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 time scales 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 3 through 10 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 11–16 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 10, 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 11). 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 13). 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 10, 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 14, 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 galaxy

population. With modern, large telescopes we can now observe galaxies out to redshifts beyond six, making it 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.

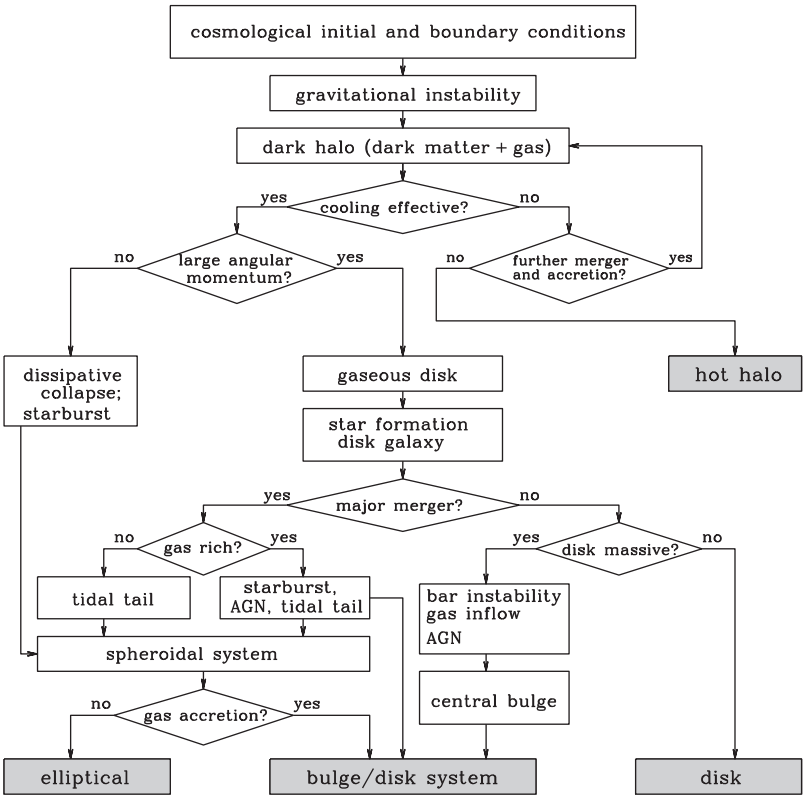

Fig. 1.1. A logic flow chart for galaxy formation. In the standard scenario, the initial and boundary conditions for galaxy formation are set by the cosmological framework. The paths leading to the formation of various galaxies are shown along with the relevant physical processes. Note, however, that 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.

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 3, 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 space-time 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[1] 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 3 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 §3.6, 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

[1] 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 §3.6, 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 4, 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 4, 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 over-dense region increases with time due to the overall expansion of the universe. Once the perturbation reaches over-density $\delta\rho/\rho \sim 1$, 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 nonlinear regime to the strongly nonlinear regime.

The outcome of the subsequent nonlinear, 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 5. Nonlinear, 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 7 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 §8.2.3 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 8 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 nonlinear 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 7, 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 11.

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 9), thus giving rise to a visible galaxy.

Unfortunately, many details of these processes are still unclear. In particular, we are still 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 10, 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 12. 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 §10.5). The radiation and the blast waves 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 §§8.6 and 10.5.

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 14). 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.

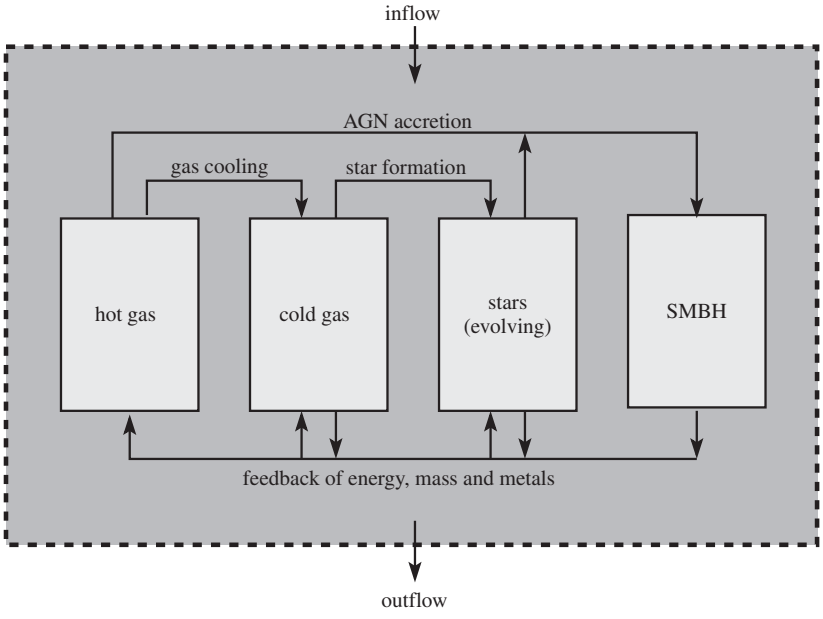

Fig. 1.2. A flow chart of the evolution of an individual galaxy. The galaxy is represented by the dashed box which contains hot gas, cold gas, stars and a supermassive black hole (SMBH). Gas cooling converts hot gas into cold gas, star formation converts cold gas into stars, and dying stars inject energy, 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.

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 blast waves 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