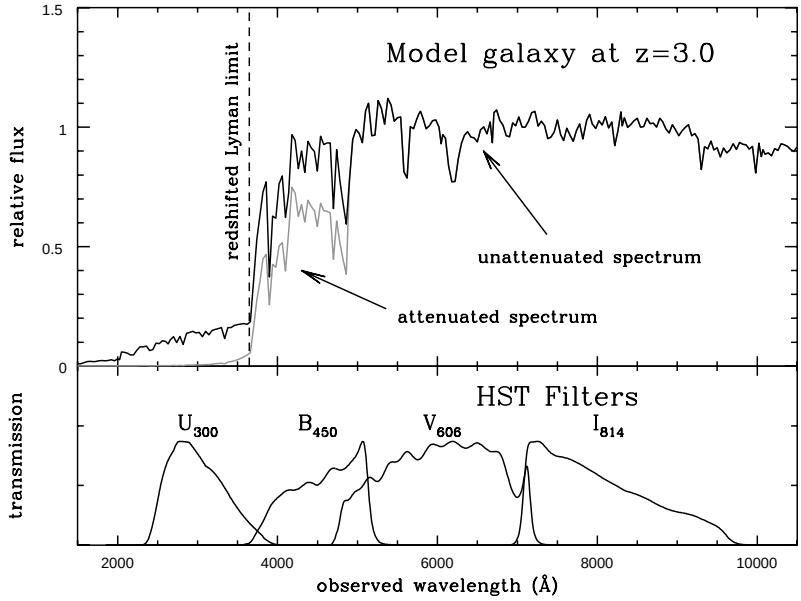


Fig. 2.34. An illustration of how the 'Lyman-break' or 'drop-out' technique can be used to select star-forming galaxies at redshifts $z \sim 3$. The spectrum of a typical star-forming galaxy has a break at the Lyman limit (912 Å), which is redshifted to a wavelength $\lambda \sim 4000$ Å if the galaxy is at $z \sim 3$. As a result, the galaxy appears very faint (or may even be undetectable) in the U band, but bright in the redder bands. [Courtesy of M. Dickinson, see Dickinson (1998)]

mutual agreement. In particular, they show that the characteristic luminosity, L^*, of the galaxy population in the rest-frame B-band becomes fainter by ~ 1.3 mag from $z = 1$ to $z = 0$ for both the red and blue populations. However, the number density of L^* galaxies, ϕ^*, behaves very differently for red and blue galaxies: while ϕ^* of blue galaxies has roughly remained constant since $z = 1$, that of red galaxies has nearly quadrupled (Bell et al., 2004; Brown et al., 2007; Faber et al., 2007). As we will see §**??**, this puts important constraints on the formation history of elliptical galaxies.

Another large redshift survey, which is being conducted at the time of writing, is the VIRMOS VLT Deep Survey (VVDS; Le Fèvre et al., 2005) which will ultimately acquire $\sim 150,000$ redshifts over ~ 4 square degrees in the sky. Contrary to DEEP2, the VVDS does not apply any color selection; rather, spectroscopic candidates are purely selected on the basis of their apparent magnitude in the $I_{\rm AB}$ band. Consequently the redshift distribution of VVDS galaxies is very broad: it peaks at $z \sim 0.7$, but has a long high-redshift tail extending all the way out to $z \sim 5$. The luminosity functions obtained from ~ 8000 galaxies in the first data of the VVDS are in excellent agreement with those obtained from DEEP2 and COMBO-17 (see Fig. 2.33).

2.6.4 Lyman-Break Galaxies

As discussed above, broad features in the SEDs of galaxies allow for the determination of photometric redshifts, and for a very successful pre-selection of candidate galaxies at $z \sim 1$ for follow-up spectroscopy. The same principle can also be used to select a special subset of galaxies at much higher redshifts. A star-forming galaxy has a SED roughly flat down to the Lyman limit at $\lambda \sim 912$ Å, beyond which there is a prominent break due to the spectra of the stellar population (see the spectra of the O9 and B0 stars in Fig. 2.5) and to intervening absorption. Physically

this reflects the large ionization cross section of neutral hydrogen. A galaxy revealing a pronounced break at the Lyman limit is called a Lyman-break galaxy (LBG), and is characterized by a relatively high star formation rate.

For a LBG at $z \sim 3$, the Lyman break falls in between the U and B bands (see Fig. 2.34). Therefore, by selecting those galaxies in a deep multi-color survey that are undetected (or extremely faint) in the U-band, but detected in the B and redder bands, one can select candidate star-forming galaxies in the redshift range $z = 2.5$-3.5 (Steidel et al., 1996). Galaxies selected this way are called UV drop-outs. Follow-up spectroscopy of large samples of UV drop-out candidates has confirmed that this Lyman-break technique is very effective, with the vast majority of the candidates being indeed star forming galaxies at $z \sim 3$.

To date more than 1000 LBGs with $2.5 \lesssim z \lesssim 3.5$ have been spectroscopically confirmed. The comoving number density of bright LBGs is estimated to be comparable to that of present-day bright galaxies. However, contrary to typical bright galaxies at $z \sim 0$, which are mainly early-type galaxies, LBGs are actively forming stars (note that they are effectively selected in the B-band, corresponding to rest frame UV at $z \sim 3$) with inferred star formation rates in the range of a few times $10\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$ up to $\sim 100\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$, depending on the uncertain amount of dust extinction (Adelberger & Steidel, 2000).

The Lyman break (or drop-out) technique has also been applied to deep imaging surveys in redder bands to select galaxies that drop out of the B-band, V-band and even the I-band. If these are indeed LBGs, their redshifts correspond to $z \sim 4$, $z \sim 5$, and $z \sim 6$, respectively. Deep imaging surveys with the HST and ground-based telescopes have already produced large samples of these drop-out galaxies. Unfortunately, most of these galaxies are too faint to follow-up spectroscopically, so that it is unclear to what extent these samples are contaminated by low redshift objects. With this caveat in mind, the data have been used to probe the evolution of the galaxy luminosity function (LF) in the rest-frame UV all the way from $z \sim 0$ (using data from the GALEX satellite) to $z \sim 6$. Over the redshift range $4 \lesssim z \lesssim 6$ this LF is found to have an extremely steep faint-end slope, while the characteristic luminosity $L^*_{\rm UV}$ is found to brighten significantly from $z = 6$ to $z = 4$ (Bouwens et al., 2007).

2.6.5 Lyα Emitters

In addition to the broad-band selection techniques mentioned above, one can also search for high-redshift galaxies using narrow-band photometry. This technique has been used extensively to search for Lyα emitters (LAEs) at redshifts $z \gtrsim 3$ for which the Lyα emission line ($\lambda = 1216\,\text{Å}$) appears in the optical.

Objects with strong Lyα are either QSOs or galaxies actively forming stars. However, since the Lyα flux is easily quenched by dust extinction, not all star forming galaxies feature Lyα emission. In fact, a large fraction of LBGs, although actively forming stars, lack an obvious Lyα emission line. Therefore, by selecting LAEs one is biased towards star forming galaxies with relatively little dust, or in which the dust has a special geometry so that part of the Lyα flux can leave the galaxy un-extincted.

One can search for LAEs at a particular redshift, $z_{\rm LAE}$, using a narrow-band filter centered on a wavelength $\lambda = 1216\,\text{Å} \times (1 + z_{\rm LAE})$ plus another, much broader filter centered on the same λ. The objects in question then show up as being particularly bright in the narrow-band filter in comparison to the broad band image. A potential problem is that one might also select emission-line galaxies at very different redshifts. For example, a galaxy with strong [OII] emission ($\lambda = 3727\,\text{Å}$) would shift into the same narrow band filter if the galaxy is at a redshift $z_{\rm [OII]} = 0.33 z_{\rm LAE} - 0.67$. To minimize this kind of contamination one generally only selects sys-

tems with a large equivalent width† in the emission line ($\gtrsim 150\,$Å), which excludes all but the rarest [OII] emitters. Another method to check whether the object is indeed a LAE at $z_{\rm LAE}$ is to use follow-up spectroscopy to see whether (i) there are any other emission lines visible that help to determine the redshift, and (ii) the emission line is asymmetric, as expected for Lyα due to preferential absorption in the blue wing of the line.

This technique can be used to search for high redshift galaxies in several narrow redshift bins ranging from $z \sim 3$ to $z \sim 6.5$, and at the time of writing ~ 100 LAEs covering this redshift range have been spectroscopically confirmed. Since these systems are typically extremely faint, the nature of these objects is still unclear.

2.6.6 Sub-Millimeter Sources

Since the Lyman-break technique and Lyα imaging select galaxies according to their rest-frame UV light, they may miss dust-enshrouded star-forming galaxies, the high-redshift counterparts of local starbursts. Most of the UV photons from young stars in such galaxies are absorbed by dust and re-emitted in the far-infrared. Such galaxies can therefore be detected in the sub-millimeter (sub-mm) band, which corresponds to rest-frame far-infrared at $z \sim 3$. Deep surveys in the sub-mm bands only became possible in the mid 1990s with the commissioning of the Sub-millimeter Common-User Bolometer Array (SCUBA, see Holland et al., 1999), operating at 450 μm and 850 μm, on the James Clerk Maxwell Telescope (JCMT). This led to the discovery of an unexpectedly large population of faint sub-mm sources (Smail et al., 1997). An extensive and difficult observational campaign to identify the optical counterparts and measure their redshifts has shown that the majority of these sources are indeed starburst galaxies at a median redshift of $z \sim 2.5$. Some of the strong sub-mm sources with measured redshifts have inferred star formation rates as high as several $100\,{\rm M}_\odot\,{\rm yr}^{-1}$, similar to those of ULIRGS at $z \simeq 0$. Given the large number density of SCUBA sources, and their inferred star formation rates, the total amount of stars formed in these systems may well be larger than that formed in the Lyman-break galaxies at the same redshift (Blain et al., 1999).

2.6.7 Extremely Red Objects and Distant Red Galaxies

Another important step forward in the exploration of the galaxy population at high redshift came with the development of large format near-infrared (NIR) detectors. Deep, wide-field surveys in the *K*-band lead to the discovery of a class of faint galaxies with extremely red optical-to-NIR colors ($R-K>5$). Follow-up spectroscopy has shown that these Extremely Red Objects (EROs) typically have redshifts in the range $0.7 \lesssim z \lesssim 1.5$. There are two possible explanations for their red colors: either they are galaxies dominated by old stellar populations with a pronounced 4000 Å break that has been shifted red-wards of the *R*-band filter, or they are starbursts (or AGN) strongly reddened due to dust extinction. Spectroscopy of a sample of ~ 50 EROs suggests that they are a roughly equal mix of both (Cimatti et al., 2002).

Deep imaging in the NIR can also be used to search for the equivalent of 'normal' galaxies at $z \gtrsim 2$. As described above, the selections of LBGs, LAEs and sub-mm sources are strongly biased towards systems with relatively high star formation rates. Consequently, the population of high redshift galaxies picked out by these selections is very different from the typical, present-day galaxies whose light is dominated by evolved stars. In order to select high-redshift galaxies in a way similar to how 'normal' galaxies are selected at low redshift, one has to go to the rest-frame optical, which corresponds to the NIR at $z \sim 2$ - 3. Using the InfraRed ExtraGalactic Survey (FIRES, Labbé et al., 2003), Franx et al. (2003) identified a population of galaxies on the

† The equivalent width of an emission line, a measure for its strength, is defined as the width of the wavelength range over which the continuum needs to be integrated to have the same flux as measured in the line (see §**??**).

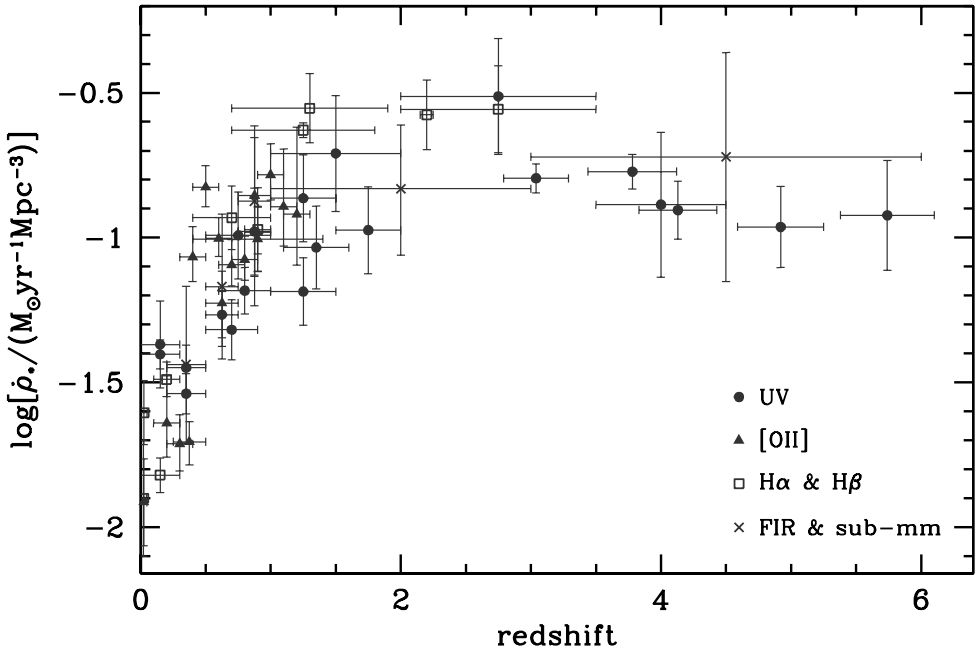


Fig. 2.35. The global star formation rate (in $M_\odot\,yr^{-1}\,Mpc^{-3}$) as a function of redshift. Different symbols correspond to different rest-frame wavelength ranges used to infer the star formation rates, as indicated. [Based on the data compilation of Hopkins (2004)]

basis of their red NIR color, $J_s - K_s > 2.3$, where the K_s and J_s filters are similar to the classical *J* and *K* filters, but centered on somewhat shorter wavelengths. The galaxies so selected are now referred to as Distant Red Galaxies (DRGs). The color criterion efficiently isolates galaxies with prominent Balmer- or 4000Å breaks at $z \gtrsim 2$, and can therefore be used to select galaxies with the oldest stellar populations at these redshifts. However, the NIR color criterion alone also selects galaxies with significant current star formation, even dusty starbursts. The brightest DRGs ($K_s < 20$) are among the most massive galaxies at $z \gtrsim 2$, with stellar masses $\gtrsim 10^{11}\,M_\odot$, likely representing the progenitors of present-day massive ellipticals. As EROs, DRGs are largely missed in UV-selected (e.g. LBG) samples. Yet, as shown by van Dokkum et al. (2006), among the most massive population of galaxies in the redshift range $2 \lesssim z \lesssim 3$, DRGs dominate over LBGs both in number density and in stellar mass density.

Using photometry in the *B*-, *z*-, and *K*-bands, Daddi et al. (2004) introduced a selection criterion which allows one to recover the bulk of the galaxy population in the redshift range $1.4 \lesssim z \lesssim 2.5$, including both active star-forming galaxies as well as passively evolving galaxies, and to distinguish between the two classes. In particular, the color criterion $BzK \equiv (z-K)_{\rm AB} - (B-z)_{\rm AB} > -0.2$ is very efficient in selecting star-forming galaxies with $1.4 \lesssim z \lesssim 2.5$, independently of their dust reddening, while the criteria $BzK < -0.2$ and $(z-K)_{\rm AB} > 2.5$ predominantly select passively evolving galaxies in the same redshift interval. At $z \sim 2$ the *BzK* - selected star-forming galaxies typically have higher reddening and higher star-formation rates than UV-selected galaxies. A comparison of *BzK* galaxies with DRGs in the same redshift range shows that many of the DRGs are reddened starbursts rather than passively evolving galaxies.

2.6.8 The Cosmic Star Formation History

The data on star-forming galaxies at different redshifts can in principle be used to map out the production rate of stars in the Universe as a function redshift. If we do not care where stars form, the star formation history of the Universe can be characterized by a global quantity, $\dot{\rho}_\star(z)$, which is the total gas mass that is turned into stars per unit time per unit volume at redshift z.

In order to estimate $\dot{\rho}_\star(z)$ from observation, one requires estimates of the number density of galaxies as a function of redshift and their (average) star formation rates. In practice, one observes the number density of galaxies as a function of luminosity in some waveband, and estimates $\dot{\rho}_\star(z)$ from

$$\dot{\rho}_\star(z) = \int \mathrm{d}\dot{M}_\star \dot{M}_\star \int P(\dot{M}_\star|L,z)\phi(L,z)\,\mathrm{d}L = \int \langle \dot{M}_\star \rangle (L,z)\,\phi(L,z)\,\mathrm{d}L\,, \tag{2.43}$$

where $P(\dot{M}_\star|L,z)\,\mathrm{d}\dot{M}_\star$ is the probability for a galaxy with luminosity L (in a given band) at redshift z to have a star formation rate in the range $(\dot{M}_\star, \dot{M}_\star + \dot{M}_\star)$, and $\langle \dot{M}_\star \rangle (L,z)$ is the mean star formation rate for galaxies with luminosity L at redshift z. The luminosity function $\phi(L,z)$ can be obtained from deep redshift surveys of galaxies, as summarized above. The transformation from luminosity to star formation rate depends on the rest-frame waveband used to measure the luminosity function, and typically involves many uncertainties (see § **??** for a detailed discussion).

Fig. 2.35 shows a compilation of various measurements of the global SFR at different redshifts, obtained using different techniques. Although there is still considerable scatter, and the data may be plagued by systematic errors due to uncertain extinction corrections, it is now well established that the cosmic star formation rate has dropped by roughly an order of magnitude from $z \sim 2$ to the present. Integrating this cosmic star formation history over time, one can show that the star-forming populations observed to date are already sufficient to account for the majority of stars observed at $z \sim 0$ (e.g. Dickinson et al., 2003).

2.7 Large-Scale Structure

An important property of the galaxy population is its overall spatial distribution. Since each galaxy is associated with a large amount of mass, one might naively expect that the galaxy distribution reflects the large-scale mass distribution in the Universe. On the other hand, if the process of galaxy formation is highly stochastic, or galaxies only form in special, preferred environments, the relation between the galaxy distribution and the matter distribution may be far from straightforward. Therefore, detailed studies of the spatial distribution of galaxies in principle can convey information regarding both the overall matter distribution, which is strongly cosmology dependent, and regarding the physics of galaxy formation.

Fig. 2.36 shows the distribution of more than 80,000 galaxies in the 2dFGRS, where the distances of the galaxies have been estimated from their redshifts. Clearly the distribution of galaxies in space is not random, but shows a variety of structures. As we have already seen in §2.5 some galaxies are located in high density clusters containing several hundreds of galaxies, or in smaller groups containing a few to tens of galaxies. The majority of all galaxies, however, are distributed in low-density filamentary or sheet-like structures. These sheets and filaments surround large voids, which are regions with diameters up to $\sim 100\,$Mpc that contain very few, or no, galaxies. One of the challenges in studying the spatial distribution of galaxies is to properly quantify the complexity of this 'cosmic web' of filaments, sheets and voids. In this section we consider the galaxy distribution as a point set in space and study the spatial correlations among these points in a statistical sense.

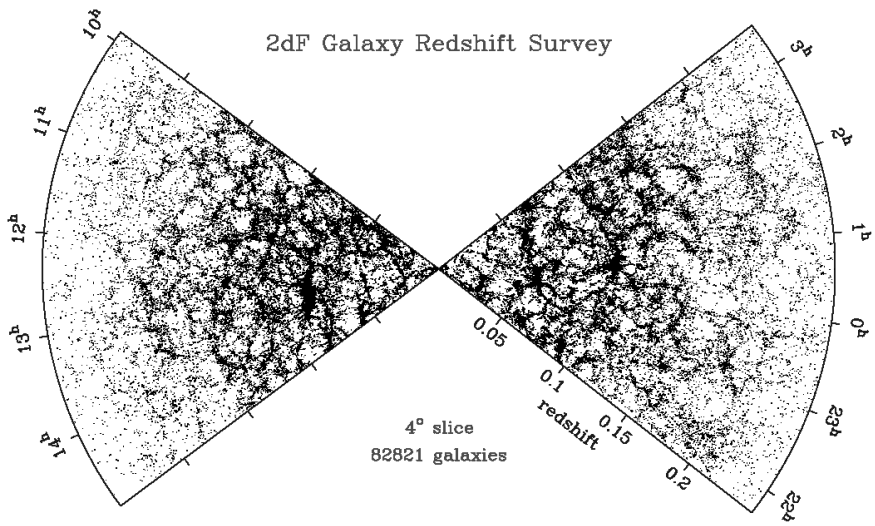


Fig. 2.36. The spatial distribution of $\sim 80,000$ galaxies in the 2dFGRS in a 4° slice projected onto the redshift/right-ascension plane. Clearly galaxies are not distributed randomly, but are clumped together in groups and clusters connected by large filaments that enclose regions largely devoid of galaxies. [Adapted from Peacock (2002)]

2.7.1 Two-Point Correlation Functions

One of the most important statistics used to characterize the spatial distribution of galaxies is the two-point correlation function, defined as the excess number of galaxy pairs of a given separation, r, relative to that expected for a random distribution:

$$\xi(r) = \frac{DD(r)\,\Delta r}{RR(r)\,\Delta r} - 1\,. \tag{2.44}$$

Here $DD(r)\,\Delta r$ is the number of galaxy pairs with separations in the range $r \pm \Delta r/2$, and $RR(r)\,\Delta r$ is the number that would be expected if galaxies were randomly distributed in space. Galaxies are said to be positively correlated on scale r if $\xi(r) > 0$, to be anti-correlated if $\xi(r) < 0$, and to be uncorrelated if $\xi(r) = 0$. Since it is relatively straightforward to measure, the two-point correlation function of galaxies has been estimated from various samples. In many cases, redshifts are used as distances and the corresponding correlation function is called the correlation function in redshift space. Because of peculiar velocities, this redshift-space correlation is different from that in real space. The latter can be estimated from the projected two-point correlation function, in which galaxy pairs are defined by their separations projected onto the plane perpendicular to the line of sight so that it is not affected by using redshift as distance (see Chapter **??** for details). Fig. 2.37 shows an example of the redshift-space correlation function and the corresponding real-space correlation function. On scales smaller than about $10\,h^{-1}$Mpc the real-space correlation function can well be described by a power law,†

$$\xi(r) = (r/r_0)^{-\gamma}\,, \tag{2.45}$$

with $\gamma \sim 1.8$ and with a correlation length $r_0 \approx 5\,h^{-1}$Mpc. This shows that galaxies are strongly clustered on scales $\lesssim 5\,h^{-1}$Mpc, and the clustering strength becomes weak on scales much larger

† Note that, because of the definition of the two-point correlation function, $\xi(r)$ has to become negative on large scales. Therefore, a power-law can only fit the data up to a finite scale.

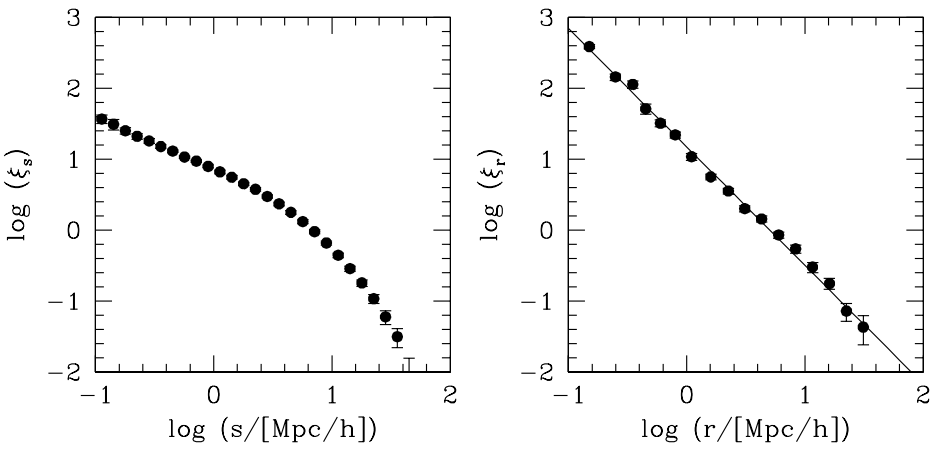


Fig. 2.37. The two-point correlation function of galaxies in redshift space (left) and real space (right). The straight line is a power law, $\xi(r) = (r/r_0)^{-\gamma}$, with $r_0 = 5.05\,h^{-1}$Mpc and $\gamma = 1.67$. [Based on data published in Hawkins et al. (2003)]

than $\sim 10\,h^{-1}$Mpc. The exact values of γ and r_0 are found to depend significantly on the properties of the galaxies. In particular the correlation length, r_0, defined by $\xi(r_0) = 1$, is found to depend on both galaxy luminosity and color in the sense that brighter and redder galaxies are more strongly clustered than their fainter and bluer counterparts (e.g. Norberg et al., 2001, 2002a; Zehavi et al., 2005; Wang et al., 2008).

One can apply exactly the same correlation function analysis to groups and clusters of galaxies. This shows that their two-point correlation functions has a logarithmic slope, γ, that is similar to that of galaxies, but a correlation length, r_0, which increases strongly with the richness of the systems in question, from about $5\,h^{-1}$Mpc for poor groups to about $20\,h^{-1}$Mpc for rich clusters (e.g. Yang et al., 2005b).

Another way to describe the clustering strength of a certain population of objects is to calculate the variance of the number counts within randomly-placed spheres of given radius r:

$$\sigma^2(r) \equiv \frac{1}{(\overline{n}V)^2} \sum_{i=i}^{M} (N_i - \overline{n}V)^2 \,, \tag{2.46}$$

where $\overline{n}$ is the mean number density of objects, $V = 4\pi r^3/3$, and N_i ($i = 1, \cdots, M$) are the number counts of objects in M randomly-placed spheres. For optically selected galaxies with a luminosity of the order of L^* one finds that $\sigma \sim 1$ on a scale of $r = 8\,h^{-1}$Mpc and decreases to $\sigma \sim 0.1$ on a scale of $r = 30\,h^{-1}$Mpc. This confirms that the galaxy distribution is strongly inhomogeneous on scales of $\lesssim 8\,h^{-1}$Mpc, but starts to approach homogeneity on significantly larger scales.

Since galaxies, groups and clusters all contain large amounts of matter, we expect their spatial distribution to be related to the mass distribution in the Universe to some degree. However, the fact that different objects have different clustering strengths makes one wonder if any of them are actually fair tracers of the matter distribution. The spatial distribution of luminous objects, such as galaxies, groups and clusters, depends not only on the matter distribution in the Universe, but also on how they form in the matter density field. Therefore, without a detailed understanding of galaxy formation, it is unclear which, if any, population of galaxies accurately traces the matter distribution. It is therefore very important to have independent means to probe the matter density field.

One such probe is the velocity field of galaxies. The peculiar velocities of galaxies are gen-

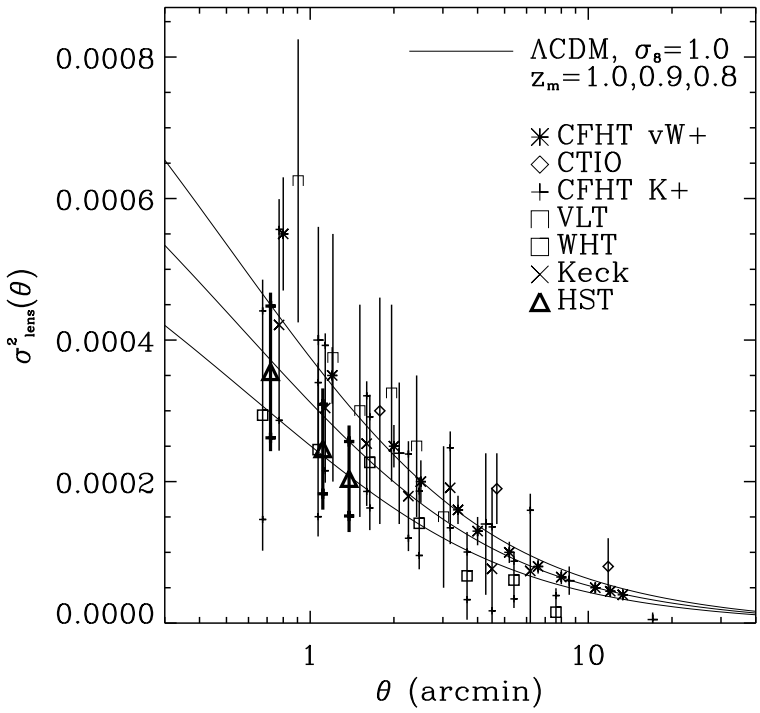


Fig. 2.38. In the limit of weak lensing, the shear field at a position in the sky is proportional to the ellipticity of the image of a circular source at that position. This plot shows the mean square of the shear field averaged within circular regions of given radius, θ, obtained from various observations. The non-zero values of this 'cosmic shear' are due to gravitational lensing induced by the line-of-sight projected mass distribution in the Universe. The solid curves are theoretical predictions (see §**??**) and are in good agreement with the data. [Adapted from Refregier et al. (2002) by permission of AAS]

erated by the gravitational field, and therefore contain useful information regarding the matter distribution in the Universe. In the past, two different methods have been used to extract this information from observations. One is to estimate the peculiar velocities of many galaxies by measuring both their receding velocities (i.e. redshifts) and their distances. The peculiar velocities then follow from Eq. (2.19), which can then be used to trace out the matter distribution. Such analyses not only yield constraints on the mean matter density in the Universe, but also on how galaxies trace the mass distribution. Unfortunately, although galaxy redshifts are easy to measure, accurate distance measurements for a large sample of galaxies are very difficult to obtain, severely impeding the applicability of this method. Another method, which is more statistical in nature, extracts information about the peculiar velocities of galaxies from a comparison of the real-space and redshift-space two-point correlation functions. This method is based on the fact that an isotropic distribution in real space will appear anisotropic in redshift space due to the presence of peculiar velocities. Such redshift-space distortions are the primary reason why the redshift-space correlation function has a shape different from that of the real-space correlation function (see Fig. 2.37). As described in detail in §**??**, by carefully modeling the redshift space distortions one can obtain useful constraints on the matter distribution in the Universe.

2.7.2 Probing the Matter Field via Weak Lensing

A very promising way to probe the mass distribution in the Universe is through weak gravitational lensing. Any light beam we observe from a distant source has been deflected and distorted due to the gravitational tidal field along the line of sight. This cumulative gravitational lensing effect due to the inhomogeneous mass distribution between source and observer is called cosmic shear, and holds useful information about the statistical properties of the matter field. The great advantage of this technique over the clustering analysis discussed above is that it does not have to make assumptions about the relation between galaxies and matter.

Unless the beam passes very close to a particular overdensity (i.e., a galaxy or cluster), in which case we are in the strong lensing regime, these distortions are extremely weak. Typical values for the expected shear are of the order of one percent on angular scales of a few arcminutes, which means that the distorted image of an intrinsically circular source has an ellipticity of 0.01. Even if one could accurately measure such a small ellipticity, the observed ellipticity holds no information without prior knowledge of the intrinsic ellipticity of the source, which is generally unknown. Rather, one detects cosmic shear via the spatial correlations of image ellipticities. The light beams from two distant sources that are close to each other on the sky have roughly encountered the same large-scale structure along their lines of sight, and their distortions (i.e., image ellipticities) are therefore expected to be correlated (both in magnitude and in orientation). Such correlations have been observed (see Fig. 2.38), and detailed modeling of these results shows that the variance of the matter density field on scales of $8\,h^{-1}$Mpc is about 0.7 - 0.9 (e.g., Van Waerbeke et al., 2001), slightly lower than that of the distribution of bright galaxies.

Since the matter distribution around a given galaxy or cluster will cause a distortion of its background galaxies, weak lensing can also be used to probe the matter distributions around galaxies and clusters. In the case of clusters, one can often detect a sufficient number of background galaxies to reliably measure the shear induced by its gravitational potential. Weak lensing therefore offers a means of measuring the total gravitational mass of an individual (massive) cluster. In the case of individual galaxies, however, one typically has only a few background galaxies available. Consequently, the weak lensing signal is far too weak to detect around individual galaxies. However, by stacking the images of many foreground galaxies (for example, according to their luminosity), one obtains sufficient signal-to-noise to measure the shear, which reflects the *average* mass distribution around the stacked galaxies. This technique is called galaxy-galaxy lensing, and has been used to demonstrate that galaxies are surrounded by extended dark matter halos with masses 10 to 100 times more massive than the galaxies themselves (e.g., Mandelbaum et al., 2006).

2.8 The Intergalactic Medium

The intergalactic medium (IGM) is the medium that permeates the space in between galaxies. In the framework laid out in Chapter 1, galaxies form by the gravitational aggregation of gas in a medium which was originally quite homogeneous. In this scenario, the study of the IGM is an inseparable part of galaxy formation, because it provides us with the properties of the gas from which galaxies form.

The properties of the IGM can be probed observationally by its emission and by its absorption of the light from background sources. If the medium is sufficiently dense and hot, it can be observed in X-ray emission, as is the case for the intracluster medium described in §2.5.1. However, in general the density of the IGM is too low to produce detectable emission, and its properties have to be determined from absorption studies.

2.8.1 The Gunn-Peterson Test

Much information about the IGM has been obtained through its absorption of light from distant quasars. Quasars are not only bright, so that they can be observed out to large distances, but also have well-behaved continua, against which absorption can be analyzed relatively easily. One of the most important tests of the presence of intergalactic neutral hydrogen was proposed by Gunn & Peterson (1965). The Gunn-Peterson test makes use of the fact that the Lyα absorption of neutral hydrogen at $\lambda_\alpha = 1216\,\text{Å}$ has a very large cross section. When the ultraviolet continuum of a distant quasar (assumed to have redshift z_Q) is shifted to 1216 Å at some redshift $z < z_\text{Q}$, the radiation would be absorbed at this redshift if there were even a small amount of neutral hydrogen. Thus, if the Universe were filled with a diffuse distribution of neutral hydrogen, photons bluer than Lyα would be significantly absorbed, causing a significant decrement of flux in the observed quasar spectrum at wavelengths shorter than $(1+z_\text{Q})\lambda_\alpha$. Using the hydrogen Ly$\alpha$ cross section and the definition of optical depth (see Chapter **??** for details), one obtains that the proper number density of HI atoms obeys

$$n_\text{HI}(z) \sim 2.42 \times 10^{-11} \tau(z) h H(z)/H_0\,\text{cm}^{-3}\,, \tag{2.47}$$

where $H(z)$ is Hubble's constant at redshift z, and $\tau(z)$ is the absorption optical depth out to z that can be determined from the flux decrements in quasar spectra. Observations show that the Lyα absorption optical depth is much smaller than unity out to $z \lesssim 6$. The implied density of neutral hydrogen in the diffuse IGM is thus much lower than the mean gas density in the Universe (which is about $10^{-7}\,\text{cm}^{-3}$). This suggests that the IGM must be highly ionized at redshifts $z \lesssim 6$.

As we will show in Chapter **??**, the IGM is expected to be highly neutral after recombination, which occurs at a redshift $z \sim 1000$. Therefore, the fact that the IGM is highly ionized at $z \sim 6$ indicates that the Universe must have undergone some phase transition, from being largely neutral to being highly ionized, a process called reionization. It is generally believed that photoionization due to energetic photons (with energies above the Lyman limit) are responsible for the reionization. This requires the presence of effective emitters of UV photons at high redshifts. Possible candidates include quasars, star-forming galaxies and the first generation of stars. But to this date the actual ionizing sources have not yet been identified, nor is it clear at what redshift reionization occurred. The highest redshift quasars discovered to date, which are close to $z = 6.5$, show almost no detectable flux at wavelengths shorter than $(1+z)\lambda_\alpha$ (Fan et al., 2006). Although this seems to suggest that the mass density of neutral hydrogen increases rapidly at around this redshift, it is not straightforward to convert such flux decrements into an absorption optical depth or a neutral hydrogen fraction, mainly because any $\tau \gg 1$ can result in an almost complete absorption of the flux. Therefore it is currently still unclear whether the Universe became (re-)ionized at a redshift just above 6 or at a significantly higher redshift. At the time of writing, several facilities are being constructed that will attempt to detect 21cm line emission from neutral hydrogen at high redshifts. It is anticipated that these experiments will shed important light on the detailed reionization history of the Universe, as we discuss in some detail in §**??**.

2.8.2 Quasar Absorption Line Systems

Although the flux blueward of $(1+z_\text{Q})\lambda_\alpha$ is not entirely absorbed, quasar spectra typically reveal a large number of absorption lines in this wavelength range (see Fig. 2.39). These absorption lines are believed to be produced by intergalactic clouds that happen to lie along the line of sight from the observer to the quasar, and can be used to probe the properties of the IGM. Quasar absorption line systems are grouped into several categories:

Table 2.8. *Properties of Common Absorption Lines in Quasar Spectra.*

System:	$\log(N_{\rm HI}/{\rm cm}^{-2})$	$b/({\rm km\,s}^{-1})$	$Z/Z_\odot$	$\log(N_{\rm HI}/N_{\rm H})$
Lyα forest	12.5 - 17	15 - 40	< 0.01	< -3
Lyman limit	> 17	~ 100	~ 0.1	> -2
sub-DLA	19 - 20.3	~ 100	~ 0.1	> -1
DLA	> 20.3	~ 100	~ 0.1	~ 0
CIV	> 15.5	~ 100	~ 0.1	> -3
MgII	> 17	~ 100	~ 0.1	> -2

Table 2.9. *Redshift Evolution of Quasar Absorption Line Systems.*

System:	z-range	γ	Reference
Lyα forest	2.0 - 4.0	~ 2.5	Kim et al. (1997)
Lyα forest	0.0 - 1.5	~ 0.15	Weymann et al. (1998)
Lyman limit	0.3 - 4.1	~ 1.5	Stengler-Larrea et al. (1995)
Damped Lyα	0.1 - 4.7	~ 1.3	Storrie-Lombardi et al. (1996a)
CIV	1.3 - 3.4	~ -1.2	Sargent et al. (1988)
MgII	0.2 - 2.2	~ 0.8	Steidel & Sargent (1992)

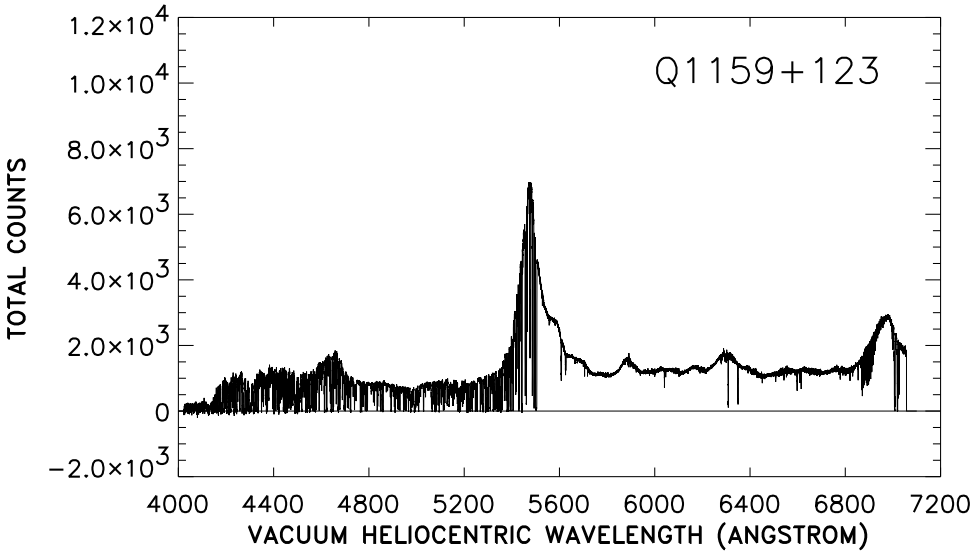


Fig. 2.39. The spectrum of a QSO that reveals a large number of absorption lines due to the IGM. The strongest peak at 5473 Å is the emission line due to Lyα at a rest-frame wavelength of 1216 Å. The numerous absorption lines at $\lambda < 5473$ Å make up the Lyα forest which is due to Lyα absorption of neutral hydrogen clouds between the QSO and the Earth. The break at 4150 Å is due to a Lyman limit cloud which is optically thick at the hydrogen Lyman edge (rest-frame wavelength of 912 Å). The relatively sparse lines to the right of the Lyα emission line are due to absorption by metal atoms associated with the absorbing clouds. [Adapted from Songaila (1998) by permission of AAS]

- Lyα forest: These are narrow lines produced by HI Lyα absorption. They are numerous and appear as a 'forest' of lines blueward of the Lyα emission line of a quasar.

- Lyman-limit systems (LLS): These are systems with HI column densities $N_{\rm HI} \gtrsim 10^{17}\,{\rm cm}^{-2}$, at which the absorbing clouds are optically thick to the Lyman-limit photons (912 Å). These systems appear as continuum breaks in quasar spectra at the redshifted wavelength $(1+z_{\rm a})\times 912\,$Å, where $z_{\rm a}$ is the redshift of the absorber.

- Damped Lyα systems (DLAs): These systems are produced by HI Lyα absorption of gas clouds with HI column densities, $N_{\rm HI} \gtrsim 2\times 10^{20}\,{\rm cm}^{-2}$. Because the Ly$\alpha$ absorption optical depth at such column densities is so large, the quasar continuum photons are completely absorbed near the line center and the line profile is dominated by the damping wing due to the natural (Lorentz) broadening of the absorption line. DLAs with column densities in the range $10^{19}\,{\rm cm}^{-2} < N_{\rm HI} < 2\times 10^{20}\,{\rm cm}^{-2}$ also exhibit damping wings, and are sometimes called sub-DLAs (Péroux et al., 2002). They differ from the largely neutral DLAs in that they are still significantly ionized.

- Metal absorption line systems: In addition to the hydrogen absorption line systems listed above, QSO spectra also frequently show absorption lines due to metals. The best known examples are MgII systems and CIV systems, which are caused by the strong resonance-line doublets MgII$\lambda\lambda 2796, 2800$ and CIV$\lambda\lambda 1548, 1550$, respectively. Note that both doublets have restframe wavelengths longer than $\lambda_{{\rm Ly}\alpha} = 1216$Å. Consequently, they can appear on the red side of the Lyα emission line of the QSO, which makes them easily identifiable because of the absence of confusion from the Lyα forest.

Note that a single absorber may be detected as more than one absorption system. For example, an absorber at $z_{\rm a}$ may be detected as a HI Lyα line at $\lambda = (1+z_{\rm a})\times 1216\,$Å, as a CIV system at $\lambda = (1+z_{\rm a})\times 1548\,$Å, if it has a sufficiently large abundance of CIV ions, and as a Lyman-limit system at $\lambda = (1+z_{\rm a})\times 912\,$Å, if its HI column density is larger than $\sim 10^{17}\,{\rm cm}^{-2}$.

In addition to the most common absorption systems listed above, other line systems are also frequently identified in quasar spectra. These include low ionization lines of heavy elements, such as CII, MgI, FeII etc, and the more highly ionized lines, such as SiIV and NV. Highly-ionized lines such as OVI and OVII are also detected in the UV and/or X-ray spectra of quasars. Since the ionization state of an absorbing cloud depends on its temperature, highly-ionized lines, such as OVI and OVII, in general signify the existence of hot ($\sim 10^6\,$K) gas, while low-ionization lines, such as HI, CII and MgII, are more likely associated with relatively cold ($\sim 10^4\,$K) gas.

For a given quasar spectrum, absorption line systems are identified by decomposing the spectrum into individual lines with some assumed profiles (e.g. the Voigt profile, see §**??**). By modeling each system in detail, one can in principle obtain its column density, b-parameter (defined as $b=\sqrt{2}\sigma$, where σ is the velocity dispersion of the absorbing gas), ionization state, and temperature. If both hydrogen and metal systems are detected, one may also estimate the metallicity of the absorbing gas. Table 2.8 lists the typical values of these quantities for the most commonly detected absorption systems mentioned above.

The evolution of the number of absorption systems is described by the number of systems per unit redshift, d$\mathcal{N}$/dz, as a function of z. This relation is usually fitted by a power law d$\mathcal{N}$/dz$\propto (1+z)^{\gamma}$, and the values of γ for different systems are listed in Table 2.9. The distribution of absorption line systems with respect to the HI-column density is shown in Fig. 2.40. Over the whole observed range, this distribution follows roughly a power law, d$\mathcal{N}$/d$N_{\rm HI}\propto N_{\rm HI}^{-\beta}$, with $\beta\sim 1.5$.

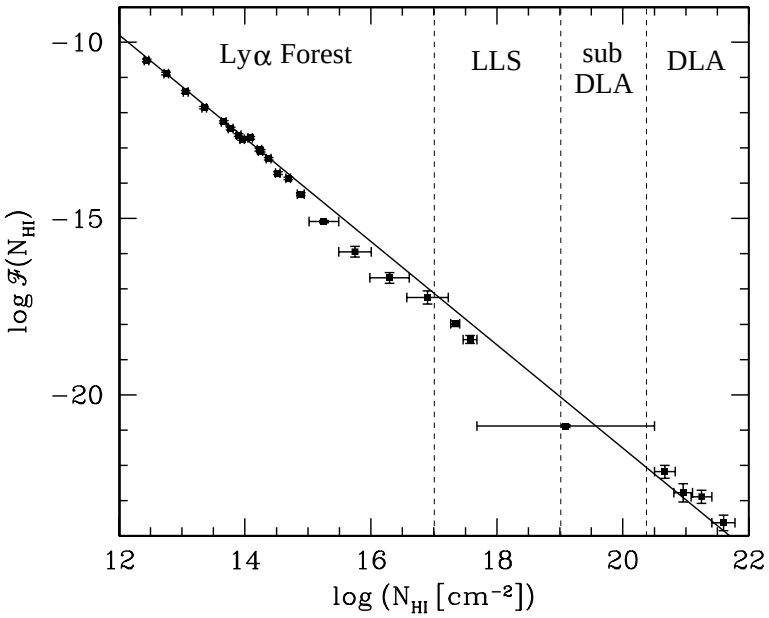


Fig. 2.40. The HI column density distribution of QSO absorption line systems. Here $\mathscr{F}(N_{\rm HI})$ is defined as the number of absorption lines per unit column density, per unit X (which is a quantity that is related to redshift according to Eq. [**??**]). The solid line corresponds to $\mathscr{F}(N_{\rm HI}) \propto N_{\rm HI}^{-1.46}$, which fits the data reasonably well over the full 10 orders of magnitude in column density. [Based on data published in Petitjean et al. (1993) and Hu et al. (1995)]

From the observed column density distribution, one can estimate the mean mass density of neutral hydrogen that is locked up in quasar absorption line systems:

$$\rho_{\rm HI}(z) = \left(\frac{{\rm d}l}{{\rm d}z}\right)^{-1} m_{\rm H} \int N_{\rm HI} \frac{{\rm d}^2\mathscr{N}}{{\rm d}N_{\rm HI}\,{\rm d}z}\,{\rm d}N_{\rm HI}\,, \tag{2.48}$$

where ${\rm d}l/{\rm d}z$ is the physical length per unit redshift at z (see §**??**). Given that ${\rm d}\mathscr{N}/{\rm d}N_{\rm HI}$ is a power law with index ~ -1.5, $\rho_{\rm HI}$ is dominated by systems with the highest $N_{\rm HI}$, i.e. by damped Lyα systems. Using the observed HI-column density distribution, one infers that about 5% of the baryonic material in the Universe is in the form of HI gas at $z \sim 3$ (e.g., Storrie-Lombardi et al., 1996b). In order to estimate the total hydrogen mass density associated with quasar absorption line systems, however, one must know the neutral fraction, $N_{\rm HI}/N_{\rm H}$, as a function of $N_{\rm HI}$. This fraction depends on the ionization state of the IGM. Detailed modeling shows that the Lyα forest systems are highly ionized, and that the main contribution to the total (neutral plus ionized) gas density comes from absorption systems with $N_{\rm HI} \sim 10^{14}\,{\rm cm}^{-2}$. The total gas mass density at $z \sim 3$ thus inferred is comparable to the total baryon density in the Universe (e.g., Rauch et al., 1997; Weinberg et al., 1997).

Quasar absorption line systems with the highest HI column densities are expected to be gas clouds in regions of high gas densities where galaxies and stars may form. It is therefore not surprising that these systems contain metals. Observations of damped Lyα systems show that they have typical metallicities about 1/10 of that of the Sun (e.g., Pettini et al., 1990; Kulkarni et al., 2005), lower than that of the ISM in the Milky Way. This suggests that these systems may be associated with the outer parts of galaxies, or with galaxies in which only a small fraction of the gas has formed stars. More surprising is the finding that most, if not all, of the Lyα forest

lines also contain metals, although the metallicities are generally low, typically about $1/1000$ to $1/100$ of that of the Sun (e.g. Simcoe et al., 2004). There is some indication that the metallicity increases with HI column density, but the trend is not strong. Since star formation requires relatively high column densities of neutral hydrogen (see Chapter **??**), the metals observed in absorption line systems with low HI-column densities most likely originate from, and have been expelled by, galaxies at relatively large distances.

2.9 The Cosmic Microwave Background

The cosmic microwave background (CMB) was discovered by Penzias and Wilson in 1965 when they were commissioning a sensitive receiver at centimeter wavelengths in Bell Telephone Laboratories. It was quickly found that this radiation background was highly isotropic on the sky and has a spectrum close to that of a blackbody with a temperature of about 3 K. The existence of such a radiation background was predicted by Gamow, based on his model of a hot big bang cosmology (see §1.4.2), and it therefore did not take long before the cosmological significance of this discovery was realized (e.g., Dicke et al., 1965).

The observed properties of the CMB are most naturally explained in the standard model of cosmology. Since the early Universe was dense, hot and highly ionized, photons were absorbed and re-emitted many times by electrons and ions and so a blackbody spectrum could be established in the early Universe. As the Universe expanded and cooled and the density of ionized material dropped, photons were scattered less and less often and eventually could propagate freely to the observer from a last-scattering surface, inheriting the blackbody spectrum.

Because the CMB is so important for our understanding of the structure and evolution of the Universe, there have been many attempts in the 1970s and 1980s to obtain more accurate measurements of its spectrum. Since the atmospheric emission is quite close to the peak wavelength of a 3 K blackbody spectrum, most of these measurements were carried out using high-altitude balloon experiments (for a discussion of early CMB experiments, see Partridge, 1995).

A milestone in CMB experiments was the launch by NASA in November 1989 of the Cosmic Background Explorer (COBE), a satellite devoted to accurate measurements of the CMB over the entire sky. Observations with the Far InfraRed Absolute Spectrophotometer (FIRAS) on board COBE showed that the CMB has a spectrum that is perfectly consistent with a blackbody spectrum, to exquisite accuracy, with a temperature $T = 2.728 \pm 0.002\,\mathrm{K}$. As we will see in §**??** the lack of any detected distortions from a pure blackbody spectrum puts strong constraints on any processes that may change the CMB spectrum after it was established in the early Universe.

Another important observational result from COBE is the detection, for the first time, of anisotropy in the CMB. Observations with the Differential Microwave Radiometers (DMR) on board COBE have shown that the CMB temperature distribution is highly isotropic over the sky, confirming earlier observational results, but also revealed small temperature fluctuations (see Fig. 2.41). The observed temperature map contains a component of anisotropy on very large angular scales, which is well described by a dipole distribution over the sky,

$$T(\alpha) = T_0 \left(1 + \frac{v}{c}\cos\alpha\right) , \tag{2.49}$$

where α is the angle of the line of sight relative to a specific direction. This component can be explained as the Doppler effect caused by the motion of the Earth with a velocity $v = 369 \pm 3\,\mathrm{km\,s^{-1}}$ towards the direction $(l,b) = (264.31^\circ \pm 0.20^\circ, 48.05^\circ \pm 0.10^\circ)$ in Galactic coordinates (Lineweaver et al., 1996). Once this dipole component is subtracted, the map of the temperature fluctuations looks like that shown in the lower left panel of Fig. 2.41. In addition to emission

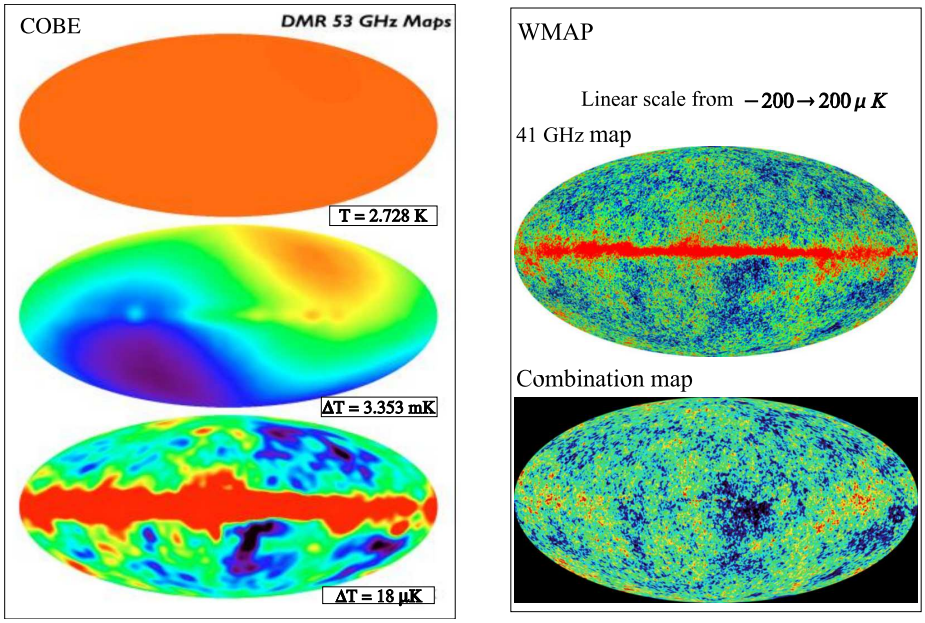


Fig. 2.41. Temperature maps of the CMB in galactic coordinates. The three panels on the left show the temperature maps obtained by the DMR on board the COBE satellite [Courtesy of NASA Goddard Space Flight Center]. The upper panel shows the near-uniformity of the CMB brightness; the middle panel is the map after subtraction of the mean brightness, showing the dipole component due to our motion with respect to the background; and the bottom panel shows the temperature fluctuations after subtraction of the dipole component. Emission from the Milky Way is evident in the bottom image. The two right panels show the temperature maps observed by WMAP from the first year of data [Courtesy of WMAP Science Team], one is from the 41 GHz channel and the other is a linear combination of 5 channels. Note that the large-scale temperature fluctuations in the COBE map at the bottom are clearly seen in the WMAP maps, and that the WMAP angular resolution (about 0.5°) is much higher than that of COBE (about 7°).

from the Milky Way, it reveals fluctuations in the CMB temperature with an amplitude of the order of $\Delta T/T \sim 2\times 10^{-5}$.

Since the angular resolution of the DMR is about 7°, COBE observations cannot reveal anisotropy in the CMB on smaller angular scales. Following the detection by COBE, there have been a large number of experiments to measure small scale CMB anisotropies, and many important results have come out in recent years. These include the results from balloon-borne experiments such as Boomerang (de Bernardis et al., 2000) and Maxima (Hanany et al., 2000), from ground-based interferometers such as the Degree Angular Scale Interferometer (DASI; Halverson et al., 2002) and the Cosmic Background Imager (CBI; Mason et al., 2002), and from an all-sky satellite experiment called the Wilkinson Microwave Anisotropy Probe (WMAP; Bennett et al., 2003; Hinshaw et al., 2007). These experiments have provided us with extremely detailed and accurate maps of the anisotropies in the CMB, such as that obtained by WMAP shown in the right panels of Fig. 2.41.

In order to quantify the observed temperature fluctuations, a common practice is to expand the map in spherical harmonics,

$$\frac{\Delta T}{T}(\vartheta,\varphi) \equiv \frac{T(\vartheta,\varphi)-\overline{T}}{\overline{T}} = \sum_{\ell,m} a_{\ell m} Y_{\ell,m}(\vartheta,\varphi)\,. \tag{2.50}$$

The angular power spectrum, defined as $C_\ell \equiv \langle |a_{\ell m}|^2\rangle^{1/2}$ (where $\langle\cdots\rangle$ denotes averaging over m),

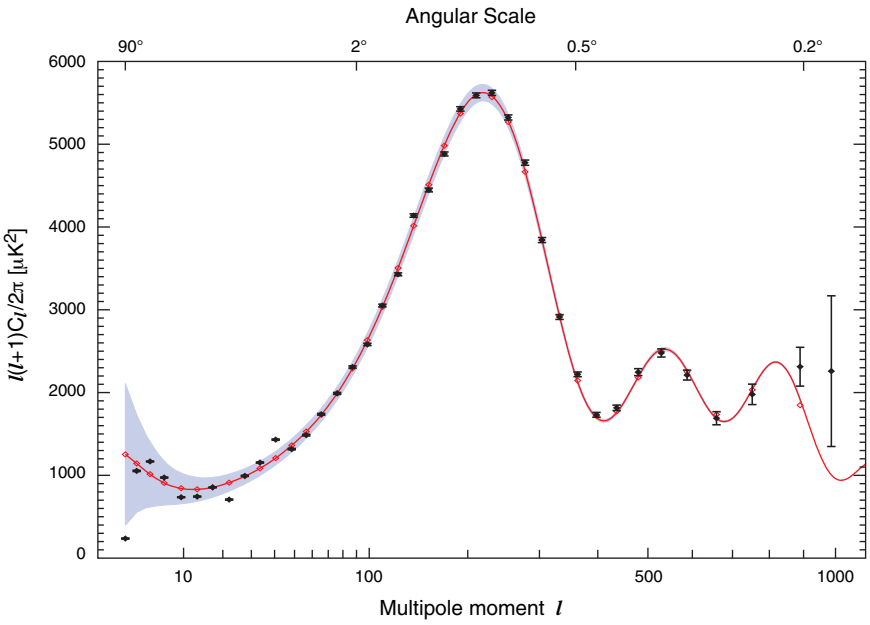


Fig. 2.42. The angular power spectrum, C_ℓ, of the CMB temperature fluctuations in the WMAP full-sky map. This shows the relative brightness of the 'spots' in the CMB temperature map vs. the size of the spots. The shape of this curve contains a wealth of information about the geometry and matter content of the Universe. The curve is the model prediction for the best-fit ΛCDM cosmology. [Adapted from Hinshaw et al. (2007) by permission of AAS]

can be used to represent the amplitudes of temperature fluctuations on different angular scales. Fig. 2.42 shows the temperature power spectrum obtained by the WMAP satellite. As one can see, the observed C_ℓ as a function of ℓ shows complex features. These observational results are extremely important for our understanding of the structure formation in the Universe. First of all, the observed high degree of isotropy in the CMB gives strong support for the assumption of the standard cosmology that the Universe is highly homogeneous and isotropic on large scales. Second, the small temperature fluctuations observed in the CMB are believed to be caused by the density perturbations at the time when the Universe became transparent to CMB photons. These same density perturbations are thought to be responsible for the formation of structures in the Universe. So the temperature fluctuations in the CMB may be used to infer the properties of the initial conditions for the formation of galaxies and other structures in the Universe. Furthermore, the observations of CMB temperature fluctuations can also be used to constrain cosmological parameters. As we will discuss in detail in Chapter **??**, the peaks and valleys in the angular power spectrum are caused by acoustic waves present at the last scattering surface of the CMB photons. The heights (depths) and positions of these peaks (valleys) depend not only on the density of baryonic matter, but also on the total mean density of the Universe, Hubble's constant and other cosmological parameters. Modeling the angular power spectrum of the CMB temperature fluctuations can therefore provide constraints on all of these cosmological parameters.

2.10 The Homogeneous and Isotropic Universe

As we will see in Chapter **??**, the standard cosmological model is based on the 'Cosmological Principle' according to which the Universe is homogeneous and isotropic on large scales. As we have seen, observations of the CMB and of the large-scale spatial distribution of galaxies offer strong support for this cosmological principle. Since according to Einstein's General Relativity the spacetime geometry of the Universe is determined by the matter distribution in the Universe, this large-scale distribution of matter has important implications for the large-scale geometry of spacetime.

For a homogeneous and isotropic universe, its global properties (such as density and pressure) at any time must be the same as those in any small volume. This allows one to study the global properties of the Universe by examining the properties of a small volume within which Newtonian physics is valid. Consider a (small) spherical region of fixed mass M. Since the Universe is homogeneous and isotropic, the radius R of the sphere should satisfy the following Newtonian equation†

$$\ddot{R} = -\frac{GM}{R^2}\,. \tag{2.51}$$

Note that, because of the homogeneity, there is no force due to pressure gradients and that only the mass within the sphere is relevant for the motion of R. This follows directly from Birkhoff's theorem, according to which the gravitational acceleration at any radius in a spherically symmetric system depends only on the mass within that radius. For a given M, the above equation can be integrated once to give

$$\frac{1}{2}\dot{R}^2 - \frac{GM}{R} = E\,, \tag{2.52}$$

where E is a constant, equal to the specific energy of the spherical shell. For simplicity, we write $R = a(t)R_0$, where R_0 is independent of t. It then follows that

$$\frac{\dot{a}^2}{a^2} - \frac{8\pi G\overline{\rho}}{3} = -\frac{Kc^2}{a^2}\,, \tag{2.53}$$

where $\overline{\rho}$ is the mean density of the Universe and $K = -2E/(cR_0)^2$. Unless $E = 0$, which corresponds to $K = 0$, we can always choose the value of R_0 so that $|K| = 1$. So defined, K is called the curvature signature, and takes the value $+1$, 0, or -1. With this normalization, the equation for a is independent of M. As we will see in Chapter **??**, Eq. (2.53) is identical to the Friedmann equation based on General Relativity. For a universe dominated by a non-relativistic fluid, this is not surprising, as it follows directly from the assumption of homogeneity and isotropy. However, as we will see in Chapter **??**, it turns out that Eq. (2.53) also holds even if relativistic matter and/or the energy density associated with the cosmological constant are included.

The quantity $a(t)$ introduced above is called the scale factor, and describes the change of the distance between any two points fixed in the cosmological background. If the distance between a pair of points is l_1 at time t_1, then their distance at some later time t_2 is related to l_1 through $l_2 = l_1 a(t_2)/a(t_1)$. It then follows that at any time t the velocity between any two (comoving) points can be written as

$$\dot{l} = [\dot{a}(t)/a(t)]l\,, \tag{2.54}$$

where l is the distance between the two points at time t. Thus, $\dot{a} > 0$ corresponds to an expanding

† As we will see in Chapter **??**, in General Relativity it is the combination of energy density ρ and pressure P, $\rho + 3P/c^2$, instead of ρ, that acts as the source of gravitational acceleration. Therefore, Eq. (2.51) is not formally valid, even though Eq. (2.53), which derives from it, happens to be correct.

universe, while $\dot{a} < 0$ corresponds to a shrinking universe; the Universe is static only when $\dot{a} = 0$. The ratio $\dot{a}/a$ evaluated at the present time, t_0, is called the Hubble constant,

$$H_0 \equiv \dot{a}_0/a_0\,, \tag{2.55}$$

where $a_0 \equiv a(t_0)$, and the relation between velocity and distance, $\dot{l} = H_0 l$, is known as Hubble's expansion law. Another quantity that characterizes the expansion of the Universe is the deceleration parameter, defined as

$$q_0 \equiv -\frac{\ddot{a}_0 a_0}{\dot{a}_0^2}\,. \tag{2.56}$$

This quantity describes whether the expansion rate of the Universe is accelerating ($q_0 < 0$) or decelerating ($q_0 > 0$) at the present time.

Because of the expansion of the Universe, waves propagating in the Universe are stretched. Thus, photons with a wavelength λ emitted at an earlier time t will be observed at the present time t_0 with a wavelength $\lambda_{\rm obs} = \lambda a_0/a(t)$. Since $a_0 > a(t)$ in an expanding universe, $\lambda_{\rm obs} > \lambda$ and so the wavelength of the photons is redshifted. The amount of redshift z between time t and t_0 is given by

$$z \equiv \frac{\lambda_{\rm obs}}{\lambda} - 1 = \frac{a_0}{a(t)} - 1\,. \tag{2.57}$$

Note that $a(t)$ is a monotonically increasing function of t in an expanding universe, and so redshift is uniquely related to time through the above equation. If an object has redshift z, i.e. its observed spectrum is shifted to the red relative to its rest-frame (intrinsic) spectrum by $\Delta\lambda = \lambda_{\rm obs} - \lambda = z\lambda$, then the photons we observe today from the object were actually emitted at a time t that is related to its redshift z by Eq. (2.57). Because of the constancy of the speed of light, an object's redshift can also be used to infer its distance.

From Eq. (2.53) one can see that the value of K is determined by the mean density $\overline{\rho}_0$ at the present time t_0 and the value of Hubble's constant. Indeed, if we define a critical density

$$\rho_{\rm crit,0} \equiv \frac{3H_0^2}{8\pi G}\,, \tag{2.58}$$

and write the mean density in terms of the density parameter,

$$\Omega_0 \equiv \overline{\rho}_0/\rho_{\rm crit,0}\,, \tag{2.59}$$

then $K = H_0^2 a_0^2(\Omega_0 - 1)$. So $K = -1$, 0 and $+1$ corresponds to $\Omega_0 < 1$, $= 1$ and > 1, respectively. Before discussing the matter content of the Universe, it is illustrative to write the mean density as a sum of several possible components:

(i) non-relativistic matter whose (rest-mass) energy density changes as $\rho_{\rm m} \propto a^{-3}$,
(ii) relativistic matter (such as photons) whose energy density changes as $\rho_{\rm r} \propto a^{-4}$ (the number density changes as a^{-3} while energy is redshifted according to a^{-1}),
(iii) vacuum energy, or the cosmological constant Λ, whose density $\rho_\Lambda = c^2\Lambda/8\pi G$ is a constant.

Thus,

$$\Omega_0 = \Omega_{\rm m,0} + \Omega_{\rm r,0} + \Omega_{\Lambda,0}\,, \tag{2.60}$$

and Eq. (2.53) can be written as

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 E^2(z)\,, \tag{2.61}$$

where

$$E(z) = \left[\Omega_{\Lambda,0} + (1-\Omega_0)(1+z)^2 + \Omega_{\rm m,0}(1+z)^3 + \Omega_{\rm r,0}(1+z)^4\right]^{1/2} \tag{2.62}$$

with z related to $a(t)$ by Eq. (2.57). In order to solve for $a(t)$, we must know the value of H_0 and the energy (mass) content ($\Omega_{\rm m,0}$, $\Omega_{\rm r,0}$, $\Omega_{\Lambda,0}$) at the present time. The deceleration parameter defined in Eq. (2.56) is related to these parameters by

$$q_0 = \frac{\Omega_{\rm m,0}}{2} + \Omega_{\rm r,0} - \Omega_{\Lambda,0}\,. \tag{2.63}$$

A particularly simple case is the Einstein-de Sitter model in which $\Omega_{\rm m,0} = 1$, $\Omega_{\rm r,0} = \Omega_{\Lambda,0} = 0$ (and so $q_0 = 1/2$). It is then easy to show that $a(t) \propto t^{2/3}$. Another interesting case is a flat model in which $\Omega_{\rm m,0} + \Omega_{\Lambda,0} = 1$ and $\Omega_{\rm r,0} = 0$. In this case, $q_0 = 3\Omega_{\rm m,0}/2 - 1$, so that $q_0 < 0$ (i.e. the expansion is accelerating at the present time) if $\Omega_{\rm m,0} < 2/3$.

2.10.1 The Determination of Cosmological Parameters

As shown above, the geometry of the Universe in the standard model is specified by a set of cosmological parameters. The values of these cosmological parameters can therefore be estimated by measuring the geometrical properties of the Universe. The starting point is to find two observables that are related to each other only through the geometrical properties of the Universe. The most important example here is the redshift-distance relation. As we will see in Chapter **??**, two types of distances can be defined through observational quantities. One is the luminosity distance, d_L, which relates the luminosity of an object, L, to its flux, f, according to $L = 4\pi d_L^2 f$. The other is the angular-diameter distance, d_A, which relates the physical size of an object, D, to its angular size, θ, via $D = d_A\theta$. In general, the redshift-distance relation can formally be written as

$$d(z) = \frac{cz}{H_0}\left[1 + \mathscr{F}_d(z;\Omega_{\rm m,0},\Omega_{\Lambda,0},\cdots)\right], \tag{2.64}$$

where d stands either for d_L or d_A, and by definition $\mathscr{F}_d \ll 1$ for $z \ll 1$. For redshifts much smaller than 1, the redshift-distance relation reduces to the Hubble expansion law $cz = H_0 d$, and so the Hubble constant H_0 can be obtained by measuring the redshift and distance of an object (ignoring, for the moment, that objects can have peculiar velocities). Redshifts are relatively easy to obtain from the spectra of objects, and in §2.1.3 we have seen how to measure the distances of a few classes of astronomical objects. The best estimate of the Hubble constant at the present comes from Cepheids observed by the HST, and the result is

$$H_0 = 100h\,{\rm km\,s^{-1}\,Mpc^{-1}}, \quad \text{with} \quad h = 0.72 \pm 0.08 \tag{2.65}$$

(Freedman et al., 2001).

In order to measure other cosmological parameters, one has to determine the non-linear terms in the redshift-distance relation, which typically requires objects at $z \gtrsim 1$. For example, measuring the light curves of Type Ia supernovae out to $z \sim 1$ has yielded the following constraints

$$0.8\Omega_{\rm m,0} - 0.6\Omega_{\Lambda,0} \sim -0.2 \pm 0.1 \tag{2.66}$$

(e.g., Perlmutter et al., 1999). Using Eq. (2.63) and neglecting $\Omega_{\rm r,0}$ because it is small, the above relation gives $q_0 \sim -0.33 - 0.83\Omega_{\rm m,0}$. Since $\Omega_{\rm m,0} > 0$, we have $q_0 < 0$, i.e. the expansion of the Universe is speeding up at the present time.

Important constraints on cosmological parameters can also be obtained from the angular spectrum of the CMB temperature fluctuations. As shown in Fig. 2.42, the observed angular spectrum C_ℓ contains peaks and valleys, which are believed to be produced by acoustic waves in the baryon-photon fluid at the time of photon-matter decoupling. As we will see in §**??**, the heights/depths

and positions of these peaks/valleys depend not only on the density of baryonic matter in the Universe, but also on the total mean density, Hubble's constant and other cosmological parameters. In particular, the position of the first peak is sensitive to the total density parameter Ω_0 (or the curvature K). Based on the observational results shown in Fig. 2.42, one obtains

$$\begin{aligned} \Omega_0 &= 1.02 \pm 0.02\,; \qquad \Omega_{\rm m,0} h^2 = 0.14 \pm 0.02\,; \\ h &= 0.72 \pm 0.05\,; \qquad \Omega_{\rm b,0} h^2 = 0.024 \pm 0.001\,, \end{aligned} \tag{2.67}$$

where $\Omega_{\rm m,0}$ and $\Omega_{\rm b,0}$ are the density parameters of total matter and of baryonic matter, respectively (Spergel et al., 2007). Note that this implies that the Universe has an almost flat geometry, that matter accounts for only about a quarter of its total energy density, and that baryons account for only ~ 17 percent of the matter.

2.10.2 The Mass and Energy Content of the Universe

There is a fundamental difficulty in directly observing the mass (or energy) densities in different mass components: all that is gold does not glitter. There may well exist matter components with significant mass density which give off no detectable radiation. The only interaction which all components are guaranteed to exhibit is gravity, and thus gravitational effects must be studied if the census is to be complete. The global gravitational effect is the curvature of spacetime which we discussed above. Independent information on the amount of gravitating mass can only be derived from the study of the inhomogeneities in the Universe, even though such studies may never lead to an unambiguous determination of the total matter content. After all, one can imagine adding a smooth and invisible component to any amount of inhomogeneously distributed mass, which would produce no detectable effect on the inhomogeneities.

The most intriguing result of such dynamical studies has been the demonstration that the total mass in large-scale structures greatly exceeds the amount of material from which emission can be detected. This unidentified 'dark matter' (or 'invisible matter') is almost certainly the dominant contribution to the total mass density $\Omega_{\rm m,0}$. Its nature and origin remain one of the greatest mysteries of contemporary astronomy.

(a) Relativistic Components One of the best observed relativistic components of the Universe is the CMB radiation. From its blackbody spectrum and temperature, $T_{\rm CMB} = 2.73\,{\rm K}$, it is easy to estimate its energy density at the present time:

$$\rho_{\gamma,0} \approx 4.7 \times 10^{-34}\,{\rm g\,cm^{-3}}, \;\; \text{or} \;\; \Omega_{\gamma,0} = 2.5 \times 10^{-5} h^{-2}\,. \tag{2.68}$$

As we have seen in Fig. 2.2, the energy density of all other known photon backgrounds is much smaller. The only other relativistic component which is almost certainly present, although not yet directly detected, is a background of neutrinos. As we will see in Chapter **??**, the energy density in this component can be calculated directly from the standard model, and it is expected to be 0.68 times that of the CMB radiation. Since the total energy density of the Universe at the present time is not much smaller than the critical density (see last subsection), the contribution from these relativistic components can safely be ignored at low redshift.

(b) Baryonic Components Stars are made up of baryonic matter, and so a lower limit on the mass density of baryonic matter can be obtained by estimating the mass density of stars in galaxies. The mean luminosity density of stars in galaxies can be obtained from the galaxy luminosity function (see §2.4.1). In the B-band, the best-fit Schechter function parameters are $\alpha \approx -1.2$, $\phi^* \approx 1.2 \times 10^{-2} h^3\,{\rm Mpc^{-3}}$ and $\mathscr{M}^* \approx -20.05 + 5\log h$ (corresponding to $L^* = 1.24 \times 10^{10} h^{-2}\,{\rm L_\odot}$), so that

$$\mathscr{L}_B \approx 2 \times 10^8 h\,{\rm L_\odot\,Mpc^{-3}}\,. \tag{2.69}$$

Dividing this into the critical density leads to a value for the mass per unit observed luminosity of galaxies required for the Universe to have the critical density. This critical mass-to-light ratio is

$$\left(\frac{M}{L}\right)_{B,\rm crit} = \frac{\rho_{\rm crit}}{\mathscr{L}_B} \approx 1500h \left(\frac{\rm M_\odot}{\rm L_\odot}\right)_B . \tag{2.70}$$

Mass-to-light ratios for the visible parts of galaxies can be estimated by fitting their spectra with appropriate models of stellar populations. The resulting mass-to-light ratios tend to be in the range of 2 to $10(\rm M_\odot/L_\odot)$. Adopting $M/L = 5(\rm M_\odot/L_\odot)$ as a reasonable mean value, the global density contribution of stars is

$$\Omega_{\star,0} \sim 0.003h^{-1}. \tag{2.71}$$

Thus, the visible parts of galaxies provide less than one percent of the critical density. In fact, combined with the WMAP constraints on $\Omega_{\rm b,0}$ and the Hubble constant, we find that stars only account for less than 10 percent of all baryons.

So where are the other 90 percent of the baryons? At low redshifts, the baryonic mass locked up in cold gas (either atomic or molecular), and detected either via emission or absorption, only accounts for a small fraction, $\Omega_{\rm cold} \sim 0.0005h^{-1}$ (Fukugita et al., 1998). A larger contribution is due to the hot intracluster gas observed in rich galaxy clusters through their bremsstrahlung emission at X-ray wavelengths (§2.5.1). From the number density of X-ray clusters and their typical gas mass, one can estimate that the total amount of hot gas in clusters is about $(\Omega_{\rm HII})_{\rm cl} \sim 0.0016h^{-3/2}$ (Fukugita et al., 1998). The total gas mass in groups of galaxies is uncertain. Based on X-ray data, Fukugita et al. obtained $(\Omega_{\rm HII})_{\rm group} \sim 0.003h^{-3/2}$. However, the plasma in groups is expected to be colder than that in clusters, which makes it more difficult to detect in X-ray radiation. Therefore, the low X-ray emissivity from groups may also be due to low temperatures rather than due to small amounts of plasma. Indeed, if we assume that the gas/total mass ratio in groups is comparable to that in clusters, then the total gas mass in groups could be larger by a factor of two to three. Even then, the total baryonic mass detected in stars, cold gas and hot gas only accounts for less than 50 percent of the total baryonic mass inferred from the CMB.

The situation is very different at higher redshifts. As discussed in §2.8, the average density of hydrogen inferred from quasar absorption systems at $z \sim 3$ is roughly equal to the total baryon density as inferred from the CMB data. Hence, although we seem to have detected the majority of all baryons at $z \sim 3$, at low redshifts roughly half of the expected baryonic mass is unaccounted for observationally. One possibility is that the gas has been heated to temperatures in the range $10^5 - 10^6\,$K at which it is very difficult to detect. Indeed, recent observations of OVI absorption line systems seem to support the idea that a significant fraction of the IGM at low redshift is part of such a Warm-Hot Intergalactic Medium (WHIM), whose origin may be associated with the formation of large-scale sheets and filaments in the matter distribution (see Chapter **??**).

An alternative explanation for the 'missing baryons' is that a large fraction of the gas detected at $z \sim 3$ has turned into 'invisible' compact objects, such as brown dwarfs or black holes. The problem, though, is that most of these objects are stellar remnants, and their formation requires a star formation rate between $z = 3$ and $z = 0$ that is significantly higher than normally assumed. Not only is this inconsistent with the observation of the global star formation history of the Universe (see §2.6.8), but it would also result in an over-production of metals. This scenario thus seems unlikely. Nevertheless, some observational evidence, albeit controversial, does exist for the presence of a population of compact objects in the dark halo of our Milky Way. In 1986 Bohdan Paczyński proposed to test for the presence of massive compact halo objects (MACHOs) using gravitational lensing. Whenever a MACHO in our Milky Way halo moves across the line-of-sight to a background star (for example, a star in the LMC), it will magnify the flux of the background star, an effect called microlensing. Because of the relative motion of source, lens

and observer, this magnification is time-dependent, giving rise to a characteristic light curve of the background source. In the early 1990s two collaborations (MACHO and EROS) started campaigns to monitor millions of stars in the LMC for a period of several years. This has resulted in the detection of about 20 events in total. The analysis by the MACHO collaboration suggests that about 20 percent of the mass of the halo of the Milky Way could consist of MACHOs with a characteristic mass of $\sim 0.5\,\mathrm{M}_\odot$ (Alcock et al., 2000). The nature of these objects, however, is still unclear. Furthermore, these results are inconsistent with those obtained by the EROS collaboration, which obtained an upper limit for the halo mass fraction in MACHOs of 8 percent, and rule out MACHOs in the mass range $0.6\times10^{-7}\,\mathrm{M}_\odot < M < 15\,\mathrm{M}_\odot$ as the primary occupants of the Milky Way Halo (Tisserand et al., 2007).

(c) Non-Baryonic Dark Matter As is evident from the CMB constraints given by Eq. (2.67) on $\Omega_{\mathrm{m},0}$ and $\Omega_{\mathrm{b},0}$, baryons can only account for ~ 15 – 20 percent of the total matter content in the Universe. And this is supported by a wide range of observations. As we will see in the following chapters, constraints from a number of other measurements, such as cosmic shear, the abundance of massive clusters, large-scale structure, and the peculiar velocity field of galaxies, all agree that $\Omega_{\mathrm{m},0}$ is of the order of 0.3. At the same time, the total baryonic matter density inferred from CMB observations is in excellent agreement with independent constraints from nucleosynthesis and the observed abundances of primordial elements. The inference is that the majority of the matter in the Universe (75 to 80 percent) must be in some non-baryonic form.

One of the most challenging tasks for modern cosmology is to determine the nature and origin of this dark matter component. Particle physics in principle allows for a variety of candidate particles, but without a direct detection it is and will be difficult to discriminate between the various candidates. One thing that is clear from observations is that the distribution of dark matter is typically more extended than that of the luminous matter. As we have seen above, the mass-to-light ratios increase from $M/L \sim 30h(M/L)_\odot$ at a radius of about $30\,h^{-1}$kpc as inferred from the extended rotation curves of spiral galaxies, to $M/L \sim 100h(M/L)_\odot$ at the scale of a few hundred kpc, as inferred from the kinematics of galaxies in groups, to $M/L \sim 350h(M/L)_\odot$ in galaxy clusters, probing scales of the order of 1 Mpc. This latter value is comparable to that of the Universe as a whole, which follows from multiplying the critical mass-to-light ratio given by Eq. (2.70) with $\Omega_{\mathrm{m},0}$, and suggests that the content of clusters, which are the largest virialized structures known, is representative of that of the entire Universe.

All these observations support the idea that galaxies reside in extended halos of dark matter. This in turn puts some constraints on the nature of the dark matter, namely that it has to be relatively cold (i.e., it needs to have initial peculiar velocities that are much smaller than the typical velocity dispersion within an individual galaxy). This coldness is required because otherwise the dark matter would not be able to cluster on galactic scales to form the dark halos around galaxies. Without a better understanding of the nature of the dark matter, we have to live with the vague term, cold dark matter (or CDM), when talking about the main mass component of the Universe.

(d) Dark Energy As we have seen above, the observed temperature fluctuations in the CMB show that the Universe is nearly flat, implying that the mean energy density of the Universe must be close to the critical density, ρ_{crit}. However, studies of the kinematics of galaxies and of large-scale structure in the Universe give a mean mass density that is only about 1/4 to 1/3 of the critical density, in good agreement with the constraints on $\Omega_{\mathrm{m},0}$ from the CMB itself. This suggests that the dominant component of the mass/energy content of the Universe must have a homogenous distribution so that it affects the geometry of the Universe but does not follow the structure in the baryonic and dark matter. An important clue about this dominant component is provided by the observed redshift-distance relation of high-redshift Type Ia supernovae. As shown in §2.10.1, this relation implies that the expansion of the Universe is speeding up at the present time. Since all matter, both baryonic and non-baryonic, decelerates the expansion of the

Universe, the dominant component must be an energy component. It must also be extremely dark, because otherwise it would have been observed.

The nature of this dark energy component is a complete mystery at the present time. As far as its effect on the expansion of the Universe is concerned, it is similar to the cosmological constant introduced by Einstein in his theory of General Relativity to achieve a stationary Universe (Einstein, 1917). The cosmological constant can be considered as an energy component whose density does not change with time. As the Universe expands, it appears as if more and more energy is created to fill the space. This strange property is due to its peculiar equation of state that relates its pressure, P, to its energy density, ρ. In general, we may write $P = w\rho c^2$, and so $w = 0$ for a pressureless fluid and $w = 1/3$ for a radiation field (see §**??**). For a dark energy component with constant energy density, $w = -1$, which means that the fluid actually gains internal energy as it expands, and acts as a gravitational source with a negative effective mass density ($\rho + 3P/c^2 = -2\rho < 0$), causing the expansion of the Universe to accelerate. In addition to the cosmological constant, dark energy may also be related to a scalar field (with $-1 < w < -1/3$). Such a form of dark energy is called quintessence, which differs from a cosmological constant in that it is dynamic, meaning that its density and equation of state can vary through both space and time. It has also been proposed that dark energy has an equation of state parameter $w < -1$, in which case it is called phantom energy. Clearly, a measurement of the value of w will allow us to discriminate between these different models. Currently, the value of w is constrained by a number of observations to be within a relatively narrow range around -1 (e.g. Spergel et al., 2007), consistent with a cosmological constant, but also with both quintessence and phantom energy. The next generation of galaxy redshift surveys and Type Ia supernova searches aim to constrain the value of w to a few percent, in the hope of learning more about the nature of this mysterious and dominant energy component of our Universe.

Bibliography

Abell G. O., 1958, ApJS, 3, 211
Abraham R. G., 1998, astro-ph/9809131
Adami C., Biviano A., Mazure A., 1998, A&A, 331, 439
Adelberger K. L., Steidel C. C., 2000, ApJ, 544, 218
Alcock C., Allsman R. A., Alves D. R., et al. 2000, ApJ, 542, 281
Bahcall N. A., McKay T. A., Annis J., et al. 2003, ApJS, 148, 243
Baldry I. K., Balogh M. L., Bower R. G., et al. 2006, MNRAS, 373, 469
Bardeen J. M., Steinhardt P. J., Turner M. S., 1983, PhRvD, 28, 679
Barnes J. E., 1988, ApJ, 331, 699
Begeman K. G., 1989, A&A, 223, 47
Bekki K., Couch W. J., Drinkwater M. J., 2001, ApJ, 552, L105
Bell E. F., Wolf C., Meisenheimer K., et al. 2004, ApJ, 608, 752
Bell E. F., Zucker D. B., Belokurov V., et al. 2008, ApJ, 680, 295
Bender R., Burstein D., Faber S. M., 1992, ApJ, 399, 462
Bender R., Surma P., Doebereiner S., et al. 1989, A&A, 217, 35
Bennett C. L., Halpern M., Hinshaw G., et al. 2003, ApJS, 148, 1
Berlind A. A., Frieman J., Weinberg D. H., et al. 2006, ApJS, 167, 1
Bernardi M., Sheth R. K., Annis J., et al. 2003, AJ, 125, 1866
Bertola F., Buson L. M., Zeilinger W. W., 1992, ApJ, 401, L79
Bertola F., Capaccioli M., 1975, ApJ, 200, 439
Bessell M. S., 1990, PASP, 102, 1181
Binggeli B., Sandage A., Tarenghi M., 1984, AJ, 89, 64
Binggeli B., Tammann G. A., Sandage A., 1987, AJ, 94, 251
Binney J., 1977, ApJ, 215, 483
Binney J., Gerhard O. E., Stark A. A., et al. 1991, MNRAS, 252, 210
Blain A. W., Smail I., Ivison R. J., Kneib J.-P., 1999, MNRAS, 302, 632
Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, 517
Blumenthal G. R., Pagels H., Primack J. R., 1982, Nature, 299, 37
Böker T., Laine S., van der Marel R. P., et al. 2002, AJ, 123, 1389
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
Bond J. R., Efstathiou G., Silk J., 1980, Phys. Rev. Lett., 45, 1980
Bond J. R., Szalay A. S., 1983, ApJ, 274, 443
Bond J. R., Szalay A. S., Turner M. S., 1982, Phys. Rev. Lett., 48, 1636
Bouwens R. J., Illingworth G. D., Franx M., Ford H., 2007, ApJ, 670, 928
Branch D., Tammann G. A., 1992, ARA&A, 30, 359
Brown M. J. I., Dey A., Jannuzi B. T., et al. 2007, ApJ, 654, 858
Bullock J. S., Dekel A., Kolatt T. S., et al. 2001, ApJ, 555, 240
Buson L. M., Sadler E. M., Zeilinger W. W., et al. 1993, A&A, 280, 409
Butcher H., Oemler Jr. A., 1978, ApJ, 219, 18
Carlberg R. G., Yee H. K. C., Ellingson E., et al. 1997, ApJ, 476, L7
Carr B. J., Bond J. R., Arnett W. D., 1984, ApJ, 277, 445
Cimatti A., Daddi E., Mignoli M., et al. 2002, A&A, 381, L68
Cole S., Aragon-Salamanca A., Frenk C. S., et al. 1994, MNRAS, 271, 781
Colless M., Dalton G., Maddox S., et al. 2001, MNRAS, 328, 1039
Cooper M. C., Newman J. A., Coil A. L., et al. 2007, MNRAS, 376, 1445

Côté P., Ferrarese L., Jordán A., et al. 2007, ApJ, 671, 1456
Côté P., Piatek S., Ferrarese L., et al. 2006, ApJS, 165, 57
Couch W. J., Ellis R. S., Sharples R. M., Smail I., 1994, ApJ, 430, 121
Cowsik R., McClelland J., 1972, Phys. Rev. Lett., 29, 669
Cox A. N., 2000, Allen's astrophysical quantities. Springer, AIP Press, New York
Daddi E., Cimatti A., Renzini A., et al. 2004, ApJ, 617, 746
Dalcanton J. J., Spergel D. N., Summers F. J., 1997, ApJ, 482, 659
Dalton G. B., Maddox S. J., Sutherland W. J., Efstathiou G., 1997, MNRAS, 289, 263
Davies R. L., Efstathiou G., Fall S. M., et al. 1983, ApJ, 266, 41
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Davis M., Faber S. M., Newman J., et al. 2003, in Guhathakurta P., ed., Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 4834, . pp 161–172
de Bernardis P., Ade P. A. R., Bock J. J., et al. 2000, Nature, 404, 955
de Grijs R., Kregel M., Wesson K. H., 2001, MNRAS, 324, 1074
de Jong R. S., 1996a, A&A, 313, 45
de Jong R. S., 1996b, A&A, 313, 377
de Vaucouleurs G., 1974, in Shakeshaft J. R., ed., The Formation and Dynamics of Galaxies Vol. 58 of IAU Symposium, . pp 1–52
Dicke R. H., Peebles P. J. E., 1979, in Hawking S. W., Israel W., eds, General Relativity: An Einstein Centenary Survey . pp 504–517
Dicke R. H., Peebles P. J. E., Roll P. G., Wilkinson D. T., 1965, ApJ, 142, 414
Dickinson M., 1998, in Livio M., et al. eds, The Hubble Deep Field . p. 219
Dickinson M., Papovich C., Ferguson H. C., Budavári T., 2003, ApJ, 587, 25
Doroshkevich A. G., 1970, Astrophysics, 6, 320
Doroshkevich A. G., 1973, Astrophys. Lett., 14, 11
Dressler A., 1980a, ApJS, 42, 565
Dressler A., 1980b, ApJ, 236, 351
Dressler A., 1984, ARA&A, 22, 185
Dressler A., Gunn J. E., 1983, ApJ, 270, 7
Dressler A., Smail I., Poggianti B. M., et al. 1999, ApJS, 122, 51
Drinkwater M. J., Gregg M. D., Hilker M., et al. 2003, Nature, 423, 519
Efstathiou G., Jones B. J. T., 1979, MNRAS, 186, 133
Efstathiou G., Silk J., 1983, Fund. of Cosmic Phys., 9, 1
Efstathiou G., Sutherland W. J., Maddox S. J., 1990, Nature, 348, 705
Einasto J., 1965, Trudy Inst. Astroz. Alma-Ata, 57, 87
Einasto J., Kaasik A., Saar E., 1974, Nature, 250, 309
Einstein A., 1917, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin), pp 142–152
Eke V. R., Baugh C. M., Cole S., et al. 2004, MNRAS, 348, 866
Ellis R. S., Colless M., Broadhurst T., et al. 1996, MNRAS, 280, 235
Fabbiano G., 1989, ARA&A, 27, 87
Faber S. M., Lin D. N. C., 1983, ApJ, 266, L17
Faber S. M., Tremaine S., Ajhar E. A., et al. 1997, AJ, 114, 1771
Faber S. M., Willmer C. N. A., Wolf C., et al. 2007, ApJ, 665, 265
Fan X., Strauss M. A., Becker R. H., et al. 2006, AJ, 132, 117
Farouki R. T., Shapiro S. L., 1982, ApJ, 259, 103
Feast M. W., Catchpole R. M., 1997, MNRAS, 286, L1
Ferguson A., Irwin M., Chapman S., et al. 2007, in de Jong R. S., ed., Island Universes - Structure and Evolution of Disk Galaxies . p. 239
Ferguson A. M. N., Irwin M. J., Ibata R. A., et al. 2002, AJ, 124, 1452
Ferguson H. C., Dickinson M., Williams R., 2000, ARA&A, 38, 667
Ferrarese L., Côté P., Dalla Bontà E., et al. 2006a, ApJ, 644, L21
Ferrarese L., Côté P., Jordán A., et al. 2006b, ApJS, 164, 334
Ferrarese L., van den Bosch F. C., Ford H. C., et al. 1994, AJ, 108, 1598
Franx M., Illingworth G., Heckman T., 1989, AJ, 98, 538
Franx M., Labbé I., Rudnick G., et al. 2003, ApJ, 587, L79
Freedman W. L., Madore B. F., Gibson B. K., et al. 2001, ApJ, 553, 47
Freeman K. C., 1970, ApJ, 160, 811
Fukugita M., Hogan C. J., Peebles P. J. E., 1998, ApJ, 503, 518

Fukugita M., Ichikawa T., Gunn J. E., et al. 1996, AJ, 111, 1748
Gallazzi A., Charlot S., Brinchmann J., et al. 2005, MNRAS, 362, 41
Gamow G., Teller E., 1939, Phys. Rev., 55, 654
Garnavich P. M., Kirshner R. P., Challis P., et al. 1998, ApJ, 493, L53
Geha M., Guhathakurta P., van der Marel R. P., 2002, AJ, 124, 3073
Geller M. J., Huchra J. P., 1983, ApJS, 52, 61
Genzel R., Pichon C., Eckart A., et al. 2000, MNRAS, 317, 348
Gerhard O. E., 1981, MNRAS, 197, 179
Gershtein S. S., Zel'Dovich Y. B., 1966, ZhETF Pis ma Redaktsiiu, 4, 174
Ghez A. M., Duchêne G., Matthews K., et al. 2003, ApJ, 586, L127
Ghez A. M., Salim S., Hornstein S. D., et al. 2005, ApJ, 620, 744
Gilmore G., Wilkinson M. I., Wyse R. F. G., et al. 2007, ApJ, 663, 948
Giovanelli R., Haynes M. P., Herter T., et al. 1997, AJ, 113, 53
Gott III J. R., Thuan T. X., 1976, ApJ, 204, 649
Graham A. W., 2001, AJ, 121, 820
Graham A. W., Guzmán R., 2003, AJ, 125, 2936
Grant N. I., Kuipers J. A., Phillipps S., 2005, MNRAS, 363, 1019
Grebel E. K., 1999, in Whitelock P., Cannon R., eds, The Stellar Content of Local Group Galaxies Vol. 192 of IAU Symposium, . p. 17
Gunn J. E., Gott J. R. I., 1972, ApJ, 176, 1
Gunn J. E., Peterson B. A., 1965, ApJ, 142, 1633
Guth A. H., Pi S.-Y., 1982, Phys. Rev. Lett., 49, 1110
Halverson N. W., Leitch E. M., Pryke C., et al. 2002, ApJ, 568, 38
Hanany S., Ade P., Balbi A., et al. 2000, ApJ, 545, L5
Harrison E. R., 1970, PhRvD, 1, 2726
Hawking S. W., 1982, Phys. Lett. B, 115, 295
Hawkins E., Maddox S., Cole S., et al. 2003, MNRAS, 346, 78
Helmi A., White S. D. M., de Zeeuw P. T., Zhao H., 1999, Nature, 402, 53
Hickson P., 1982, ApJ, 255, 382
Hinshaw G., Nolta M. R., Bennett C. L., et al. 2007, ApJS, 170, 288
Holland W. S., Robson E. I., Gear W. K., et al. 1999, MNRAS, 303, 659
Hopkins A. M., 2004, ApJ, 615, 209
Hoyle F., 1949, in Problems of Cosmical Aerodynamics. Dayton Ohio: Central Air Documents Office
Hu E. M., Kim T.-S., Cowie L. L., et al. 1995, AJ, 110, 1526
Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168
Hubble E., Humason M. L., 1931, ApJ, 74, 43
Ibata R. A., Gilmore G., Irwin M. J., 1994, Nature, 370, 194
Illingworth G., 1977, ApJ, 218, L43
Impey C. D., Sprayberry D., Irwin M. J., Bothun G. D., 1996, ApJS, 105, 209
Jaffe W., Ford H. C., O'Connell R. W., et al. 1994, AJ, 108, 1567
Jeans J. H., 1902, Philosophical Transactions of the Royal Society of London, 199, 1
Jerjen H., Binggeli B., 1997, in Arnaboldi M., et al. eds, The Nature of Elliptical Galaxies; 2nd Stromlo Symposium Vol. 116 of ASP, . p. 239
Jing Y. P., Suto Y., 2002, ApJ, 574, 538
Jørgensen I., Franx M., Kjaergaard P., 1996, MNRAS, 280, 167
Kaiser N., 1984, ApJ, 284, L9
Katz N., 1992, ApJ, 391, 502
Katz N., Gunn J. E., 1991, ApJ, 377, 365
Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201
Kauffmann G., White S. D. M., Heckman T. M., et al. 2004, MNRAS, 353, 713
Kim T.-S., Hu E. M., Cowie L. L., Songaila A., 1997, AJ, 114, 1
Klinkhamer F. R., Norman C. A., 1981, ApJ, 243, L1
Klypin A., Gottlöber S., Kravtsov A. V., Khokhlov A. M., 1999, ApJ, 516, 530
Koester B. P., McKay T. A., Annis J., et al. 2007, ApJ, 660, 239
Komatsu E., Dunkley J., Nolta M. R., et al. 2009, ApJS, 180, 330
Kormendy J., 1985, ApJ, 295, 73
Kormendy J., 2001, in Funes J. G., Corsini E. M., eds, Galaxy Disks and Disk Galaxies Vol. 230 of ASP, . pp 247–256
Kormendy J., Bender R., 1996, ApJ, 464, L119

Kulkarni V. P., Fall S. M., Lauroesch J. T., et al. 2005, ApJ, 618, 68
Labbé I., Franx M., Rudnick G., et al. 2003, AJ, 125, 1107
Lacey C., Cole S., 1993, MNRAS, 262, 627
Larson R. B., 1974a, MNRAS, 166, 585
Larson R. B., 1974b, MNRAS, 169, 229
Larson R. B., 1975, MNRAS, 173, 671
Larson R. B., 1976, MNRAS, 176, 31
Lauer T. R., Ajhar E. A., Byun Y.-I., et al. 1995, AJ, 110, 2622
Le Fèvre O., Vettolani G., Garilli B., et al. 2005, A&A, 439, 845
Lifshitz E. M., 1946, Journal of Physics (Moscow), 10, 116
Lilly S. J., Le Fevre O., Crampton D., et al. 1995, ApJ, 455, 50
Lin Y.-T., Mohr J. J., Stanford S. A., 2004, ApJ, 610, 745
Lineweaver C. H., Tenorio L., Smoot G. F., et al. 1996, ApJ, 470, 38
Lumsden S. L., Nichol R. C., Collins C. A., Guzzo L., 1992, MNRAS, 258, 1
Lyubimov V. A., Novikov E. G., Nozik V. Z., et al. 1980, Phys. Lett. B, 94, 266
MacArthur L. A., Courteau S., Bell E., Holtzman J. A., 2004, ApJS, 152, 175
MacArthur L. A., Courteau S., Holtzman J. A., 2003, ApJ, 582, 689
Maddox S. J., Efstathiou G., Sutherland W. J., Loveday J., 1990a, MNRAS, 242, 43P
Maddox S. J., Efstathiou G., Sutherland W. J., Loveday J., 1990b, MNRAS, 243, 692
Madore B. F., Freedman W. L., 1991, PASP, 103, 933
Mandelbaum R., Seljak U., Kauffmann G., et al. 2006, MNRAS, 368, 715
Martin N. F., Ibata R. A., Chapman S. C., et al. 2007, MNRAS, 380, 281
Mason B. S., Cartwright J. K., Padin S., et al. 2002, in Gurzadyan V. G., et al. eds, The Ninth Marcel Grossmann Meeting . pp 2171–2172
Mateo M. L., 1998, ARA&A, 36, 435
Mathews W. G., Brighenti F., 2003, ARA&A, 41, 191
McGaugh S. S., de Blok W. J. G., 1997, ApJ, 481, 689
Miller C. J., Nichol R. C., Reichart D., et al. 2005, AJ, 130, 968
Misner C. W., 1968, ApJ, 151, 431
Mo H. J., Mao S., White S. D. M., 1998, MNRAS, 295, 319
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Moore B., Governato F., Quinn T., et al. 1998, ApJ, 499, L5
Navarro J. F., Benz W., 1991, ApJ, 380, 320
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Navarro J. F., White S. D. M., 1994, MNRAS, 267, 401
Negroponte J., White S. D. M., 1983, MNRAS, 205, 1009
Nolthenius R., White S. D. M., 1987, MNRAS, 225, 505
Norberg P., Baugh C. M., Hawkins E., et al. 2001, MNRAS, 328, 64
Norberg P., Baugh C. M., Hawkins E., et al. 2002a, MNRAS, 332, 827
Norberg P., Cole S., Baugh C. M., et al. 2002b, MNRAS, 336, 907
Odenkirchen M., Grebel E. K., Dehnen W., et al. 2002, AJ, 124, 1497
Ostriker J. P., 1980, Comments on Astrophysics, 8, 177
Ostriker J. P., Peebles P. J. E., Yahil A., 1974, ApJ, 193, L1
Pahre M. A., Djorgovski S. G., de Carvalho R. R., 1998, AJ, 116, 1591
Partridge R. B., 1995, 3K: The Cosmic Microwave Background Radiation. Cambridge University Press, Cambridge
Pasquali A., van den Bosch F. C., Rix H.-W., 2007, ApJ, 664, 738
Peacock J. A., 2002, in Metcalfe N., Shanks T., eds, A New Era in Cosmology Vol. 283 of ASP, . p. 19
Peebles P. J. E., 1965, ApJ, 142, 1317
Peebles P. J. E., 1971, A&A, 11, 377
Peebles P. J. E., 1982, ApJ, 263, L1
Peebles P. J. E., Yu J. T., 1970, ApJ, 162, 815
Peletier R. F., Balcells M., 1996, AJ, 111, 2238
Peletier R. F., Davies R. L., Illingworth G. D., et al. 1990, AJ, 100, 1091
Penzias A. A., Wilson R. W., 1965, ApJ, 142, 419
Perlmutter S., Aldering G., Goldhaber G., et al. 1999, ApJ, 517, 565
Péroux C., Dessauges-Zavadsky M., Kim T., et al. 2002, Astrophys. Space Science, 281, 543
Petitjean P., Webb J. K., Rauch M., et al. 1993, MNRAS, 262, 499
Pettini M., Boksenberg A., Hunstead R. W., 1990, ApJ, 348, 48

Phillips A. C., Illingworth G. D., MacKenty J. W., Franx M., 1996, AJ, 111, 1566
Phillips M. M., Lira P., Suntzeff N. B., et al. 1999, AJ, 118, 1766
Pierce M. J., Tully R. B., 1992, ApJ, 387, 47
Pohlen M., Dettmar R.-J., Lütticke R., 2000, A&A, 357, L1
Press W. H., Schechter P., 1974, ApJ, 187, 425
Quintana H., 1979, AJ, 84, 15
Rauch M., Miralda-Escude J., Sargent W. L. W., et al. 1997, ApJ, 489, 7
Ravindranath S., Ho L. C., Peng C. Y., et al. 2001, AJ, 122, 653
Rees M. J., Ostriker J. P., 1977, MNRAS, 179, 541
Refregier A., Rhodes J., Groth E. J., 2002, ApJ, 572, L131
Reines F., Sobel H. W., Pasierb E., 1980, Phys. Rev. Lett., 45, 1307
Rest A., van den Bosch F. C., Jaffe W., et al. 2001, AJ, 121, 2431
Roberts M. S., Haynes M. P., 1994, ARA&A, 32, 115
Roberts M. S., Hogg D. E., Bregman J. N., et al. 1991, ApJS, 75, 751
Roberts M. S., Rots A. H., 1973, A&A, 26, 483
Rubin V. C., Thonnard N., Ford Jr. W. K., 1978, ApJ, 225, L107
Rubin V. C., Thonnard N., Ford Jr. W. K., 1980, ApJ, 238, 471
Sackett P. D., Morrison H. L., Harding P., Boroson T. A., 1994, Nature, 370, 441
Saglia R. P., Burstein D., Baggley G., et al. 1997, MNRAS, 292, 499
Sandage A., Binggeli B., 1984, AJ, 89, 919
Sandage A., Visvanathan N., 1978, ApJ, 223, 707
Sargent W. L. W., Boksenberg A., Steidel C. C., 1988, ApJS, 68, 539
Sato H., 1971, Prog. of Theoretical Phys., 45, 370
Sato H., Takahara F., 1980, Prog. of Theoretical Phys., 64, 2029
Schechter P., 1976, ApJ, 203, 297
Schmidt M., 1959, ApJ, 129, 243
Schneider D. P., Gunn J. E., Hoessel J. G., 1983, ApJ, 268, 476
Schödel R., Ott T., Genzel R., Eckart A., et al. 2003, ApJ, 596, 1015
Schramm D. N., Steigman G., 1981, ApJ, 243, 1
Scott D., 2000, in Courteau S., Willick J., eds, Cosmic Flows Workshop Vol. 201 of ASP, . p. 403
Sérsic J. L., 1968, Atlas de Galaxias Australes. Observatorio Astronomico; Cordoba, Argentina
Sevenster M. N., 1996, in Buta R., et al. eds, IAU Colloq. 157: Barred Galaxies Vol. 91 of ASP, . p. 536
Shen S., Mo H. J., White S. D. M., et al. 2003, MNRAS, 343, 978
Sievers J. L., Bond J. R., Cartwright J. K., et al. 2003, ApJ, 591, 599
Silk J., 1968, ApJ, 151, 459
Silk J., 1977, ApJ, 211, 638
Simcoe R. A., Sargent W. L. W., Rauch M., 2004, ApJ, 606, 92
Simien F., de Vaucouleurs G., 1986, ApJ, 302, 564
Smail I., Ivison R. J., Blain A. W., 1997, ApJ, 490, L5
Smoot G. F., Bennett C. L., Kogut A., et al. 1992, ApJ, 396, L1
Soifer B. T., Rowan-Robinson M., Houck J. R., et al. 1984, ApJ, 278, L71
Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087
Songaila A., 1998, AJ, 115, 2184
Spergel D. N., Bean R., Doré O., et al. 2007, ApJS, 170, 377
Spergel D. N., Verde L., Peiris H. V., et al. 2003, ApJS, 148, 175
Stanek K. Z., Mateo M., Udalski A., et al. 1994, ApJ, 429, L73
Starobinsky A. A., 1982, Phys. Lett. B, 117, 175
Steidel C. C., Giavalisco M., Pettini M., et al. 1996, ApJ, 462, L17
Steidel C. C., Sargent W. L. W., 1992, ApJS, 80, 1
Stengler-Larrea E. A., Boksenberg A., Steidel C. C., et al. 1995, ApJ, 444, 64
Storrie-Lombardi L. J., Irwin M. J., McMahon R. G., 1996a, MNRAS, 282, 1330
Storrie-Lombardi L. J., McMahon R. G., Irwin M. J., 1996b, MNRAS, 283, L79
Swaters R. A., Madore B. F., Trewhella M., 2000, ApJ, 531, L107
Tisserand P., Le Guillou L., Afonso C., et al. 2007, A&A, 469, 387
Toomre A., Toomre J., 1972, ApJ, 178, 623
Tran H. D., Tsvetanov Z., Ford H. C., et al. 2001, AJ, 121, 2928
Tremaine S., Gunn J. E., 1979, Phys. Rev. Lett., 42, 407
Tremonti C. A., Heckman T. M., Kauffmann G., et al. 2004, ApJ, 613, 898
van den Bosch F. C., Pasquali A., Yang X., et al. 2008, arXiv:0805.0002

van der Marel R. P., Magorrian J., Carlberg R. G., et al. 2000, AJ, 119, 2038
van Dokkum P. G., Franx M., 1995, AJ, 110, 2027
van Dokkum P. G., Quadri R., Marchesini D., et al. 2006, ApJ, 638, L59
Van Waerbeke L., Mellier Y., Radovich M., et al. 2001, A&A, 374, 757
Verolme E. K., Cappellari M., Copin Y., et al. 2002, MNRAS, 335, 517
Wagoner R. V., Fowler W. A., Hoyle F., 1967, ApJ, 148, 3
Walcher C. J., van der Marel R. P., McLaughlin D., et al. 2005, ApJ, 618, 237
Wang Y., Yang X., Mo H. J., et al. 2008, ApJ, 687, 919
Warren M. S., Quinn P. J., Salmon J. K., Zurek W. H., 1992, ApJ, 399, 405
Wegner G., Colless M., Saglia R. P., et al. 1999, MNRAS, 305, 259
Weinberg D. H., Miralda-Escude J., Hernquist L., Katz N., 1997, ApJ, 490, 564
Weinberg S., 1971, ApJ, 168, 175
Weinmann S. M., van den Bosch F. C., Yang X., Mo H. J., 2006, MNRAS, 366, 2
Weymann R. J., Jannuzi B. T., Lu L., et al. 1998, ApJ, 506, 1
White S. D. M., 1978, MNRAS, 184, 185
White S. D. M., Davis M., Frenk C. S., 1984, MNRAS, 209, 27P
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
White S. D. M., Rees M. J., 1978, MNRAS, 183, 341
Whitelock P., Catchpole R., 1992, in Blitz L., ed., The Center, Bulge, and Disk of the Milky Way Vol. 180 of ASSL, . p. 103
Willman B., Blanton M. R., West A. A., et al. 2005, AJ, 129, 2692
Willmer C. N. A., Faber S. M., Koo D. C., et al. 2006, ApJ, 647, 853
Wirth G. D., Koo D. C., Kron R. G., 1994, ApJ, 435, L105
Wolf C., Meisenheimer K., Rix H.-W., et al. 2003, A&A, 401, 73
Yang X., Mo H. J., van den Bosch F. C., et al. 2005a, MNRAS, 362, 711
Yang X., Mo H. J., van den Bosch F. C., et al. 2007, ApJ, 671, 153
Yang X., Mo H. J., van den Bosch F. C., Jing Y. P., 2005b, MNRAS, 357, 608
Yanny B., Newberg H. J., Grebel E. K., et al. 2003, ApJ, 588, 824
Yoachim P., Dalcanton J. J., 2006, AJ, 131, 226
York D. G., Adelman J., Anderson Jr. J. E., et al. 2000, AJ, 120, 1579
Young J. S., Scoville N. Z., 1991, ARA&A, 29, 581
Zehavi I., Zheng Z., Weinberg D. H., et al. 2005, ApJ, 630, 1
Zeldovich Y. B., 1972, MNRAS, 160, 1P
Zhao H., Spergel D. N., Rich R. M., 1995, ApJ, 440, L13
Zheng Z., Shang Z., Su H., et al. 1999, AJ, 117, 2757
Zibetti S., White S. D. M., Brinkmann J., 2004, MNRAS, 347, 556