Supervisor: Dr. A. Babul

Abstract

The dynamical evolution of substructure within dark matter halos is of central importance in determining many aspects of galaxy formation and galaxy evolution in cold dark matter cosmologies. The overall sequence in which the different stellar components of galaxies are assembled, the survival of galactic disks, the number of dwarf satellites orbiting giant galaxies, and the nature of stellar material in galactic halos all depend on the dynamics of halo substructure. In this thesis, I develop an analytic description of the evolution of substructure within a dark matter halo, and use it to construct a semi-analytic model of the formation and evolution of disk galaxies.

Substructure within an individual halo is modelled as a set of distinct subhalos, orbiting in a smooth background. These subhalos evolve through three main processes: dynamical friction, tidal mass loss, and tidal heating. By including analytic descriptions of these three processes explicitly in a simple orbital integration scheme, it is possible to reproduce the results of high-resolution numerical simulations at a fraction of the computational expense. The properties of a subhalo can be estimated with an accuracy of 20%, until it has lost most of its mass or been disrupted.

Using this description of satellite dynamics, I construct a semi-analytic model for the the evolution of a galaxy or cluster halo. I show that this model reproduces the basic features of numerical simulations, and use it to investigate two major problems in current galaxy formation scenarios: the prediction of excessive substructure in galaxy halos, and the survival of galactic disks in halos filled with substructure.

I show that the small number of dwarf galaxies observed in the Local Group can be explained by considering the effects of reionisation on star formation in small halos. The stellar luminosities predicted in this case match the observed luminosities of local satellites. The predicted spatial distribution, sizes and characteristic velocities of dwarf galaxies are also consistent with those observed locally.

Many of these satellite galaxies are disrupted by tidal stripping or encounters. I investigate the properties of their debris, and show that its total mass and spatial distribution are similar to those of the stellar halo of the Milky Way. Furthermore, the stars in this debris are mainly old, satisfying another observational constraint on models of galaxy formation. Some satellites have been disrupted fairly recently, however, suggesting that coherent tidal streams may still be visible at the present day.

Finally, I investigate the effects of encounters on the central disk within the main halo. I find that the rate of disruptive encounters drops off sharply after the galaxy is assembled, such that the typical disk has remained undisturbed for the past 8–10 billion years. Less disruptive encounters are more common, and disks are often heated as they re-form after their last disruption, producing components like the thick disk of the Milky Way. These results may resolve the long-standing uncertainty about disk ages in hierarchical, cold dark matter cosmologies. It is less clear whether the bulge-to-disk mass ratios predicted by the model, for the currently favoured LCDM cosmology, are consistent with observations. The relative mass of the bulge in typical disk galaxies may place an upper limit on the age of their stellar contents.

The Formation and Survival of Disk Galaxies

by

James E. Taylor

University of Victoria

Table of Contents

List of Tables

List of Figures

Acknowledgements

This thesis combines the efforts and contributions of many people. The original idea for this project was suggested by my supervisor, Arif Babul. It is to his credit that it bore fruit in the end. I thank Hector Velázquez, Simon White, Eric Hayashi, Julio Navarro, Tom Quinn, Sebastiano Ghigna, and their collaborators for providing me with the numerical data required to calibrate my model, in chapters 2, 3, 4 and 6. Some of the material in part I will appear shortly in the Astrophysical Journal (Taylor & Babul 2001). I wish to thank Josh Barnes, Andrew Benson, Tsafrir Kolatt and Rosemary Wyse for their careful reading of a draft of that paper, and Ray Carlberg, Lucio Mayer, Julio Navarro, Jerry Ostriker, Joachim Stadel and Simon White for useful advice on many other technical points. I am also grateful to my supervisor, Arif Babul, and to the members of my committee, in particular Julio Navarro, for their academic, scientific and editorial guidance.

My studies were funded in part by the department of Physics and Astronomy, the School of Graduate Studies, and the President's Office of the University of Victoria, as well as by the Natural Science and Engineering Research Council of Canada. I am most grateful for their financial support.

Within the department, I would like to thank the other students for their help and good company over the years, and recognise Eric, Robert, Kathleen, John and Steven in particular for their invaluable advice on Latex, sm, and other computer problems. I also thank Stevenson Yang for his heroic efforts keeping the computers up, Dave Balam for initiating me into the secrets of the DAO, and the departmental staff, notably Geri, for their assistance with all matters administrative.

My deepest thanks go to my friends and family. To my roommates, particularly to Dáithí, for some timely reminders about partial derivatives, to Vanessa, for the Latex template that produced this document, and most of all to Hinrich, for my sanity, and a bit of exercise (both physical and intellectual). To my many other friends and allies in Victoria, notably John, Duane, Su, Margaret, Signe, Ann, Mélanie, Sandy, Paul, Katrin, and Virginie, and to Jean-Jacques and Chantal Lefebvre, whose warmth and hospitality were overwhelming.

Thanks to my parents, who engendered all of this, to my siblings, Pegatha, Kate, Sarah, and Andrew, who saw me through it, and to my grandmother, who had faith in me throughout. Finally, my work over the past year would not have been possible without the unfailing love and support given me by Nathalie-Anne Lefebvre. Without her, this thesis would have been much less than it is, and less than I wanted. I hope she will take some credit for the finished product.

– *For my Parents* –

Preface

A Cautionary Tale

My supervisor likes to tell the story of a great city, that was once struck by a terrifying earthquake. As the aftershocks died away, the survivors gathered in the streets, amid the rubble of their shattered dwellings, only to be overcome by a new terror; looking up at the night sky, they saw that it was divided in two by a great band of light, and concluded this must truly be the end of the world, for the cataclysm had split the sky itself. The punchline to this story is its setting – not, as one might imagine, four or five thousand of years ago, in a collection of mud huts somewhere in the cradle of civilization, nor even hundreds of years ago in a city of medieval Europe, but modern-day Los Angeles, after the earthquake of 1994. With the power out all over the city, its inhabitants saw, in most cases for the first time, what was once a familiar sight; the starry vault of a dark sky, neatly transected by the great rift of the Milky Way. Our ancestors would not, in fact, have made the same mistake. It is only recently that we have lost our familiarity with the heavens. Ironically, however, it is also only recently that we have gained a scientific understanding of our local environment in the universe, and that we have started to appreciate the nature of the great disk galaxy which is our home.

This thesis aims to explore the nature and origins of our Galaxy[1] , imagining it to be representative of the countless other galaxies we see throughout the universe. In particular, it attempts to explain the most prominent feature of the Milky Way, observed now for millennia; the flatness of the Galactic disk, which gives our Galaxy

[1] I will follow the convention of capitalising 'galaxy' when referring to our galaxy, the Milky Way.

its characteristic appearance of a stream of stars, dividing the sky in two. This flatness, while clearly observed, has seemed hard to explain naturally, in recent models of galaxy formation. This was, in fact, part of my original motivation to explore the question of our Galaxy's origins. More generally, I hoped to gain a better understanding of the physical universe that surrounds us, its layout and its history, and the processes which govern its evolution. I hoped, in particular, to appreciate more fully our local environment in the universe, which provided us with the Sun and the planets, the stars, and all the splendors of the night sky.

The original point in this preface was that the average person today is ignorant of the appearance of the night sky. Many people once thought the world was flat, and now the idea is a synonym of ignorance. Yet it has been 70 years since Kapteyn and Shapley, and then Hubble, established the scale of our Galaxy and of the universe beyond it, and still this knowledge does not seem to have reached the public at large. Even the astronomy students I teach in introductory courses routinely confuse the Galaxy with the universe, and have no sense of the scale of either, while in the process of explaining Astronomy to the public during open houses and guided tours, I have sometimes had to update them on the revolutionary ideas of the 16th century!

The purpose of this preface is to address this ignorance, since there is no point in presenting the dynamical history of our Galaxy to readers unfamiliar with the term 'galaxy' itself. In the next few pages I hope to introduce the lay reader to our current understanding of the universe; its layout, its history and its workings. While the details of the thesis may remain obscure to those outside the field, and probably won't interest them much anyway, I hope that reading this initial section will at least give them a foothold in this subject, and provide basic vocabulary with which to tackle the rest of this work. At the end of the Preface I also provide an overview of the problem I considered in the thesis, and the methods I used to address it. I then refer the casual or confused reader to the final section of the thesis, where I summarise its more technical points, and conclude on the structure and evolution of our Galaxy.

Readers with a background in Astronomy will already be familiar with the material in this preface, and may wish to proceed directly to the introduction instead.

The Geography of the Universe

The material contents of the universe, beyond the confines of our planet, are organised into structures on a set of progressively larger scales. These structures have been discovered in succession, historically, prompting the development of Mathematics and Physics in the process. Unfortunately, only the first few rungs of this cosmic ladder are familiar to the lay reader. Since this thesis concerns some of the largest structures in the universe, I will describe each of the intervening scales below.

Material in the immediate vicinity of our planet forms the **solar system**, so named because it revolves around the Sun, as was first posited by Copernicus. Most of the transient phenomena observed in the night sky occur within the solar system, which houses the Moon, the planets and their moons, the asteroid belt, comets and various other minor bodies, along with less easily seen distributions of dust, radiation and magnetic fields. The Sun is the central organising element of this system; all the other material in the solar system is thought to be debris left over from its formation, 4.5 billion years ago. The Sun, the Moon and the planets are all distributed in roughly the same plane in space. As a result, seen in projection on the sky, they follow a single arc (called the *ecliptic*) to within ten or twenty degrees. As we shall see below, this flat spatial distribution suggests something about how the solar system formed.

The scales of these astronomical structures are of course, astronomical, but I will mention them here to give the reader a point of comparison for the even larger structures discussed later. The Earth is roughly 13,000 km in diameter, and at a mean density of 5.4 g/cm^3 weighs 6×10^{27} – that is, 6,000,000,000,000,000,000,000,000,000 – grams. It takes light .13 seconds to circle the Earth, while a car moving at 100 km/hr would take 17 days to do so. The Sun is about 100 times bigger than the Earth, and slightly less dense ($1.4\,g/cm^3$, a little denser than water), giving it a mass of 2×10^{33} grams, or 300,000 times the mass of the Earth. The mass of the Sun

defines a convenient unit with which to measure all masses in Astronomy; the **solar mass**, denoted $M_\odot$. Light takes 15 seconds to travel around the Sun, while our driver would take 5 years to do so. While the Sun is much larger than the Earth, the Earth's orbit is in turn much larger than the Sun – the mean radius of the orbit is 200 times bigger than the radius of the Sun, corresponding to 8 minutes of light travel time, or 170 years of driving time. This scale is in fact defined as a unit of length, the **astronomical unit**, denoted A.U. . The Earth's average speed in this orbit is about 30 km/s. Pluto's orbit has a maximum diameter of 25 A.U., while the outer reaches of the solar system lie at 100 A.U. from the Sun. Thus it would take our intrepid driver 17,000 years to reach the outer edge of the solar system, while light travels this distance in 14 hours. Even this local part of the universe is very, very large.

Beyond the solar system lie other stars (and other planetary systems). The stars have been known to lie outside the solar system for over 300 years, and over the course of the 19th century were established to be balls of ionised plasma like the Sun, though varying in size and mass, temperature and in chemical composition. Stars in the neighbourhood of the Sun typically lie a few light years from one another (a **light year** is a unit of distance, equal to the distance light travels in a year); for technical reasons, distances on this scale are measured in units called **parsecs** or pc, one parsec being equal to 206,250 A.U., or 3.26 **light years**. Here again, we note that space is large and very empty, relative to things on Earth – the average distance between stars is 2,000 times the size of our solar system.

There is one last scale on which the structure of our environment is apparent to the unaided observer (albeit the unaided observer living away from artificial lighting). While the brighter stars in the sky are distributed more or less uniformly, there is a fainter band of light which divides the sky in two. The ancient Greeks named this band the **Milky Way** (*galactos*), claiming it was milk spilt from the breast of the goddess Hera, while she was feeding the infant Hercules. With the earliest observations by telescope, Galileo determined that the Milky Way consisted of the combined light

of thousands of apparently faint stars, which might be normal stars lying at great distances from us. The conclusion of the hypothesis is worth considering carefully. If most stars have roughly the same brightness, and the bright stars are distributed uniformly across the sky, while the fainter stars are concentrated in a single band of light, then we must be living within a flattened distribution, a disk or plane, whose edge is close enough to see (figure 1).

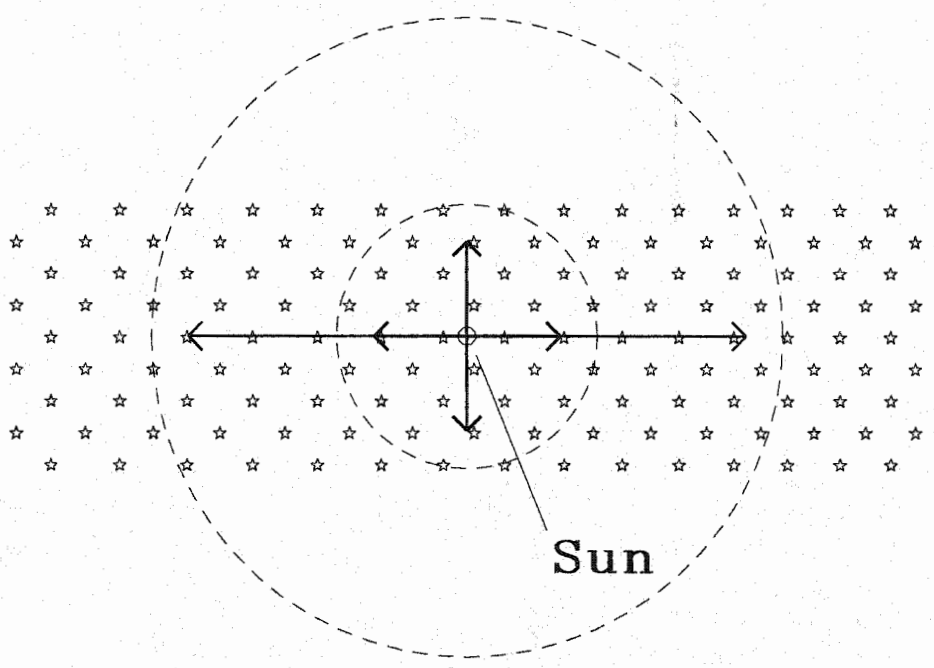

Figure 1 A schematic representation of the distribution of stars in the neighbourhood of the Sun. Nearby stars are distributed uniformly around the sky (smaller circle), while more distant stars are concentrated in particular directions (larger circle).

I will not discuss here the long and contentious debate that took place in the 1920s, over how the stellar contents of the Galaxy were distributed in space, but say simply that it was established that we live in the middle of a great disk of gas and stars, slightly offset from the centre of our Galaxy. The **Galactic Disk** is about 5,000 light years, or 1,500 pc, in thickness, and about 100,000 light years, or 30,000 pc, in diameter. Thus the axis ratio of the disk is 20:1, which is why it is described as

'thin'. The number of stars in the Galaxy is a few times 10^{10}, but the total mass of the Galaxy is much larger than one might expect from this – 2×10^{12} $M_\odot$ or so – due to the mysterious 'dark' matter. The Milky Way also has several other stellar components – a central **Bulge**, and a tenuous **Stellar Halo** (also known as the **Corona** or the **Spheroid**) distributed spherically around the disk, in particular. The disk rotates at roughly 200 km/s, while the other components have little net rotation. While our view of the Milky Way is partly obscured by dust in the plane of the disk, it no doubt resembles other external disk galaxies like the one shown in figure 2.

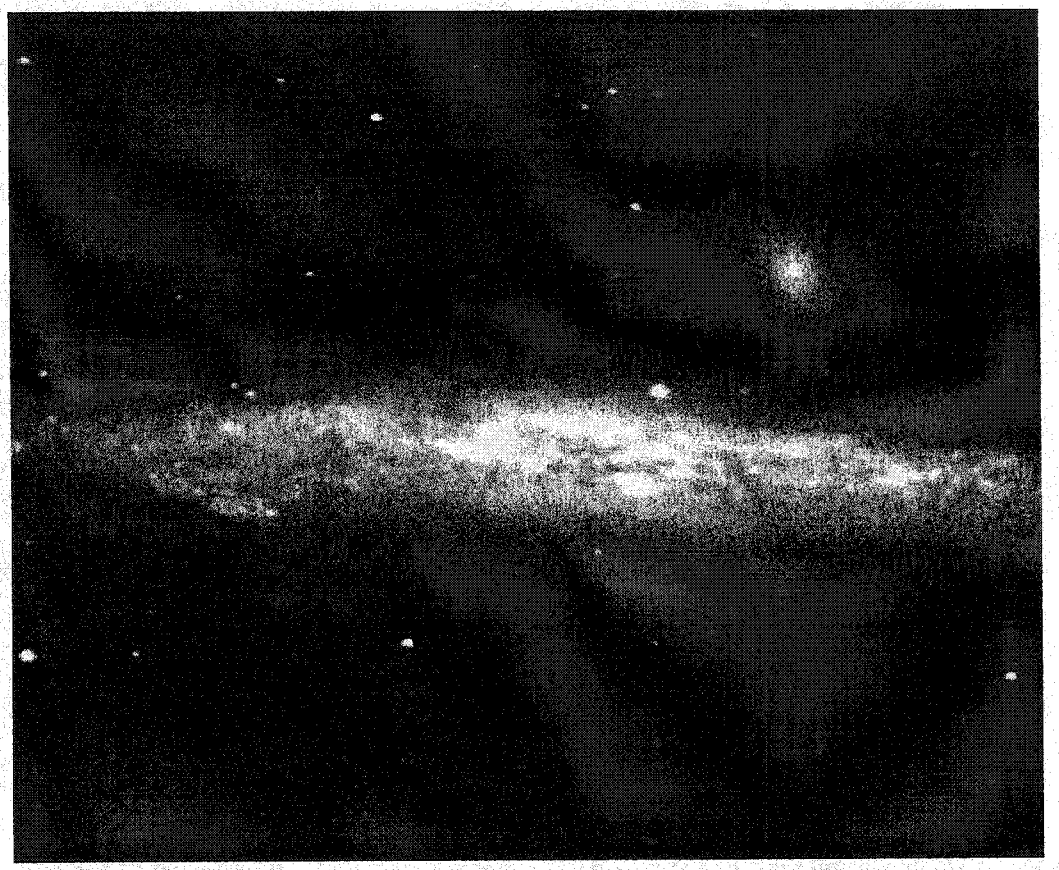

Figure 2 Arp 281, a nearby disk galaxy, similar to the Milky Way. Note the disk, the central bulge, and the small satellite galaxy (above and to the right). Image taken by the author and D. Balam with the 1.88-m Plaskett telescope at the Dominion Astrophysical Observatory.

What of the universe beyond our Galaxy? Until the early part of this century, our Galaxy was often thought to constitute most of the universe, although some inspired minds, including Immanuel Kant, had supposed otherwise. Only in the 1930s, with the work of Hubble, were the nebulae, thought to be clouds of gas within our own Galaxy, demonstrated to be separate galaxies, lying at great distances from ours. Just as gravity organises the solar system and the Milky Way, it organises the Milky Way's nearest galactic neighbours into a structure called the **Local Group**. The Local Group includes three large spiral galaxies, the **Andromeda galaxy**, (also known as 'M 31'), the most massive of the three, the Milky Way, and a smaller galaxy known as M 33. Each of these large galaxies has several smaller companions, while a few other independent galaxies also form part of the group. The current membership includes about 40 galaxies, ranging in mass from $10^6\,M_\odot$ to $10^{12}\,M_\odot$. In total, the Local Group is thought to measure 1 Mpc (one million parsecs) in size, and to weigh $\simeq 3 \times 10^{12}\,M_\odot$. Typical velocities within the Local Group are several hundred km/s, comparable to the rotation speed of the disk.

And beyond this? The Local Group is only one of many associations amongst the billions of galaxies we can see throughout the universe, with the aid of telescopes. Sometimes galaxies group together in much larger numbers, forming **(galaxy) clusters**, with hundreds of large members and thousands of smaller members, measuring several Mpc across, and weighing 10^{14}–$10^{15}\,M_\odot$ – hundreds or thousands of times the mass of the Milky Way. Sometimes, at the other extreme, galaxies appear to be isolated, though it may be that they have small companions we have not detected. Individual galaxies, and groups and clusters of galaxies, tend to be associated in larger structures which, while they are not gravitationally bound, show up distinctly in large-scale surveys. This **large-scale structure**, on scales of tens or hundreds of Mpc, represents the biggest inhomogeneities in the universe. On all scales larger than these, right out to the light horizon, approximately 13 billion light years (4 billion pc) from us, the universe is relatively smooth and uniform.

The History of the Universe

As the different scales which characterise the universe were discovered progressively, from the smallest to the largest, our sense of the history of the universe became clearer. We know it now as the history of structure formation, the formation of atoms and nuclei; of gas clouds, planets and stars; of galaxies, groups and clusters of galaxies; and finally of the largest structures. I will outline our current knowledge of these processes in the following section, and return to some of the underlying physics in the next.

The galaxies and groups of galaxies described previously do not float motionless in intergalactic space, nor do they move about at random like molecules in a gas; their motion is overwhelmingly away from us, with an average velocity proportional to their distance. This was the great discovery of Hubble in the 1930s. Assuming that our vantage point in the universe is typical (an assumption called the **Cosmological Principle**), this implies that everything in the universe is moving away from everything else, or that the universe is in uniform expansion. By implication, the material contents of the universe used to be much more tightly compressed, so much so that the temperature and density at any place in the universe would once have been greater than those in the heart of the Sun. The expansion from this early, hot, dense phase is referred to as the **Big Bang**, although analogies with an explosion are generally misleading, as the expansion of the universe is uniform and has no centre.

Since light travels at a finite speed, looking out to some distance corresponds to looking back to some early epoch, and in this sense we know that the Big Bang occurred, because if we look far enough away, we can see it going on. In fact, this statement is not strictly true; at early times, the material in the universe was so hot and dense that it was an ionised plasma, opaque to light, just as the Sun is today. With the expansion, its temperature and density dropped to the point where electrons combined with atomic nuclei to form electrically neutral atoms, and the universe

became transparent to light, about 100,000 years after the Big Bang[2] . When we look to large distances, we can see up to the start of this era, but no further. The light from earlier times, shifted in frequency by the expansion of the universe, forms the **Cosmic Microwave Background** (CMB).

At the start of this new era, the universe was very nearly homogeneous. Local fluctuations in density around the mean value, observed as changes in the brightness of the CMB, were less than 1 part in 100,000. Since this time, the effects of gravity have made the universe much less homogeneous on small scales; a typical galactic disk is 100,000 times denser than the average density of the universe. While it is hard to measure the distribution of fluctuations in the CMB on all scales, theoretical models suggest that the (spatially) small fluctuations should have been more extreme than the large ones. In turn, gravity would have acted faster on these regions, causing positive fluctuations in the local density to acquire more material and become even denser. Eventually, this process made certain regions sufficiently dense that rather than expand with the rest of the universe, they recollapsed in on themselves. The continued collapse of such regions was then halted only when star formation or some other process injected enough energy or angular momentum into them to support their mass against further collapse. The early universe should thus have contained many small, condensed regions where the first stars were forming, perhaps a few million years after the Big Bang.

At later times, larger, less extreme fluctuations would start to collapse gravitationally. These objects would already contain many small star-forming regions within their volume, however, so this collapse would be far from smooth, but involve many violent mergers between dense lumps of material. It is at this stage that we expect galaxies to have formed, first on small scales, producing objects like the dwarf

[2]Later in the history of the universe, high-energy radiation from hot stars and active galaxies **re-ionised** many of these atoms, such that intergalactic gas in the present-day universe consists mainly of charged particles.

galaxies of the Local Group, and then on larger scales, producing objects like the Milky Way. Finally, galaxies would have been attracted together to form the largest bound objects, groups and clusters of galaxies. This process of collapse and merger on progressively larger scales is known as **hierarchical structure formation**.

It is not quite clear where to place the formation of the different components of our Galaxy in this sequence. The disk of the Galaxy continues to form stars very actively to the present day, and contains stars of all ages, going back 12 billion years, or four-fifths of the age of the universe. The other components, the **bulge** and the corona (or **stellar halo**), appear to be made up of old stars, as old as any we see in the universe. One of the goals of galaxy formation models is to explain these observations. Closer to home, our Sun, with an age of 4.5 billion years, would have formed when the universe was two-thirds of its present age, while the planets formed shortly thereafter. Life on Earth, in all its wonderful variety, is more recent still, only two to three billion years old.

The Physics of the Universe

How then was all this accomplished? While the physical processes that formed atoms, stars and galaxies cannot always be studied in the lab, they have been partially determined over the past century through theoretical modelling and the detailed observation of the universe, greatly aided by work in related fields of physical science.

On the largest scales, the behaviour of the universe is relatively simple, and can be described by a few quantities, known as the **cosmological parameters**, which determine its age, size, and expansion rate. These parameters include the **Hubble constant**, $\mathbf{H_0}$, which describes the relative expansion rate of the universe, or equivalently the time that has elapsed since the Big Bang, the **density parameter**, $\mathbf{\Omega}$, the density of the universe relative to the critical density needed to stop its expansion, and the **cosmological constant**, $\mathbf{\Lambda_0}$ (sometimes denoted $\mathbf{\Omega_\Lambda}$), which measures the response of empty space to the expansion. The global behaviour depends on the

values of these parameters; in particular, if $\Omega > 1$ and $\Lambda_0 = 0$, the universe will eventually recollapse, whereas if $\Omega < 1$ and $\Lambda_0 = 0$, or if $\Lambda_0 > 0$, the universe will expand forever. One of the original goals of modern cosmology was to attempt to determine the values of these global parameters; indeed, much of contemporary galaxy formation theory was developed as part of this endeavour. While galaxy formation models may still help determine the values of the Hubble constant, the density parameter and the cosmological constant, other methods such as observation of the CMB have now superseded them in this task. Galaxy formation does play a crucial role, however, in the understanding of the material contents of the universe.

The density parameter Ω is known from studies of the CMB, galaxy clusters and several other observations to be approximately 0.3 ± 0.1; that is the universe has only three-tenths of the density it would need to stop its continued expansion. The theory of **nucleosynthesis**, the production of atomic nuclei and by extension the different chemical elements in the Big Bang, predicts that the density of normal matter in the universe should be around 0.1 or less (this material is called **baryonic**, baryons being protons and neutrons, the massive particles which make up atomic nuclei). Thus, two-thirds of the matter in the universe must be in some other, unknown form. Where is this stuff? In fact we have good evidence that it surrounds and permeates galaxies and clusters of galaxies. The Milky Way, for instance, spins so fast (roughly 200 km/s at the position of the Sun) that the outer parts of its disk would fly apart if not held in place by much more mass than can be accounted for in the visible stars. The same is true of galaxy clusters – the ratio of the mass needed to keep them gravitationally bound to the light we receive from them is roughly 100 times greater than the mass-to-light ratio of the Sun. Thus, since clusters do contain normal stars as well, most of their mass must be 'dark', in the sense that it does not interact with light and is effectively invisible. Determining the nature of the **dark matter**, first discovered by Fritz Zwicky in the 1940s, remains one of the greatest problems facing modern astrophysics.

How does one go about studying material which is, by definition, invisible? Dark matter has at least one definite property, and that is mass. Like normal matter, dark matter in the early universe would have been subject to gravitational instability: regions with slightly more material initially would have exerted a stronger gravitational pull on their surroundings, attracting more material and becoming yet more dense. The conclusion of the process would have been the collapse of regions into dense, gravitationally bound lumps called **halos**. Since normal matter and dark matter do not interact strongly except through gravity, the two would initially have been mixed through these regions. Normal matter does interact with light, however, and can thereby radiate away, or **dissipate**, energy. This process is also referred to as 'cooling', since it corresponds to the cooling of a gas in the conventional sense, that is the expulsion of microscopic kinetic energy by radiation. Once confined to a dense region, where collisions between atoms were common and the consequent radiation and dissipation of kinetic energy was possible, the normal material would have collapsed further than the dark matter, forming a concentrated residue sitting in the centre of each dark matter halo. The final distribution of baryons would be determined by two processes. The first is the conservation of angular momentum, which cannot be lost through conventional radiation and would force collapsing material to form a rapidly rotating disk if it had any initial spin at all. The second is star formation, which would become an important source of energy once the baryons reached the densities typical of our Galactic disk.

Within this dense star-forming object, or proto-galaxy, the first generation of stars would be born from the densest regions of gas, would evolve – burning through nuclear reactions which transmute hydrogen to helium, and helium to heavier elements (referred to collectively as **metals** in astrophysics, in contrast with the conventional definition) – and would die, leaving slow burning remnants in the case of low-mass stars, or exploding as **supernovae** in the case of high-mass stars. These supernovae, occurring a few million years after the first stars formed, would have expelled their

metal content into the surrounding interstellar medium, changing its chemical composition permanently.

From considering these basic processes, one gets a rough picture of how galaxies first formed. (Although clearly the whole process would have been much more complicated than the early stages of structure formation, when gravity was the only important force.) Within halos of dark matter, gaseous baryons cooled and collapsed into centrally concentrated objects dense enough to form the first generation of stars. The most massive of these stars evolved quickly and exploded as supernovae after only a few million years, reinjecting enough energy into the gas to halt further collapse. Depending on the initial state of the baryons, these early star-forming regions may have been rotating disks, but subsequent merging with other halos containing other star-forming regions would have disrupted this structure. Disks could then have reformed as more gas fell into the halo and lost its energy, while preserving its angular momentum, possibly to be disrupted again by subsequent mergers. Understanding, from this rather complex and muddled picture of galaxy formation, how our Galaxy came to look as it does today is the goal of this thesis.

A Summary of the Thesis

At the beginning of this preface, I reminded the reader of the one most easily observed properties of our Galaxy, namely its thinness. The axis ratio of about 20:1 that characterises the disk of our Galaxy is mirrored by a corresponding feature in the motions of disk stars: the average random velocity of stars in and out of the plane of the disk is about 20 km/s, while their orbital motion carries them around the Galaxy at 200 km/s, or ten times as fast. Since the kinetic energy associated with motion is proportional to the square of the velocity, this means that the average star has 100 times more energy associated with its orbital motion than with its random motion in and out of the plane of the disk. If one likens the motions of stars to the motions of gas molecules, random motion is analogous to heat; in this sense, the disk is **cold** - that is its stellar motions are highly organised, rather than random.

A cold, highly organised disk is the natural product of dissipation and the conservation of angular momentum, if the disk forms from the collapse of a single ball of gas with some initial spin. Yet the clumpiness of dark matter on small scales implies that galaxy formation was nothing like this. Galaxies formed from lumps, which formed from smaller lumps, and so on, and a galaxy the size of the Milky Way would have had many progenitors in this process, each with their own stars and distinct structure. One might reasonably expect the outcome of such a chaotic history to look like a mess. Indeed, some galaxies, a minority of about 10%, are very mixed up – these are the **elliptical** galaxies. But **spiral galaxies**, like the Milky Way and most of the isolated galaxies we see in the sky, are not messy but highly structured.

This problem can be rephrased in a slightly different way. The Milky Way formed out of the material occupying a certain volume of the universe. Models of the early distribution of dark matter predict that this region would already have contained many smaller halos of collapsed material, either entirely dark, or containing dwarf galaxies. If the dense central cores of these halos survived disruption as the Milky Way formed, then they would still be orbiting our galaxy, presenting a hazard to the stability of the Galactic disk. The thinness and age of the disk of the Milky Way places limits on the 'lumpiness' of the material from which it formed, as any large lumps would have stirred it up, scattering its stars out of the plane of the disk. It is not immediately clear whether these limits are consistent with the lumpiness predicted by current cosmological theories.

To attack this problem requires a model of galaxy formation that covers a range of scales, from the thickness of the disk to the diameter of the Local Group; that describes both the conditions of the background universe and those in the main part of our Galaxy; and that can be run many times, to determine how likely it is to reproduce the specific properties of our Galaxy in a large set of realisations. Existing models of galaxy formation tend to be largely **numerical**, that is they generate complex results using many simple calculations, treating galaxies as clouds of point

masses which interact only through gravity. While conceptually simple, and robust in their predictions, such models are very expensive to run computationally, requiring weeks or months of computer time to generate a single set of results. An alternate approach to studying astrophysical problems is to do the calculation with pen and paper; this is called the **analytic method**. This has the advantage of speed, but can only be used to solve the simplest problems, typically those with a very simple spatial symmetry. My strategy, in this thesis, has been to use the bastard child of these two approaches, a **semi-analytic method**, which mixes pen-and-paper analytic results with a certain amount of raw numerical computation. The result is a relatively complex model, but one that runs 5,000–10,000 times faster than purely numerical methods. While semi-analytic models are only approximate, and must be checked carefully against numerical results when possible, they provide the speed required to study the statistics of galaxy formation, for many different scenarios and choices of parameters.

In the first part of this thesis, I outline an analytic description of minor mergers between lumps in a larger halo, that forms the basis for my model. In the second part, I then describe how to combine this analytic description with various other components to produce a complete model of galaxy formation, and then use this model to investigate the problem of disk survival, and several related problems. The details of these two sections are fairly technical, but the conclusion offers a summary of this work which I hope will be more accessible to the lay reader.

Chapter 1

Introduction

Astronomy is an effort to understand both the contents of the universe, that is, the organised structures we see within it, and the workings of the universe, that is, the processes by which these structures formed. This endeavour started long ago, with the first attempts to explain the mysterious regularities of the Sun, the Moon and the planets, and has inspired many developments in mathematics and physics over the centuries. If we pursue it to this day, it is partly in the hope that it will continue to teach us new mathematics and new physics, but also, perhaps, because we sense we are close to the end of a chapter in the book. We can now observe most of the volume of the universe that will ever be visible to us, and we have instruments capable of detecting most forms of electromagnetic radiation, and even gravitational radiation and massive particles from space; that is to say, we can detect most of the information we will ever receive from it. The challenge now is to begin to understand the universe we observe by these means.

Observations show us organised structure on many scales in the universe, from the sub-atomic to the sub-horizon, material arranged by the interplay of particles and forces. The largest structures, on scales where gravity is the dominant force, are simplest to understand, and it is here that the current picture of cosmogenesis seems most reliable. Moving to smaller scales, structure becomes easier to observe, but harder to explain, as electromagnetism and the other forces start to play an important role. Galaxies, in particular, sit at the threshold of this new regime, which is why understanding galaxy formation is one of the main stumbling blocks in current astrophysics. Yet galaxies provide a vital connection between the global laws that govern the universe as a whole, and the detailed appearance of our immediate surroundings.

This thesis, which explores galaxy formation, is one small part of a broader attempt to bridge the gap between local observations and universal laws, relating the detailed structure of our Galaxy to the formation of galaxies and larger-scale structure at a more general level.

1.1 Hierarchical Galaxy Formation in a Universe Dominated by Cold Dark Matter

Our current picture of structure formation is based on a few key elements. First, several independent lines of argument – the theoretical behaviour of space and matter predicted by General Relativity, the observed expansion of the universe around us, and the existence of a cold, uniform background radiation, the Cosmic Microwave Background (CMB) – all imply that the universe was once much hotter and denser than it is now, and that the moment corresponding to the origin of the expansion, the Big Bang, occurred at some finite time in the past. While the initial moments cannot be observed directly, we can see back to the moment when the universe became electrically neutral and transparent to light, the epoch of recombination, and observe its state at the time by looking at the CMB. Several features of the CMB are striking; notably, it is uniform to one part in 10^5 (Smoot *et al.* 1991) on large scales, indicating that the universe as a whole is extremely uniform within its current horizon. Equally important, however, is the fact that it does show some small variations, whose amplitude varies with scale in a characteristic way. These fluctuations provide the initial conditions required for the subsequent growth of structure.

A second essential element of current structure formation models is dark matter. The dynamics of galaxies and clusters, which probe the distribution of mass in the universe on the largest scales, are inconsistent with the mass distributions we would infer from their light alone. While gas in various phases, or very low-mass stars, can

account for some of the missing mass, the discrepancy between the estimated and required masses is huge, typically a factor of ten or more. The one other probe of the gravitational potential available on these scales, gravitational lensing, or the bending of light as mass curves space-time around it, confirms these estimates. Whatever accounts for the missing mass, it must by definition be 'dark', that is it cannot interact with electromagnetism, since it has to account for a deficit in mass per unit light. We have another constraint on its properties: the reaction networks which predict the abundances of elements forged in the latter stages of the Big Bang are extremely sensitive to the overall density of the universe in protons and neutrons, the baryons that account for most of the mass of normal matter. Nucleosynthesis limits the baryonic content of the universe to about a third of the total density due to matter, and since much of this can be observed directly in the form of stars and gas, most dark matter must be non-baryonic (Krauss 2001). Finally, since dark matter dominates the mass density of the universe, its dynamics at early times will strongly affect structure formation. Dark matter particles which were still relativistic in the early universe could stream out of dense regions, erasing density fluctuations until late times. In this 'hot dark matter' (HDM) picture, the only surviving structure after this era would be on very large scales, and clusters and galaxies would have to form through the fragmentation of these larger objects, producing numbers and distributions of structure quite different from those observed (White, Frenk & Davis 1983). Current models therefore favour 'warm' or 'cold' dark matter (WDM and CDM respectively), particles that became non-relativistic earlier on in the expansion of the universe, and allowed small fluctuations to survive.

The initial fluctuations seen in the CMB grow in amplitude through gravitational instability, and eventually collapse, becoming gravitationally bound. In a universe dominated by CDM, fluctuations on small scales have larger initial amplitudes on average, and therefore collapse first. Any collapsing volume is therefore likely to contain smaller regions which have already collapsed, and structure grows larger with

time through successive mergers between these regions. Within these regions, gas becomes dense enough to cool by radiation, collapsing further and forming stars. This process is normally called hierarchical structure formation. While the details of the process become very uncertain once stars start to form, the first objects likely to have formed stars in the early universe would probably have contained about a million solar masses of material. A large galaxy such as the Milky, with a total mass of $10^{12}\,M_{\odot}$, would have formed from the merger of a whole spectrum of smaller objects, including hundreds or thousands of these smallest star-forming halos. Mergers between collapsed regions of dark matter, commonly called 'halos', by analogy with the dark matter distributions surrounding present-day galaxies, are only one part of the galaxy formation process. After their parent halos have merged, the baryonic contents rearrange themselves, dissipating energy and forming condensed central regions, the galaxies observed today. Even ignoring the complexities of the dissipative processes involved, however, some parts of this picture seem inconsistent with observations.

Most large galaxies – roughly 80% of them – have disks (Binggeli, Sandage & Tammann 1988), and most disks are fairly thin (van der Kruit & Searle 1982; Shaw & Gilmore 1990). The disk of the Milky Way, for instance, has an axis ratio of roughly 20:1 (Bahcall & Soneira 1980). It is also fairly old – the oldest parts of the thin disk are at least two-thirds the age of the universe (Carraro *et al.* 2001). Yet old, flat disks are vulnerable to heating and disruption. The disk of the Milky Way has roughly 100 times as much energy in its in-plane rotation as in its vertical (out-of-plane) motion. The addition of energy from an object with even 1% of the disk's mass could change this balance, heating the disk appreciably. If the Milky Way formed through a series of mergers, how has its thin disk survived continual disruption for so long? Hierarchical galaxy formation in CDM cosmologies leads to other problems as well. The density profiles of CDM halos have been studied extensively (NFW 1996, 1997; Moore *et al.* 1998), and are generically predicted to have high-density cusps at their centres. These cusps survive mergers as distinct structures, leading to

a general lumpiness of CDM halos. Even if these lumps do not collide directly with the disk of a central galaxy, they may heat it indirectly, by causing fluctuations in the potential of the halo. These lumps can also exchange angular momentum with the disk; in simulations of disk formation, this process leads to the disk losing most of its angular momentum, and collapsing down to a fraction of the size of observed disks with similar rotation velocities (Steinmetz & Navarro 1999). If the halo of our galaxy is full of lumps of dark matter, it also raises another question: why do we see no evidence for associated stars? The halo of the Milky Way does contain 11 other luminous objects, dwarf galaxies ranging in size from 1% to 0.001% of the mass of our Galaxy (Mateo 1998). The predicted number of dark subhalos down to the same mass limits is on the order of 1000, however. Did only a handful of these objects form galaxies, or is the CDM prediction incorrect?

There are alternatives to cold dark matter cosmologies. Dark matter could be slightly 'warmer', or have some self-interaction (Spergel & Steinhardt 2001; Hannestad & Scherrer 2000), the power spectrum of initial density fluctuations could be truncated on small scales by an additional mechanism (Gramann & Hütsi 2000), dark matter could decay with time (Cen 2001), or gravity itself could work differently on large scales (Bekenstein & Milgrom 1984; Milgrom 1986). To distinguish between these radical possibilities and more mundane solutions to the apparent problems of hierarchical galaxy formation, one requires a model of galaxy formation that can connect the assumed properties of matter to observable features of galaxies, groups and clusters. I will discuss two main classes of such models in the next section.

1.2 Current Models of Galaxy Formation

1.2.1 Numerical Models

The evolution of matter under the effects of gravity can only be described analytically for systems with very simple geometry: two isolated point masses, for instance, or a single, spherically symmetric distribution of material. To follow the formation of more general structures through the effects of gravity, astrophysicist have used numerical models since the time of the earliest computers. The simplest of these numerical models, N-body codes, use sets of point particles to represent continuous distributions of matter, and calculate the gravitational interactions between these particles to determine the evolution of the system. The number of particles, the number of timesteps, and the accuracy of the force calculations determine the overall accuracy with which a model reproduces the evolution of a given physical system, although many subtleties exist in representing continuous distributions with discrete particles, choosing appropriate timesteps and estimating forces to sufficiency accuracy without incurring huge computational expense.

Early N-body code calculated all the inter-particle forces directly, requiring N^2 calculations to advance N particles over a single timestep. More sophisticated codes developed subsequently calculated forces on a grid of sample points, or used similar algorithms to separate the force calculations from the underlying representation of the matter distribution, reducing the order of the calculation to roughly $N \log(N)$ per timestep. Numerical codes have also evolved to take advantage of technological advances, in particular the recent development of large parallel machines. The detailed implementation of a code, the accuracy of the force calculations, and the timestepping criterion all affect the actual run time of a given simulation. Nonetheless, the number of particles used in an N-body or related simulation remains indicative of its overall computational expense, so I will use this as a point of comparison in the discussion

that follows. Typical simulations with $\simeq 10^5$–10^6 of particles are currently possible on desktop machines, while the largest supercomputer simulations have 10^8–10^9 particles.

In the simplest cases, simulations by N-body or similar techniques are a fairly accurate representation of collisionless dynamics, approximate only in the sense that they are often used to represent systems with far more individual 'particles'. Thus large galaxies, for instance, contain 10^{10}–10^{11} stars, while the underlying dark matter distribution may consist of huge numbers of microscopic particles. This should not be a problem provided the behaviour of the system is insensitive to the exact number of particles used. Typically, this sets a lower limit on the number of particles required to model a given system accurately. Studies of minor mergers, for instance, have found that mass loss from the tidal disruption of a satellite can be determined accurately using $\simeq$ 1000 particles (Klypin *et al.* 1999a; Velázquez & White 1999; Hayashi & Navarro 2001). Similar studies of dynamical friction, the drag force exerted on a satellite by the wake it produces in the halo, indicate that the background halo must be resolved at a comparable level to produce accurate results. Thus, in concrete terms, to study the evolution of satellites orbiting in a galactic halo, down to the size of the smallest satellites of the Milky Way, would require roughly 1000 particles for each satellite, or one million particles in total, and a few million particles for the background halo, whose total mass is several times larger. Modelling such a halo in its surrounding cosmological context would require on the order 10^7 particles in total. Modelling galactic disks is comparably difficult; not only are these structures extremely thin, but their fragility makes them very sensitive to spurious numerical heating, if they are represented by too few particles. Thus an estimated 10^6–10^7 particle simulation is required to produce a disk model which remains stable for a Hubble time (Quinn, Hernquist & Fullagar 1993). While simulations of this size are becoming more common, and should be trivial to run on small computers in another five or ten years, their computational expense at the present time, corresponding to

weeks or months of CPU time on a workstation, prohibits routine use.

The limitations are even more restrictive if one adds the expense of hydrodynamical calculations to these estimates. The basic evolution of gas is better described by numerical models which represent the underlying matter distribution by cells with some finite size. These smoothed-particle hydrodynamics (or SPH) techniques typically demand a computational effort exceeding that of N-body or similar codes. Including even more complex physics, such as energy injection from star formation and supernovae, radiative transfer and magnetic fields, which are required to model the densest regions of galaxies self-consistently, will remain prohibitively expensive in the near future. The addition of these physical processes to simulations is currently only possible through approximations which detract somewhat from the simplicity and robustness of the numerical approach. Overall, numerical simulations provide the highest-quality and most detailed galaxy formation models currently available, but do so at considerable expense.

1.2.2 Semi-analytic Models

Given the computational expense of purely numerical models of galaxy formation (and the added complexity of numerical techniques which include dissipative physics), it seems worthwhile exploring alternative approaches to modelling galaxy formation. While there is a long legacy of important analytic work on galaxy formation (e.g. Gunn & Gott 1972; Gunn 1977; Rees & Ostriker 1977; Silk 1977; White & Rees 1978; Fall & Efstathiou 1980; White & Frenk 1991), the general problem of determining, say, the spatial distribution of material in a galaxy over the course of its evolution, is too complicated to tackle with any single analytic formalism. A recent approach, developed by several groups, has been to tie together analytic results with numerical calculations, producing hybrid, *semi-analytic* models. These models are necessarily approximate, typically containing many free parameters that need to be determined

by comparison with observations, but they offer a tremendous advantage in speed over purely numerical models.

Existing semi-analytic models of galaxy use a fundamental result in the statistics of structure formation to determine the sequence of mergers by which halos build up, in contrast numerical simulations, which simply start from initial conditions, and follow the growth of structure directly. This result is from the work of Press and Schechter (1974), who pointed out a way of calculating the number of collapsed objects as a function of mass and of redshift, based on a few simple assumptions. Their approximate result was confirmed by more rigorous work subsequently (Bower 1991; Bond *et al.* 1991), and from it, Bower (1991) and Lacey and Cole (1993) derived conditional probabilities for mergers between halos. Given these probabilities, it is possible to construct algorithms which generate random merger histories, tracing how the material in a halo at the present day was distributed in smaller halos at earlier times (e.g. Cole 1991; Kauffmann & White 1993; Kauffmann, White & Guiderdoni 1993; Cole *et al.* 1994; Somerville & Kolatt 1999). In semi-analytic (SA) models, these histories, or 'merger trees', are then combined with approximate descriptions of gas cooling, star formation, feedback and other processes to produce a general description of galaxy formation. In general, these models contain no spatial information about halos or their contents, but predict only spatially averaged quantities, focusing their attention on the process of dissipation, rather than dynamical issues. Since their invention, SA models have had some success in predicting the luminosity function of field galaxies (Somerville & Primack 1999), the colours of ellipticals (Kauffmann 1996), and several other properties of galaxy populations as a whole. Only very recently, however, have they been used to consider structure on sub-halo scales in any detail (Bullock, Kravtsov, & Weinberg 2000, 2001a). This has limited their application to the problems of hierarchical galaxy formation discussed in the introduction, such as the survival of galactic disks or the excess substructure in galactic halos, which require careful consideration of the distribution and dynamics

of material within individual halos.

1.3 An Alternative Approach

To summarise the previous section, two main methods of modelling the formation and evolution of galaxies like the Milky Way have been considered to date. Numerical models represent the assembly of material into a galaxy halo in terms of the dissipationless dynamics of a set of point masses (N-body or similar codes), or the hydrodynamical evolution of a number of grid cells (smoothed-particle hydrodynamics, or SPH codes). In the former case, an estimated 10^7 particles are required to follow the dynamical evolution of substructure in a galactic halo, on the scale of the smallest dwarf galaxies, and to study disk evolution over the age of the universe; models of the second type are at least as expensive, and much more complex. Overall, numerical models produce robust results based on few assumptions, but the computational expense they incur limits their application, and their simplicity breaks down in its representation of certain processes, such as star formation, which are intrinsically complex.

The semi-analytic models of galaxy formation developed recently are much faster, permitting the investigation of statistical problems in galaxy evolution. These methods use the basic statistics of structure formation, together with simple, empirical recipes for the complex physical processes that determine galaxy evolution, to predict the global properties of galaxies. They do not follow the dynamics of galaxy interactions in detail, however, nor do they resolve structure on sub-halo scales. Thus, they are unsuitable for addressing the dynamical problems which arise in hierarchical galaxy formation.

In this thesis, I outline a third approach to modelling galaxy formation, the construction of a semi-analytic model of the dynamical evolution of halo substructure.

Semi-analytic modelling of dynamical processes, which are complex and well suited to numerical methods, might seem counterintuitive. I show, however, that the dynamics of halo substructure are determined by a combination of a few processes, each of which can be modelled at much less expense, using analytic results, than they could be in a numerical simulation. In essence, the savings result from treating each lump in the halo as a single object, orbiting in a smooth potential, rather than representing it as a set of thousands particles, embedded within a halo resolved to a comparable degree. As a result, the method is roughly 5000–10000 times faster than a numerical simulation. The model can be accurately calibrated by comparing its output to the few existing high-resolution simulations of galactic halos currently available, and then run hundreds or thousands of times in order to explore galaxy formation in different cosmologies, or using different physical assumptions, as well as to estimate the statistical likelihood of processes such as disk disruption.

The other main advantage of this approach is more subtle. Physical models serve to determine how much of the complexity of natural systems can be explained by simple underlying principles. A successful model should not only reproduce what is observed, but should also do so in a way which makes the observation easy to understand. If a model can reproduce observations using a more abstract or synthetic description of the relevant phenomena, this can provide valuable insights into their workings. This advantage must of course be weighed against the increased robustness of models based only on low-level physics. The approach I outline in this thesis includes explicit descriptions of physical processes such as dynamical friction and tidal heating, rather than generating them as a by-product of the gravitational interactions of multi-particle systems. In this sense, it gives concise and definite expressions for these processes, which can be tested in future simulations, and applied in future analytic work.

1.4 Outline of the Thesis

This thesis is divided into two main parts, which describe an analytic model of satellite dynamics, and a semi-analytic model of galaxy formation based on this analytic work, respectively. The first part consists of three chapters, which present analytic treatments of dynamical effects, and test and calibrate these components of the model by comparison with numerical simulations. Chapter 2 explains how to calculate satellite orbits, and account for the orbital decay produced by dynamical friction. Chapter 3 describes a simple method for determining mass-loss rates, while chapter 4 discusses the heating of satellites by tidal shocks. I summarise the analytic model for dynamical evolution at the end of chapter 4.

The second part of the thesis then uses this analytic theory to construct a full model of galaxy formation. To determine initial conditions for the evolution of individual halos, I generate merger trees, using a method discussed in chapter 5. In chapter 6, I explain how to relate these merger trees to the process of galaxy formation, and outline a basic model of halo assembly, which I compare to dissipationless numerical simulations. In chapters 7 and 8, I then apply this model to the two specific questions raised in the first part of this introduction. In chapter 7, I study the formation of the dwarf satellites of the Milky Way, as well as the origin of its stellar Corona, under a variety of different assumptions about the star formation rate in dark halos. In chapter 8, I then determine how many galaxies like the Milky Way would have survived disruption by infalling satellites, when the last major merger is likely to have taken place, and how much minor heating should occur in the disk over time. I conclude with a summary of the semi-analytic model, its advantages and limitations, and the insights it provides into the process of galaxy formation.

Chapter 2

Calculating Satellite Orbits

In this chapter I describe the physical context of the analytic model of subhalo dynamics presented in the first part of this thesis, and introduce two sets of numerical simulations that will be used to test the accuracy of the model. I explain the orbital calculations of the analytic model and describe their implementation in an simple code. In simulations with a live halo, the amplitudes of the satellite orbits decay rapidly due to dynamical friction, the response of the halo to the presence of the satellite. I introduce an analytic description of this effect and include it in the orbital calculation. Finally, I show that mass loss must be accounted for in order to reproduce the properties of satellite orbits observed in the simulations.

The growth of dark matter halos has been studied extensively over past decade, using high-resolution numerical simulations (see Moore (2000) for a recent review). In CDM cosmologies, dark matter halos form in regions of space containing smaller halos which have already collapsed as well as larger, low-density structures which surround and connect these smaller lumps. Halos grow by a process of accretion and merging which disrupts much of this structure, but preserves the dense cores of pre-existing halos as distinct objects within the larger halo. These surviving remnants of smaller, older halos, which had already collapsed before they merged with the larger system, are referred to as 'subhalos'. In numerical simulations, subhalos are typically small, compact and roughly spherical structures, and exist in large numbers within any well-resolved halo. A halo like that of the Milky Way, for instance, contains roughly 1000 subhalos with masses greater than those of the smallest dwarf galaxies (Klypin *et al.* 1999b; Moore *et al.* 1999). The total mass of a halo is much larger than the combined mass of its subhalos; the remainder of this material is distributed smoothly

throughout the system, in a regular density profile which is similar for halos on all scales and at all epochs (Navarro, Frenk & White 1996, 1997; Moore *et al.* 1998).

These basic features suggest a simple model of halo substructure in which subhalos are represented by point-particles, orbiting in the potential of the main halo, and the rest of the matter distribution is taken to be smooth and roughly static over short timescales. The goal of the first section of this thesis is to determine the dynamical evolution of the satellites in such a model. In the next three chapters, I describe a basic analytic model of satellite dynamics, and test its accuracy by comparison with two sets of simulations of small satellites orbiting in the potential of a large companion. I focus on the evolution of a satellite's orbit in this chapter, study the change in its bound mass in the next chapter, and determine the effects of rapidly fluctuating tidal forces on satellite structure in chapter 4.

2.1 An Overview of the Numerical Simulations

Given the overall complexity of satellite dynamics, any analytic model of the underlying physical processes requires testing and calibration by comparison with numerical simulations. Simple simulations offer the clearest test of analytic descriptions of satellite dynamics, and should be easiest to reproduce. On the other hand, since the ultimate goal of the first section of this thesis is to describe satellite evolution in a cosmological setting, it is equally important to test the effectiveness of the method described here in more complex and realistic situations. Numerical simulations are also subject to their own limitations and uncertainties, and can produce quite different results depending on the mass resolution, softening, and the accuracy of the force calculations involved. To act as an effective comparison, the numerical simulations should be sufficiently robust in this sense to give confidence in their results, and they should cover the broadest possible range of masses, densities and orbital parameters.

With these considerations in mind, I chose to compare the predictions of the analytic model to two sets of simulations, a set of simulations of satellite disruption in an analytic potential by Hayashi & Navarro (2001) (see also Hayashi (2001)), and a set of self-consistent simulations of encounters between disk galaxies and small satellites by Velázquez & White (1999).

2.1.1 The Hayashi & Navarro Simulations

The simulations by Hayashi & Navarro (HN hereafter) follow the evolution of a small satellite with a Navarro, Frenk & White (NFW) density profile, orbiting in a smooth, static potential generated by another NFW profile with the same concentration, but a larger scale radius and mass. The NFW profile has the form:

$$\rho(r) = \rho_0 \, r_s^3 / r(r + r_s)^2, \tag{2.1}$$

where the scale radius $r_s = 1$ for the satellite and 10 for the main halo, and the integrated mass within 10 r_s was 1 for the satellite (which was truncated outside this radius) and 300 for the main halo ($G = 1$ in these simulations). The HN simulations do not include the effects of dynamical friction, since the background potential is static, nor do they explore the effects of different satellite density profiles or complex background potentials, but they are much simpler to model analytically as a result, and cover a wide range of orbital parameters. They also use carefully chosen softening lengths, a large enough number of particles, and sufficiently short timesteps to guarantee robust results. The orbital parameters for these simulations are listed in table 1, while the circular-velocity curve for the background potential, that is, the velocity of a circular orbit, $V_c = \sqrt{rF_r}$, as a function of r, is shown in the top panel of figure 2.1.

Table 2.1 Summary of Simulations from Hayashi & Navarro (2001)

Name	$r_{\rm apo}$	$r_{\rm peri}$	$\epsilon_{\rm circ}$	Name	$r_{\rm apo}$	$r_{\rm peri}$	$\epsilon_{\rm circ}$
HN300:15	300	15	0.261	HNc15	15	15	1.000
HN300:30	300	30	0.430	HNc30	30	30	1.000
HN300:60	300	60	0.751	HNc60	60	60	1.000
HN100:15	100	15	0.505	HNc75	75	75	1.000
HN100:30	100	30	0.751	HNc100	100	100	1.000
HN100:60	100	60	0.949	HNc150	150	150	1.000
HN60:15	60	15	0.665	HNc300	300	300	1.000
HN75:15	75	15	0.592				
HN150:15	150	15	0.398				

2.1.2 The Velázquez & White Simulations

The simulations by Velázquez and White (1999, VW hereafter) are much more complex, having been designed to study disk heating in minor mergers. While harder to reproduce, they test the success of the analytic model in describing the more complex phenomena associated with the planar geometry of the disk, and with interactions between the satellite and a responsive background. These simulations are fully self-consistent, including a realistic galaxy potential consisting of a dark halo, a stellar spheroid/bulge, and a disk, each represented by many thousands of particles. Their three satellite models have density profiles similar to those of dwarf spheroidal galaxies (although they are much more massive than typical dwarf spheroidals). VW consider eight different orbits, with cosmologically representative circularities, many of them for two or three different satellite models, and record the radial coordinate and bound mass of the satellite over the course each simulation. Their results show evidence for dynamical friction from the halo and the disk, tidal mass loss, shocking

as the satellite passes through the disk, and many other phenomena, so they offer a rigorous test of the analytic model presented here.

The details of these simulations are given in VW, but I will repeat them here for clarity. The background potential consists of three components, a truncated isothermal halo with a core, a stellar bulge, and an exponential disk. The density profiles of the three components are:

$$\begin{aligned}
\rho_{\rm h}(r) &= \frac{M_{\rm h}\alpha}{2\pi^{3/2}r_{\rm cut}}\frac{\exp(-r^2/r_{\rm cut}^2)}{r^2+\gamma^2} \\
\rho_{\rm b}(r) &= \frac{M_{\rm b}}{2\pi}\frac{a}{r(a+r)^3} \\
\rho_{\rm d}(r) &= \frac{M_{\rm d}}{4\pi R_{\rm d}^2 z_0}\exp(-R/R_{\rm d})\,{\rm sech}^2(z/z_0)
\end{aligned}$$

where the masses and scale lengths of the components are:

$$\begin{aligned}
M_{\rm h} &= 7.84\times10^{11}\,{\rm M}_\odot, \gamma = 3.5\,{\rm kpc}, r_{\rm cut} = 84\,{\rm kpc}, \\
M_{\rm b} &= 1.87\times10^{10}\,{\rm M}_\odot, \quad a = 525\,{\rm pc}, \\
M_{\rm d} &= 5.6\times10^{10}\,{\rm M}_\odot, R_{\rm d} = 3.5\,{\rm kpc}, \; z_0 = 700\,{\rm pc}\ .
\end{aligned}$$

The circular-velocity curve for this potential, calculated in the plane of the disk, is shown in the bottom panel of figure 2.1, along with the contributions from the disk, the bulge and the dark halo. The satellites S1, S2, and S3 are King models, with the core radii and initial concentrations listed table 2.2. VW carried out fifteen simulations in the potential considered here, with various combinations of satellite model and initial orbital parameters, as listed in table 2.3. (The orbits are specified by their initial size and circularity, and by their initial inclination to the disk. $r_{\rm p}$ and $r_{\rm a}$ are the initial pericentric and apocentric radii of the orbit, respectively. The circularity $\epsilon_J \equiv J/J_{\rm c}$, where J is the initial angular momentum of the orbit and $J_{\rm c}$ is the angular momentum of a circular orbit with the same energy. $\theta_{\rm i}$ is the angle between the initial angular momentum vector of the satellite and the angular momentum vector of the disc.)

Table 2.2 Velázquez & White (1999) Satellite Models

Satellite	Model	Density Profile	Mass ($M_\odot$)	r_c (pc)	r_t (kpc)
VW S1	King	$\simeq \rho_0 r_c^2/(r^2+r_c^2)$ out to r_t	5.6×10^9	1000	6.3
VW S2	"	"	5.6×10^9	500	6.3
VW S3	"	"	1.12×10^{10}	850	8.5

Table 2.3 Summary of Simulations from Velázquez & White (1999)

Name	Satellite model	θ_i	ϵ_J	r_p (kpc)	r_a (kpc)
G1S1	S1	45°	0.33	5.25	59.0
G1S2	S1	0°	0.55	10.5	55.0
G1S3	S1	45°	0.55	10.5	55.0
G1S4	S1	90°	0.55	10.5	55.0
G1S5	S1	135°	0.55	10.5	55.0
G1S6	S1	180°	0.55	10.5	55.0
G1S7	S1	0°	0.82	21.0	46.5
G1S8	S1	45°	0.82	21.0	46.5
G1S9	S2	0°	0.55	10.5	55.0
G1S10	S2	45°	0.55	10.5	55.0
G1S11	S2	90°	0.55	10.5	55.0
G1S12	S2	135°	0.55	10.5	55.0
G1S13	S2	180°	0.55	10.5	55.0
G1S14	S3	45°	0.55	10.5	55.0
G1S15	S3	135°	0.55	10.5	55.0

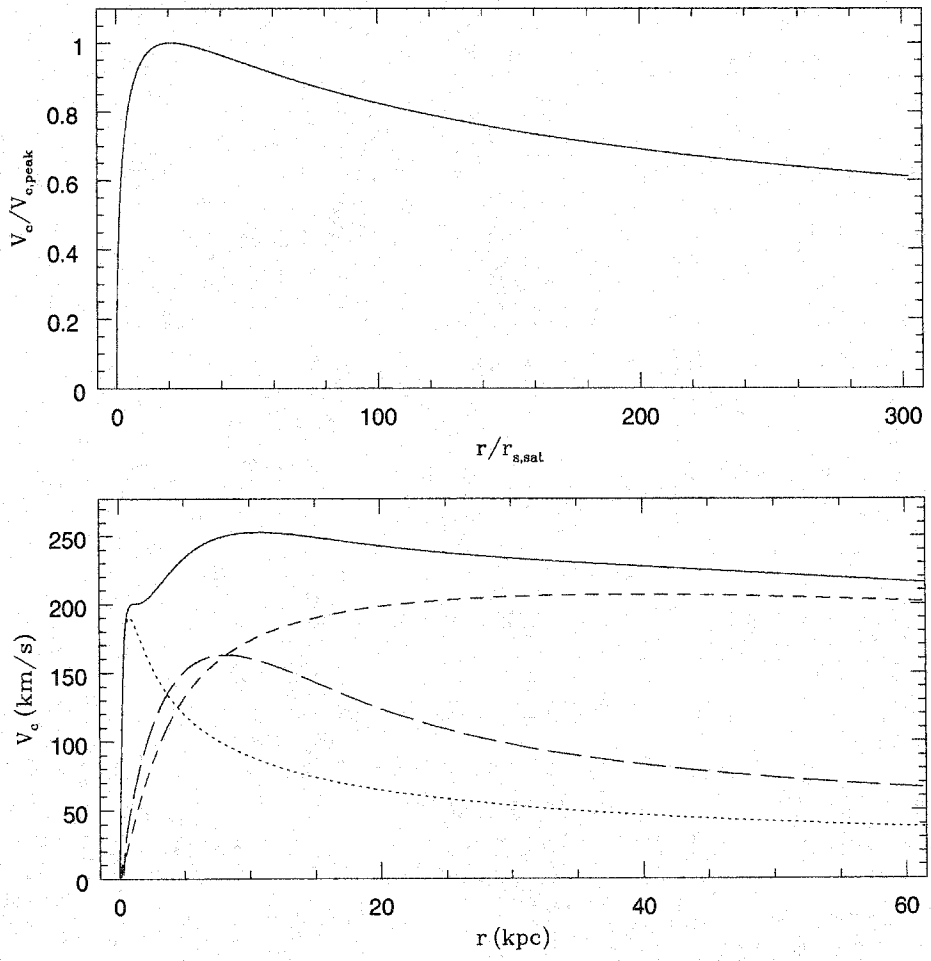

Figure 2.1 Circular velocity vs. radius for the two model potentials considered in this work. The top figure shows the circular velocity, normalised to its peak value, in the NFW halo of Hayashi & Navarro (2001). The bottom figure shows the total circular velocity for the model of Velázquez & White (1999), as well as the individual contributions from the dark halo, disk and bulge (short-dashed, long-dashed and dotted lines respectively).

2.2 Basic Orbital Calculations

2.2.1 The Integration Scheme

The orbit of a satellite, or more specifically the trajectory of its centre of mass, can be calculated by treating the satellite as a single point particle, moving in the

smooth potential generated by the main halo. This approximation will accurately reproduce the actual motion of the centre of mass of an extended distribution of particles in a smooth potential, provided the scale over which the potential changes is large compared to the scale of this distribution. Thus a method based on this approximation should be able to reproduce the results of numerically simulated orbits, provided the pericentre of the orbit is larger than the size of the satellite itself. In a general implementation of this method, the gravitational force field, or the gradient of the potential, is calculated at the position of the fictitious particle, and the corresponding acceleration is then applied to its motion.

The problem of integrating motion has been investigated extensively in several different contexts, including solar-system calculations, where long-term accuracy is essential (Saha, Stadel, & Tremaine 1997), studies of globular cluster dynamics, which involve less accurate force calculation for a larger number of objects (Aarseth 1985), and N-body simulations, where this trade-off becomes very important (Quinn *et al.* 1997). In the application considered here, a simple integration scheme is preferable for several reasons. First, each orbital calculation must be reasonably fast, to limit the computational burden on the full model, where several thousand satellites orbit within the halo for roughly a thousand timesteps. Second, the need to explore orbits in many different potentials, including some which can not be expressed in analytic terms, precludes the use of any method which relies on optimising calculations for a specific potential, or requires analytic expressions for the force or potential. Finally, and perhaps most importantly, energy and angular momentum need not be conserved to high precision in the orbital calculations, since these quantities will be substantially modified by dynamical friction in any case, by amounts uncertain to within 10% or so, and since the background potential will also change substantially over the course of a typical orbital period in the full model of halo formation.

Given these considerations, I chose to use a very simple, second-order Runge-Kutta integration scheme, the simplest that would satisfy these accuracy requirements. This

scheme produced excellent results, conserving energy to within 0.01% per orbit for a circular orbit resolved with 1000 timesteps per orbital period, and to within 1% per orbit for a circular orbit with 100 timesteps per orbital period. Given that the full model of galaxy formation described in part II uses several thousand timesteps in all (one for each new satellite), and that an average satellite spends only a few orbital periods in the main halo, this accuracy is sufficient to guarantee that most orbits will be resolved with more than enough timesteps. (To correct for numerical instabilities in orbits with small pericentres, reduced timesteps can also be used for these orbits in the full semi-analytic model, as explained in chapter 6.) The forces were interpolated from a look-up table, so that forces for non-analytic potentials could be calculated ahead of time. To optimise the code, all the variables related to the satellite's position which were required for other calculations, such as the local density and circular velocity, were also stored in the same look-up table, reducing the overhead associated with each timestep.

2.2.2 Properties of Orbits in an Axisymmetric Potential

Orbits within general spherical or axisymmetric mass distributions have integrals of motion, related to the more familiar integrals of motion associated with Keplerian orbits[1] (Binney & Tremaine 1987, BT hereafter, p. 105). In spherical mass distributions (or in the plane of symmetry of axisymmetric systems), orbits conserve energy and the three components of their angular momentum vector independently, and are therefore confined to a plane. Bound orbits take the form of rosettes, corresponding to ellipses whose major axis precesses at a regular rate around the centre of the potential. The top panels in figure 2.2 show two projections of a typical rosette orbit in the plane of the potential used by VW. These orbits have well-defined apocentres and peri-to-apocentre ratios, determined by their total energy and angular momentum

[1]That is, orbits in the potential of a point mass.

respectively, just as Keplerian orbits do.

Inclined orbits in axisymmetric potentials are no longer confined to a plane, and are consequently more complicated to describe. The inclined orbits considered in the VW simulations are of the 'box' type; Their integrals of motion confine them to a volume which is approximately bounded by two cones, corresponding to a maximum angle of inclination with respect to the plane of symmetry, and two spheres, corresponding to their minimum and maximum radial coordinates[2] . They typically pass through the plane of symmetry of the potential in a series of points, which vary in radius between the limiting values. The bottom panels in figure 2.2 show an inclined orbit in the VW potential, while figure 2.3 shows the same orbit plotted in the cylindrical coordinates R and z. The four bounding surfaces, and varying radius of the disk crossings are apparent in this projection.

2.2.3 Comparison with Simulations

Having constructed the potentials described in the previous section, it is possible to test the integration scheme in a realistic context, comparing the evolution of analytic and numerical orbits with the same initial parameters. Figure 2.4 shows the radial position of satellites as a function of time, in six simulations by VW (refer to table 2.3 for a summary of the orbit parameters). The points are the simulation results, while the solid curves show the corresponding analytic result. We see that while the amplitude and duration of the orbit is well matched by the analytic calculation over the first orbital period, the agreement deteriorates progressively, until after a few orbital periods, the analytic and numerical orbits are completely different both in phase and in amplitude. This cannot be due to any gross error in our integration scheme, since it preserves the amplitude and period of the orbit, or equivalently its energy and angular momentum, reasonably well, while the numerical orbit decays

[2]In general, the limiting surfaces have a more complicated shape. See BT, p. 114 ff. for a detailed discussion of the integrals of motion for these orbits

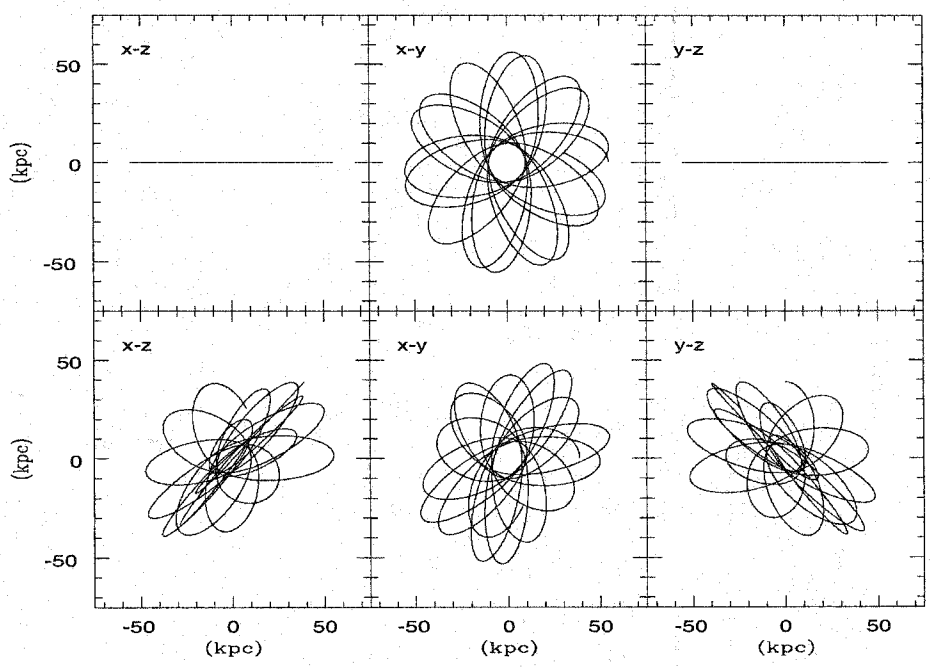

Figure 2.2 A typical orbit in the plane of symmetry of an axisymmetric potential (top panels), and a similar orbit out of the plane of symmetry (bottom panels). The potential used is that of VW.

quickly. Comparing simulations G1S3 and G1S14, which differ only in the mass of the satellite used in either case, provides a clue as to the source of this discrepancy. The orbit of the more massive satellite decays much faster, falling into the centre of the potential in half the time it takes the less massive one. This is evidence of a frictional force, whose overall amplitude is greater for more massive satellites.

2.3 Dynamical Friction

2.3.1 Chandrasekhar's Formula

The existence of a frictional force due to gravity alone was anticipated long before it was ever seen in numerical simulations. Chandrasekhar (1943), examining the dynamics of a particle moving through a uniformly background of other particles,

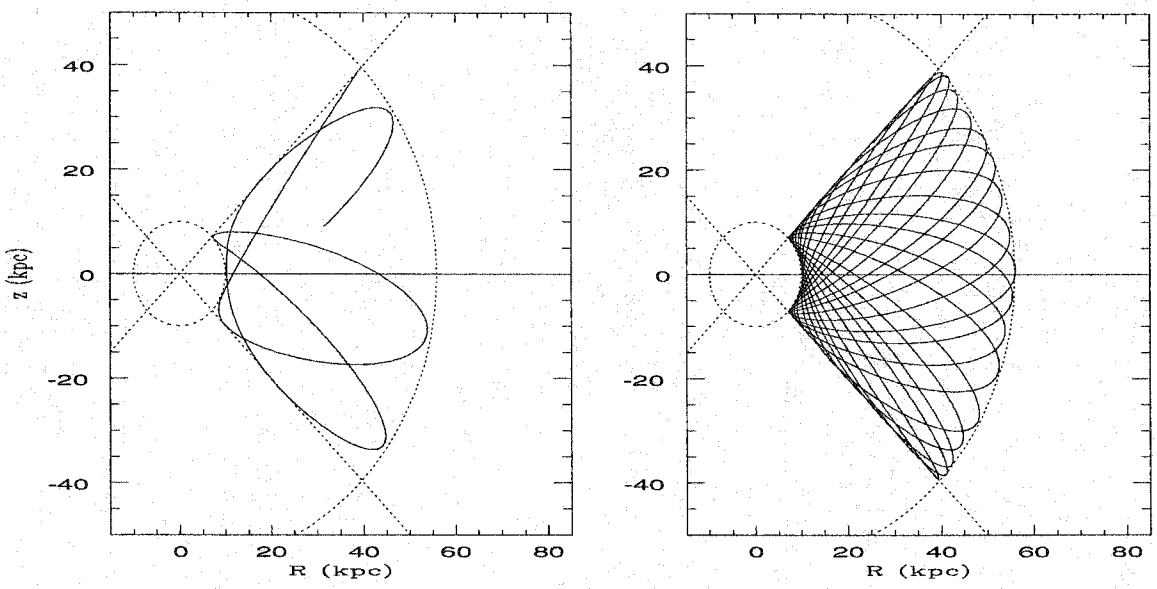

Figure 2.3 The orbit shown in the bottom panel of figure 2.2, projected onto the $R - z$ plane. (The left-hand figure shows the first few orbital periods for clarity.) Note the four bounding curves which restrict its radius and inclination.

pointed out that it would experience a net force opposing its motion, even if gravity was the only source of interactions between the particle and the background. There are several ways of understanding this result. One is to consider the response of the background particles and its effect on the moving particle. The moving particle attracts background particles towards its path, so that an overdense region forms in its wake; this wake then exerts a gravitational force on the particle, resisting its forward motion. More generally, one can view dynamical friction as a form of equipartition, driven by the collisional terms in the interaction between a single particle and a background system of many particles. Successive random encounters between the single particle and the background particles tend to transfer the former's forward momentum to the background, slowing it down progressively.

Chandrasekhar's result is derived in detail in chapter 7 of BT; I will reproduce the main parts of this derivation here for clarity. (It is possible to derive similar

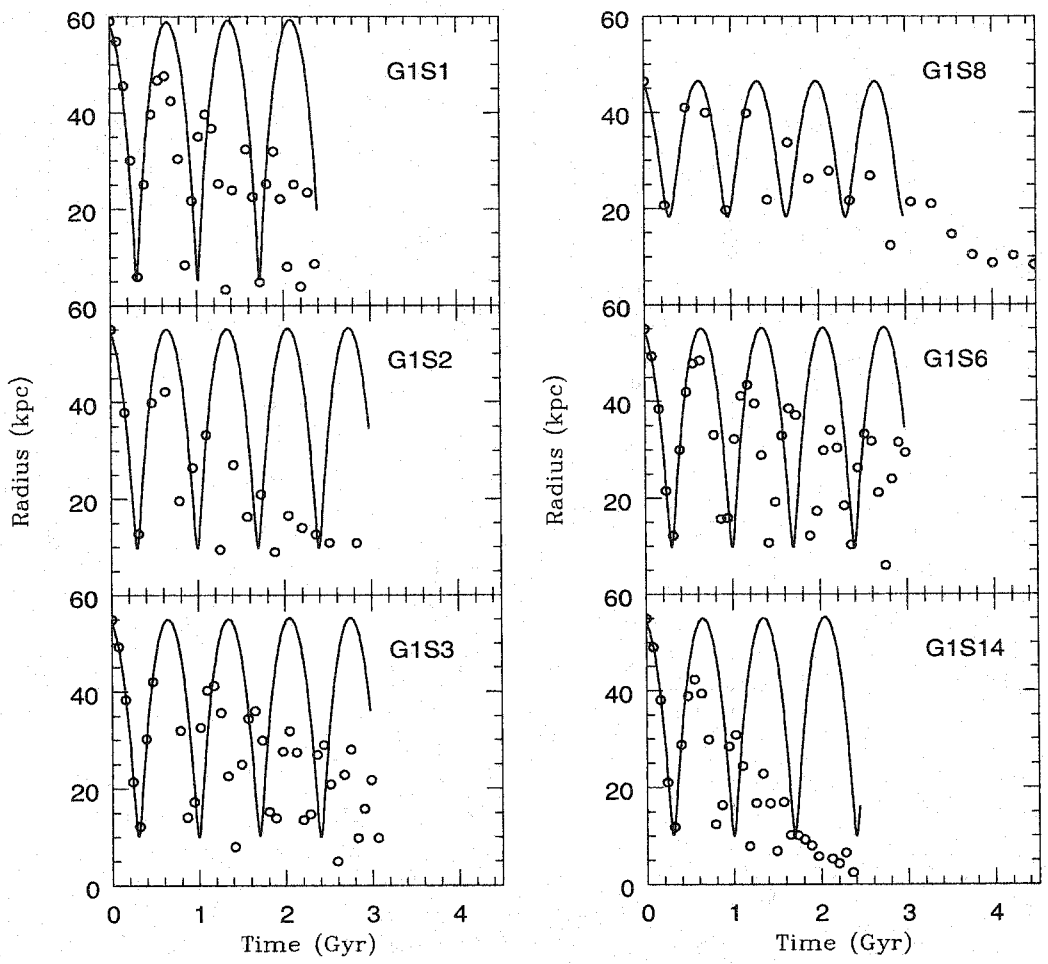

Figure 2.4 A basic orbital calculation, compared with the numerical results of VW. Note the orbital decay apparent in the simulations.

expressions using several different approaches, including the fluctuation-dissipation theorem; see Colpi, Mayer & Governato (1999), Bekenstein & Maoz (1992), and Maoz (1993) and references therein for examples of these approaches.) If we consider a single, hyperbolic encounter between a particle of mass M and a background particle of mass m, characterised by the initial relative velocity $\mathbf{V}_0 = \mathbf{V}_m - \mathbf{V}_M$ and the impact parameter b, the resulting changes in the velocity of M are:

$$|\Delta V_{M\perp}| = \frac{2mbV_0^3}{G(M+m)^2}\left[1+\frac{b^2V_0^4}{G^2(M+m)^2}\right]^{-1}, \tag{2.2}$$

$$|\Delta V_{M\|}| = \frac{2mV_0}{M+m}\left[1+\frac{b^2V_0^4}{G^2(M+m)^2}\right]^{-1} \tag{2.3}$$

(BT, eq. 7–10). If the particle M moves through a sea of similar particles, distributed symmetrically around the axis of its motion, the contributions to $|\Delta V_{M\perp}|$ from individual encounters will average out, while the contributions to $|\Delta V_{M\|}|$ will produce a net deceleration. If the phase-space density of background particles can be written as $f(b,\mathbf{v}) = D(b)n(\mathbf{v})$, then the rate at which the moving particle encounters background particles with velocities in the velocity-space volume $d^3\mathbf{v}$ and impact parameters between b and $b+db$ is:

$$2\pi b db \times V_M \times D(b)n(\mathbf{v})d^3\mathbf{v}\,. \tag{2.4}$$

Thus the rate of change of $\mathbf{V_M}$ due to these encounters is:

$$\left.\frac{d\mathbf{V_M}}{dt}\right|_{\mathbf{v}} = \mathbf{V_0}n(\mathbf{v})d^3\mathbf{v}\int\frac{2mV_0}{M+m}\left[1+\frac{b^2V_0^4}{G^2(M+m)^2}\right]^{-1}2\pi bD(b)db\,. \tag{2.5}$$

In the case where $D(b)$ is constant between 0 and $b_{\max}$, and zero beyond this, integrating this expression over all impact parameters gives:

$$\left.\frac{d\mathbf{V_M}}{dt}\right|_{\mathbf{v}} = 2\pi\ln(1+\Lambda^2)G^2m(M+m)D_0d^3\mathbf{v}\frac{(\mathbf{v}-\mathbf{V}_M)}{|\mathbf{v}-\mathbf{V}_M|^3}\,, \tag{2.6}$$

where:

$$\Lambda \equiv \frac{b_{\max}V_0^2}{G(M+m)}\,. \tag{2.7}$$

As explained in BT, Λ is very large in typical applications, so the approximation $\ln(1+\Lambda^2) \simeq 2\ln\Lambda$ is usually made.

If Λ is constant, and n is a function of $|\mathbf{v}|$ only, then equation (2.6) is similar in form to the expression for the force at a point within a spherical mass distribution. As

a result, the integral over velocities can be transformed into the following expression, by analogy with Newton's first theorem:

$$\frac{d\mathbf{V}_M}{dt} = -16\pi^2 \ln \Lambda G^2 m(M+m) D_0 \frac{\int_0^{V_M} n(v) v^2 dv}{V_M^3} \mathbf{V}_M, \tag{2.8}$$

that is the contribution to the drag force comes from particles with velocities less than V_M. This result is known as Chandrasekhar's formula. Finally, evaluating this integral for a Maxwellian velocity distribution of dispersion σ, and writing $mD_0 = \rho_0$, we have, for a massive object ($M >> m$),

$$\frac{d\mathbf{V}_M}{dt} = \frac{-4\pi \ln \Lambda G^2 \rho_0 M}{V_M^3} [\mathrm{erf}(x) - x\mathrm{erf}'(x)] \mathbf{V}_M, \tag{2.9}$$

where $\mathrm{erf}(x)$ is the error function:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp^{-t^2} dt, \tag{2.10}$$

and $X \equiv V/\sqrt{2}\sigma$. Dynamical friction in the analytic model will be calculated using this more compact form of Chandrasekhar's formula.

2.3.2 Calculating the Coulomb Logarithm

The derivation of Chandrasekhar's formula above assumed a massive point particle, moving through a homogeneous background of much lighter particles with an isotropic Maxwellian velocity distribution of zero mean, extending to some large distance $b_{\max}$. In particular, interactions between the moving particle and the background particles were treated as isolated two-body encounters. Numerous studies of satellite dynamics (Weinberg 1986; Cora, Muzzio & Vergne 1997; Bontekoe & van Albada 1987; van den Bosch *et al.* 1999; VW; Colpi, Mayer & Governato 1999), have shown the formula to be more widely applicable, however, providing a good approximation to the drag force on an extended satellite in a more complex background system, provided that the Coulomb logarithm is adjusted appropriately. Clearly, if the satellite is comparable to background system either in mass or in spatial extent,

the approximations used in the derivation of(2.9) will no longer be valid. Based on the studies mentioned above, Binney & Tremaine (1987) suggest that Chandrasekhar's formula will be fairly accurate provided that the mass of the satellite does not exceed 20% of the mass of the larger system, and that the orbit of the satellite lies neither outside the larger system nor completely within its core (BT, p. 427).

The argument of the Coulomb logarithm can be expressed as the ratio of two characteristic scales: $\Lambda = b_{\max}/b_{\min}$, where $b_{\max}$ and $b_{\min}$ are measures of the maximum and the minimum impact parameters of the background particles contributing to the wake, or equivalently the scales of the background system and the satellite respectively. For a background system with an arbitrary density profile, $b_{\max}$ is equal to the spatial extent of the system, weighted by the density at a given impact parameter, as in the integral in equation (2.5). Equations (2.2–2.3) are only strictly valid for systems of infinite extent, however, in which case this integral diverges. This introduces some uncertainty into the appropriate choice of limiting value. In practice, $b_{\max}$ is instead taken to be a scale characteristic of the system. Possible choices for a spherically symmetric system include the half-mass radius of the system (e.g. Quinn & Goodman 1986), the distance over which the background density changes by a factor of two (Binney & Tremaine 1987), and the tidal radius of the halo or the distance between the satellite's position and the centre of background system (Colpi *et al.* 1999). For the special case of an object at the centre of a spherically symmetric system with a density profile $\rho(r)$, for instance, Maoz (1993) gives the expression:

$$\ln \Lambda = \int_{b_{\min}}^{b_{\max}} \frac{\rho(r')}{\rho_0} d(\ln b) \tag{2.11}$$

where $b_{\max}$ is the maximum extent of the larger system and $b_{\min}$ is the minimum impact parameter discussed below. When the characteristic scale of the main system is comparable to, or smaller than, the satellite itself, the magnitude of dynamical friction should be greatly reduced. This can be accounted for by reducing the Coulomb

logarithm or averaging the background density over the scale of the satellite, as discussed below, although Chandrasekhar's formula is no longer strictly valid in these cases.

The value of $b_{\rm min}$ is equally ill-defined. We saw above that for a point mass satellite, $b_{\rm min} \equiv G(M+m)/V^2$ (that is the radius of a circular orbit of velocity V in the reduced potential), where m is the mass of the background particles and V is the relative velocity of the satellite in a typical encounter. For rapidly moving satellites, V will be approximately equal to the velocity of the satellite relative to the background (BT), while for slowly moving satellites, it will be approximately equal to the local velocity dispersion (Chandrasekhar 1943). For extended satellites, the contribution to dynamical friction from background particles with impact parameters smaller than the size of the satellite itself will be greatly reduced, however. This can be accounted for by increasing $b_{\rm min}$ beyond the point-mass value. White (1976), for instance, derived an expression for $b_{\rm min}$ which is approximately equal to $0.2\,r_{\rm t}$ (or very roughly the half-mass radius) for a wide range of King profiles. A good general approach is to take $b_{\rm min}$ to be the larger of the half-mass radius of the satellite and the point-mass value $G(M_{\rm sat}+m)/V^2$ (e.g. Quinn & Goodman 1986), although this is only likely to work well for small systems (Maoz 1993).

In summary, although the Coulomb logarithm is well defined in the original derivation above, its value in realistic cases is uncertain. In practice, it is easiest to estimate values for $\ln\Lambda$ from the expressions above, but calibrate them using numerical simulations of appropriate systems. It is worth noting one general feature of the logarithm, however. For most of the definitions of $b_{\rm min}$ and $b_{\rm max}$ discussed above, the argument of the Coulomb logarithm should scale roughly as $\Lambda \propto (M_{\rm sat}/M_{\rm halo})^{-1}$. We can see this most easily in the case of one isothermal halo orbiting within another; the characteristic scale of either system is proportional to its mass, and thus Λ is proportional to the ratio of the masses from the arguments above. This point is worth noting, given both the wide range of mass scales relevant to the question of halo substructure, and the

fact that an individual satellite's mass can change substantially over the course of its orbit. I will use this scaling when fitting individual orbits from the VW simulations in chapter 4, and in the full semi-analytic model in chapter 6.

Finally, it is not clear whether Chandrasekhar's formula adequately describes dynamical friction in a non-uniform or non-isotropic distribution of background particles, such as a galactic disk, and if so how to estimate the Coulomb logarithm in these cases. By the arguments given earlier, such systems should still exert a drag force on a particle or satellite moving through them. It is possible to calculate dynamical friction for a uniform background of particles with an ellipsoidal velocity distribution, using the approach outlined above. In this case the frictional force is strongest in the direction of the smallest principal axis of the distribution (BT). Maoz (1993) and Domínguez-Tenreiro & Gómez-Flechoso (1998) also derived formulae for the magnitude of the energy loss produced by an arbitrary background distribution with a uniform velocity dispersion, but could not specify the direction of the corresponding frictional force. In the interactions considered in this work, disk friction is of secondary importance compared with halo friction, so I will limit myself to using Chandrasekhar's formula to calculate its approximate direction and magnitude.

2.4 Adding Friction to the Orbital Calculations

2.4.1 Analytic Prescription

Dynamical friction can be added to the orbital calculations described above using expression (2.9), with local values for ρ and σ, and a value of $\ln\Lambda$ adjusted to fit the orbital decay rate seen in simulations. All three of the background components in the VW simulations will contribute to dynamical friction. Since the disk particles are both kinematically and spatially distinct from the other two components, their contribution should be calculated separately, while the contribution of the central

stellar bulge can be combined with the much larger effect of the halo. Thus the total frictional deceleration is:

$$\begin{aligned} \mathbf{F}_{\rm df} &= \mathbf{F}_{\rm df,halo} + \mathbf{F}_{\rm df,disk} \\ &= -4\pi G^2 M^2 \times \sum_{i={\rm h,d}} \rho_i(< V_{{\rm rel},i}) \ln\Lambda_i \frac{\mathbf{V}_{{\rm rel},i}}{|\mathbf{V}_{{\rm rel},i}|^3} \end{aligned} \tag{2.12}$$

where $\mathbf{V}_{\rm rel,h} = \mathbf{V}_{\rm sat}$, $\mathbf{V}_{\rm rel,d} = \mathbf{V}_{\rm sat} - \mathbf{V}_{\rm rot}$,

$$\rho_i(< V_{{\rm rel},i}) = \rho_i(\mathbf{r})\left[{\rm erf}(X_i) - X_i\ {\rm erf}'(X_i)\right],$$

and $X_i \equiv |\mathbf{V}_{{\rm rel},i}|/(\sqrt{2}\sigma_i)$.

Here M is the mass of the satellite, $\mathbf{r}$ is its position, $\mathbf{V}_{\rm sat}$ is its velocity, $\mathbf{V}_{\rm rot}$ is the local circular velocity of the disk, $\rho_{\rm h}$ is the local density of the spherical (halo/bulge) component, $\rho_{\rm d}$ is the local density of the disk, $\ln\Lambda_{\rm h}$ and $\ln\Lambda_{\rm d}$ are the Coulomb logarithms for the halo/bulge and the disk, and $\sigma_{\rm h}$ and $\sigma_{\rm d}$ are the one-dimensional velocity dispersions of the halo/bulge particles and the disk particles respectively.

The spatial extent of the VW satellites is larger than the scale height of the disk, so the use of (2.9) to describe disk friction is highly approximate, particularly when they orbit in the plane of the disk. In particular, equation (2.9) was derived assuming that the density of the satellite's wake is constant, and equal to the background density at the centre of the satellite. One way of accounting for the expected reduction in the magnitude of dynamical friction is to average the density of the disk over the scale of the satellite. Since this scaling will vary from one satellite to another, I chose instead to smooth the density of the disk in the vertical direction by a fixed scale length, equal to two times the disk scale height, noting that this smoothing length is on the order of the satellite half-mass radius, for satellites with masses in the range where dynamical friction has a substantial effect. The disk velocity dispersion was taken to

be:

$$\sigma_{\rm d} \;=\; V_{\rm c,d}/\sqrt{2} = \sigma_{\rm o} \exp(-R/2R_{\rm d})\,, \tag{2.13}$$

where $V_{\rm c,d}$ is the circular velocity of the disk, the disk scale length is $R_{\rm d} = 3.5\,{\rm kpc}$ as above, and $\sigma_{\rm o} = 143\,{\rm km\,s^{-1}}$. This expression is the expected functional form for the velocity dispersion of an isothermal sheet, normalised to the value measured by VW.

The sum of the halo and bulge densities was used to calculate the other friction term, since these components are kinematically similar. The combined velocity dispersion of the halo and stellar spheroid was taken to be:

$$\sigma_{\rm h} \;=\; (V_{\rm c,h}^2 + V_{\rm c,b}^2)^{1/2}/\sqrt{2}\,, \tag{2.14}$$

where $V_{\rm c,h}$ and $V_{\rm c,b}$ are the circular velocities of the halo and the bulge, respectively. This expression is approximately valid over the range of radii of interest.

2.4.2 Comparison with Numerical Simulations

Figure 2.5 shows analytic orbits calculated as in figure 2.4, but with the effects of dynamical friction included. The Coulomb logarithms used, $\ln\Lambda_{\rm h} = 2.4$ and $\ln\Lambda_{\rm d} = 0.5$, were adjusted to match the decay rate seen in the first few orbits. To indicate the overall sensitivity to the precise values of the logarithms, the dotted curves show the analytic predictions if they are increased or decreased by 20%. Except for the orbits in the plane of the disk, most of the change in orbital decay is due to the change in $\ln\Lambda_{\rm h}$.

We see that the change both in the period and in the amplitude of the orbit with time is much better predicted than in figure 2.4. The match to the first few orbital periods is particularly good. The orbital decay rate is slightly overestimated, however, towards the end of each orbit. An obvious explanation for this is that the mass used in equation (2.12) should be adjusted as the satellite loses mass. Figure 2.6 shows the bound mass of each of the satellites in figures 2.4–2.5, as determined by VW. We

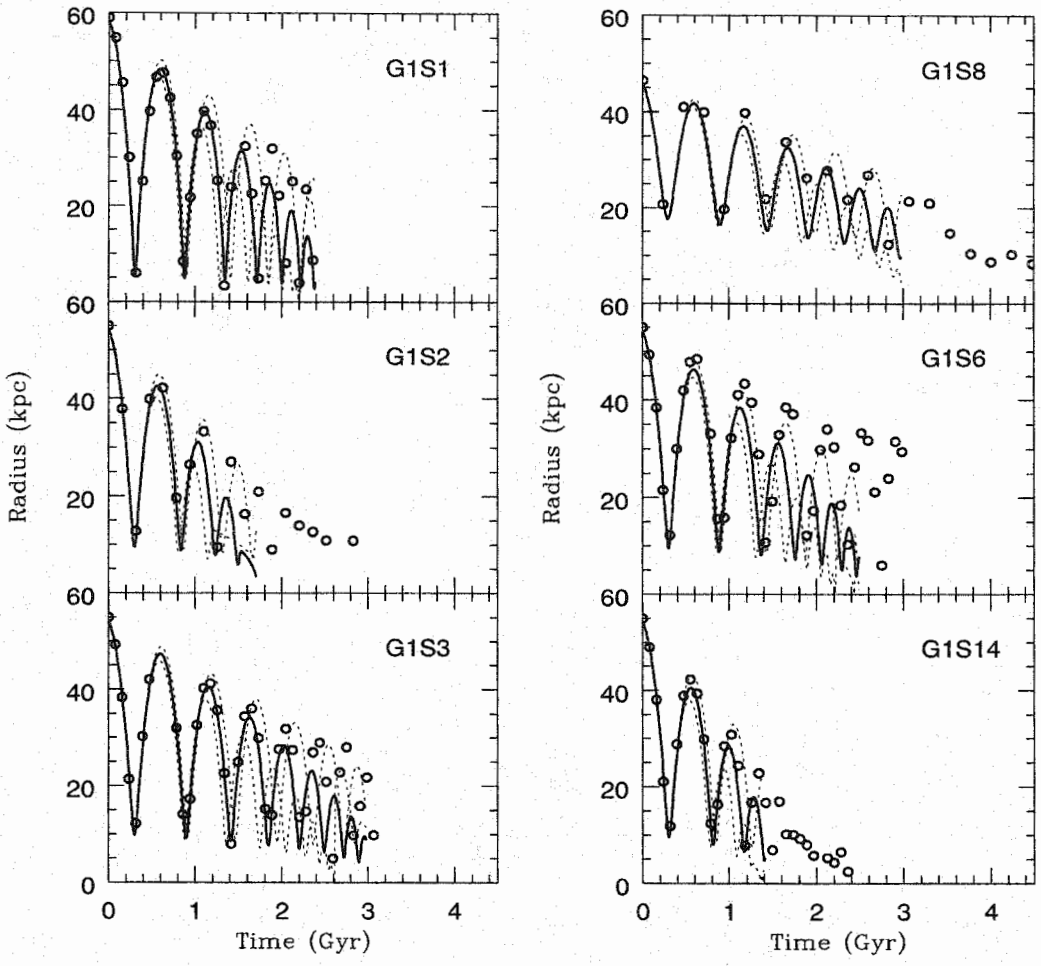

Figure 2.5 Analytic predictions compared with numerical orbits, with dynamical friction taken into account. Compare with figure 2.4.

see that the satellite can lose a substantial fraction of its mass before its orbit has decayed completely. This leads to a miscalculation of dynamical friction, and thus of the decay rate itself. To estimate the orbital decay rate accurately, one must account not only for dynamical friction, but also for mass loss. This is particularly true for less massive or lower density satellites, which will experience substantial mass loss over the course of their orbits, and fall into the potential more slowly as a result. The next chapter explores this aspect of dynamical evolution. The Coulomb logarithms may also increase slightly as a satellite loses mass, as discussed in the previous section.

This effect is less pronounced, however, so I will only include in the model once I have established an good description for mass loss and heating.

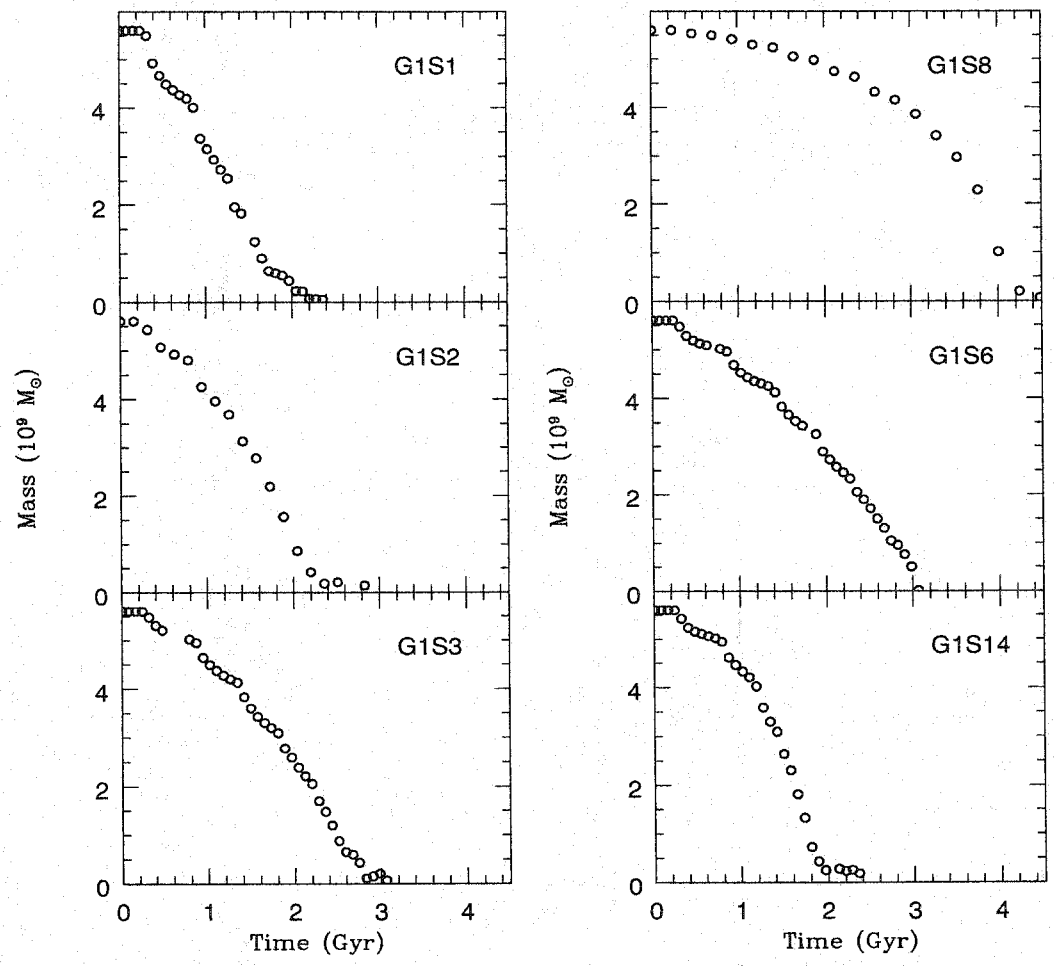

Figure 2.6 The bound mass for each of the satellites plotted in figures 2.4 and 2.5. Note the mass loss at each pericentric passage.

Finally, since the precise magnitude of dynamical friction has to be determined from simulations, it is worth making several cautionary remarks about this calibration. First, dynamical friction is a collisional effect, and should therefore depend sensitively on the fine structure of the background system used in a simulation. In particular, systems with too few particles may not resolve the wake of the satellite,

reducing the strength of dynamical friction measured in the simulation from its expected value (van Kampen 2000); in the limit of a static, unresponsive background system, simulations should show no orbital decay at all. Interestingly enough, however, some simulations with static potentials, notably those of HN, do show orbital decay. Figure 2.7 shows several of their orbits, with the apocentre-to-pericentre ratios indicated. We see that while the more circular orbit conserves its orbital energy, the more radial orbit, which loses the most mass, decays noticeably. Since the system as a whole must conserve orbital energy, the missing orbital energy must be transferred to the material stripped off the satellite. This effect has been noted in other simulations of very radial encounters (Heisler & White 1990), but is often ignored. It should also occur in fully self-consistent simulations with live halos, where it will add to the effects of dynamical friction. While this effect is generally less important than dynamical friction, it would lead to substantial errors in estimating $\ln \Lambda$ for orbits in a live halo if they were as radial as HN 300:15. Finally, some of the orbit-to-orbit variation seen in figure 2.5 may be may be an indication of the higher-order effects predicted by more sophisticated treatments of dynamical friction (Maoz 1993; Domínguez-Tenreiro & Gómez-Flechoso 1998). In this sense, the Coulomb logarithms determined here are average values, which should be roughly applicable to a wide range of orbits and potentials.

2.5 Summary

In this chapter I have introduced a basic model of halo substructure, consisting of small, dense satellites, orbiting in a smooth background potential. The orbital evolution of the satellites is determined using a simple integration scheme, which calculates the trajectory of the centre of mass of each satellite, as if it were a point mass. I show that while this method may provide an adequate description of satellite

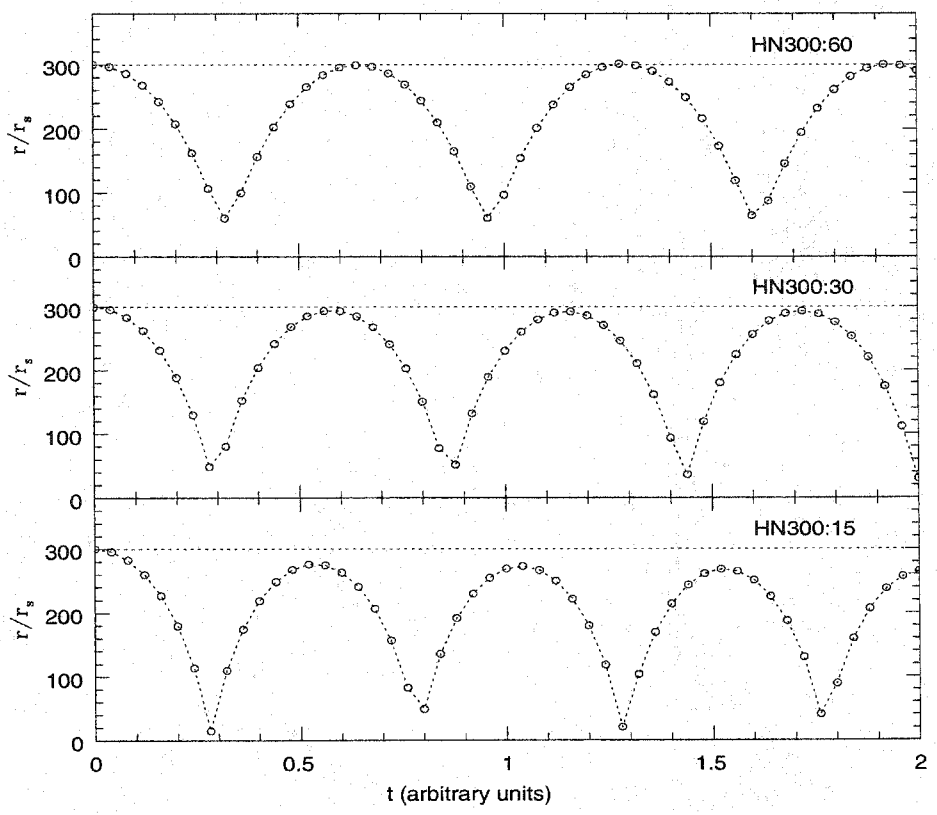

Figure 2.7 The radius vs. time in three simulations by HN. The horizontal line indicates the radius of the original apocentre. Note the orbital decay in the most radial orbit.

orbits in a static potential, it fails to reproduce the strong orbital decay seen in simulations with a 'live', or responsive, potential. This decay is the result of dynamical friction, the local response of the background to a moving satellite. The magnitude of dynamical friction can be estimated analytically, using Chandrasekhar's formula. By adding this drag force, with suitably chosen Coulomb logarithms, to the orbital calculations, it is possible to reproduce the behaviour of satellites in a live potential much more accurately. The overall evolution of the satellites still differs from this simple model, however, because they lose a substantial fraction of their mass during each orbital period, and therefore experience progressively less friction over the course of the simulation. I will attempt to account for this mass loss in the next chapter.

Chapter 3

Mass Loss

The material initially bound to a subhalo may become unbound as the subhalo orbits in the potential of a larger system. This mass loss represents an important correction to dynamical friction, as discussed in the previous chapter. A theory of mass loss is also required to reproduce the substructure seen in simulated halos, and to follow the evolution of material stripped from satellites in the halo formation process. In this chapter, I present two basic descriptions of mass loss, one applicable to static systems, and the other to rapidly changing systems. I develop a general model which interpolates between these two descriptions, and compare its predictions to the mass-loss rates seen in numerical simulations. The general model predicts the mass-loss rates for some of these orbits reasonably well, consequently improving estimates of the orbital decay rates over those predicted in the last chapter. Most of the simulations show faster mass loss than is predicted by the model, however. This is a consequence of tidal heating, which will be discussed in chapter 4.

Dynamical friction, as noted in the previous chapter, is the result of the main halo's response to a satellite moving through it, and depends sensitively on the mass of the satellite. We saw that orbital decay rates will be overestimated if the mass of a satellite is taken to be constant throughout its orbit, because this mass actually decreases with time. When a satellite subhalo first merges with the main halo, it consists of some initial distribution of material. This material may not remain spatially associated with the satellite as it orbits, however, particularly if it becomes unbound, that is, acquires a net positive energy with respect to the rest of the satellite. It is not clear whether dynamical friction depends on the bound mass of the satellite, or simply on the mass of material which is associated with it in space, but the distinction is largely semantic since any material not bound to the satellite will quickly be removed from its

vicinity. In any case, a reasonable estimate of satellite mass is required to implement dynamical friction properly, and should also be useful in relating the predictions of the analytic model to the subhalos seen in numerical simulations, which show evidence of mass loss (Ghigna *et al.* 1998), and to study the properties of the tidally stripped material itself.

Mass loss turns out to be much harder to describe in analytic terms than dynamical friction. The motion of a given particle, initially associated with a satellite in a simulation, is fundamentally chaotic, and there is no simple test to determine how long its orbit will remain associated with the motion of the satellite's centre of mass. The total mass or bound mass of a satellite can also be defined in several different ways, making comparison between published numerical results problematic. In this chapter, I present two models of mass loss, based on the tidal-limit approximation and on the impulse approximation respectively. The tidal-limit approximation is derived by considering the steady-state behaviour of a satellite on a circular orbit in a spherically symmetric potential, while the impulse approximation applies to encounters which are instantaneous compared with the internal dynamical time of the satellite. I show how these two models can be combined into a single, general description of mass loss on arbitrary orbits, and compare the predictions of this general model with simulation results.

3.1 Measuring Mass in Simulations

Before discussing the two extremes of mass loss, it is worth mentioning some of the minor complications of studying mass loss numerically. The mass associated with a specific substructure within a halo can be defined in one of two ways. Either it is the mass of material which is spatially associated, that is concentrated in one position with respect to the rest of the background material, or it is the mass of material

which is associated in phase-space, that is both in the same place and exhibiting the same dynamics. Algorithms to identify substructure in a halo that only use a local overdensity criterion, for instance, measure masses of the former type, while methods that also check whether material is bound measure masses of the second type. In practice one would expect both masses to be similar for objects much denser than the background, but they may differ in detail. A satellite moving on a circular orbit will ultimately be stripped to the point where it is 3 times denser than the mean background density interior to the radius of its orbit, or 9 times denser than the average density at its radius, for instance, if it orbits in an isothermal potential. Thus, including all the material spatially associated with the satellite would produce as mass estimate 10% higher than the bound mass alone.

Most numerical results use a mass estimate of the latter type, since it can be easily determined from simulation data, and is in some sense more physical; as noted above, this is probably the best estimate to use for dynamical friction calculations, though this has not been demonstrated rigorously. In any case, even the bound mass of a subhalo is not uniquely defined, since the criterion used to determine membership for an individual particle, namely that the sum of its kinetic and potential energies in some frame of reference and some potential be negative, depends on the frame of reference and the potential used. In practice, bound masses are usually estimated by selecting a group of particles based on their correlated positions, determining the potential generated by these particles alone, using the centre of mass or the most bound particle within the group to define a frame of reference, removing unbound particles from the group, and finally iterating through this process until the solution has converged to a final set of particles. The details of these steps vary from one implementation to the next, however, and may produce small differences in the estimated mass (e.g. Klypin *et al.* 1999a; HN). In more general situations, the issue of substructure within substructure (the 'cloud-in-cloud' problem) further complicates this process. While not an issue in the simulations considered here, this problem does

arise in analysing the full cosmological simulations considered in part II.

3.2 The Tidal Limit Approximation

3.2.1 Derivation

Given a fixed, static potential with several local minima, it is straightforward to determine whether particle orbits will remain trapped around these minima. For some particularly simple cases, the potential generated by a satellite orbiting in a larger halo is in fact static, when viewed in a frame of reference moving with the satellite. If the satellite's contribution to the potential can be built up self-consistently from trapped orbits, then this technique provides a way of estimating the total mass of material that will remain bound to the satellite for all time. I call this the *tidal limit approximation*, since the orbits are restricted to a specific volume at all times, producing a sharp edge or tidal limit to the density profile of the satellite.

The detailed derivation of the tidal limit, and its corresponding mass estimate, are described in chapter 7 of BT. I will reproduce the essential parts of their derivation here. Consider a satellite of mass m on a circular orbit of radius D, and co-rotating at its orbital frequency, in a spherically symmetric background potential. I shall initially consider the case were the main system has a mass M, entirely contained within the orbit of the satellite, and extend the result to more general cases later. The mass distribution in this system, as seen in a frame co-rotating with the orbit of the satellite, is static, and therefore the potential is static. We can define an integral of motion, Jacobi's integral, which is conserved along any orbit in this potential:

$$E_{\rm J} = \frac{1}{2}v^2 + \Phi(x) - \frac{1}{2}|\Omega \times x|^2 \tag{3.1}$$

$$= \frac{1}{2}v^2 + \Phi_{\rm eff}(x) \tag{3.2}$$

where $\mathbf{\Omega}$ is the vector $(0, 0, \Omega)$, and

$$\Omega = \sqrt{\frac{G(M+m)}{D^3}}$$

is the angular speed at which the systems orbit around their common centre of mass. Since $v^2 \geq 0$, a particle whose Jacobi integral is E_J will never enter into the region where $\Phi_\mathrm{eff}(x) > E_\mathrm{J}$, so curves of constant $\Phi_\mathrm{eff}(x)$ form impenetrable boundaries to the motion of particles in the combined system. There is a saddle point in $\Phi_\mathrm{eff}(x)$ between the satellite and the centre of the combined system, which marks the boundary between iso-potential contours centred around the satellite, and contours centered around the main system. This is shown schematically in figure 3.1. Taking the contour which passes through this point as the tidal limit of the system, one can calculate the approximate size of the satellite by determining the distance to this saddle point, that is the distance to the point where $(\partial\Phi_\mathrm{eff}/\partial x) = 0$. In the limit where $M >> m$, this gives a radius:

$$r_\mathrm{J} \simeq \left(\frac{m}{3M}\right)^{\frac{1}{3}} D \tag{3.3}$$

called the Jacobi limit, which is approximately equal to the tidal limit of the system (BT, equation 7-84).

3.2.2 Alternate Formulations

There are two alternative interpretations of the Jacobi limit which will be useful in considering satellite evolution. First, if we take the cube of equation (3.3) and rearrange terms, we obtain the result:

$$\frac{3m}{4\pi r_\mathrm{J}^3} = \bar{\rho}_\mathrm{s}(r_\mathrm{J}) = 3\left(\frac{3M}{4\pi D^3}\right) = 3\bar{\rho}_\mathrm{g}(D), \tag{3.4}$$

that is, the mean density of the satellite within its tidal limit exceeds the mean density of the main system interior to its orbit by a factor $\eta \equiv \bar{\rho}_\mathrm{s}(r_\mathrm{J})/\bar{\rho}_\mathrm{g}(D) = 3$. This overdensity criterion provides a particularly simple description of tidal stripping,

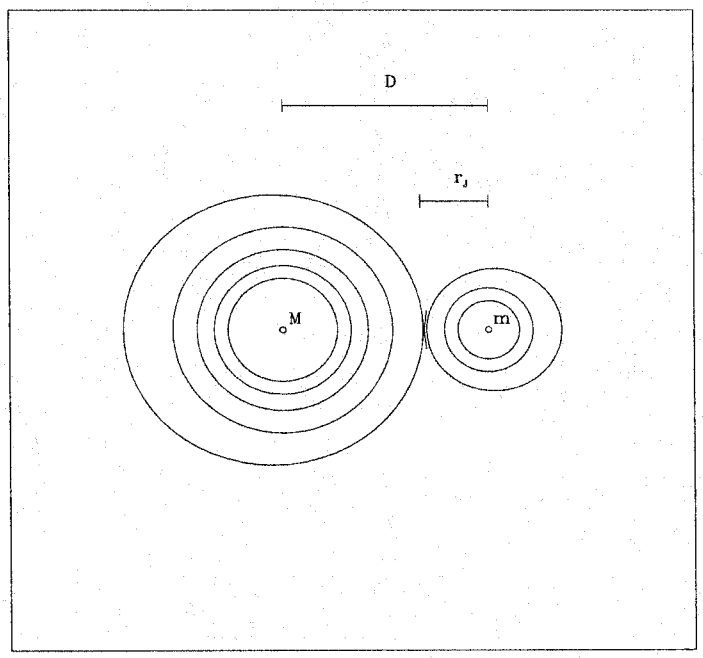

Figure 3.1 A schematic illustration of the effective potential in a binary system (a small satellite of mass m in a circular orbit around a larger system of mass M). The Jacobi limit r_J corresponds to the distance from m to the interior saddle point in the effective potential.

well suited to comparison with numerical results. The internal orbital period of the satellite at its outer radius, $t_{\rm int}(r_J)$, can similarly be related to the orbital period of the satellite within the larger system, $\tau_{\rm orb}$:

$$t_{\rm int}(r_J) = \eta^{-1/2}\tau_{\rm orb}. \tag{3.5}$$

This correspondence hints at some of the more complex physical processes at work in tidal stripping, namely resonances between the tidal forces of the external system and the internal orbital structure of the satellite.

We can also consider several useful generalisations of the Jacobi limit. First of all, the derivation above amounted to equating the binding energy of the satellite at some radius r to the work done by the tidal field of the larger system. Doing this in

an inertial frame gives:

$$-\frac{Gm}{r} = \Delta F\, r = \frac{\partial F}{\partial D}\frac{\partial D}{\partial r} r^2 = -\frac{2GM}{D^3} r^2, \quad \text{or} \quad \frac{m}{r^3} = \frac{2M}{D^3} \tag{3.6}$$

that is the overdensity η is 2 in this case, because the additional contribution of the differential centrifugal force has not been included. In practice, this term increases mass loss and raises the overdensity on circular orbits, and reduces it on very hyperbolic orbits. On a general orbit, the additional term, equal to 1 for a circular orbit, turns out to be ω^2/ω_c^2, where ω is the instantaneous angular velocity of the satellite and ω_c is the instantaneous angular velocity of a circular orbit at the same radius. For satellites which are not co-rotating with their orbital motion, ω^2 in this expression can be replaced by $\omega_{\rm orb}^2$, the square of the angular velocity in the non-rotating frame. Finally, we see from the second derivation above how to extend the Jacobi limit to systems orbiting within an extended potential. In this case, including the appropriate expression for the tidal force introduces a derivative of the force of the main system with respect to r, or equivalently a second derivative of its potential Φ. Thus the most general expression for the overdensity is:

$$\eta = \left(\frac{\omega^2}{\omega_c^2} - \frac{1}{\omega_c^2} \frac{{\rm d}^2\Phi}{{\rm d}r^2} \right). \tag{3.7}$$

This is close to the original expression derived by King (1962).

Even with these corrections and generalisations, the Jacobi limit is still approximate in several ways. First, the actual shape of the iso-potential contours in the combined system considered originally is not strictly spherical; Innanen, Harris and Webbink (1982) have calculated, for instance, that the length of the short axis will be approximately $(2/3)^{1/2}$ smaller than the distance to the saddle point calculated above. Furthermore, $E_{\rm J}$ is only one integral of motion; for many general orbits, there exist other, non-classical integrals of motion that can confine particles with apocentres beyond $r_{\rm J}$ to the vicinity of the satellite. The overall phase-space density of trapped orbits does drop quickly around $r_{\rm J}$, however, so the Jacobi limit provides a

good estimate of the limiting radius of the system (see BT, p. 452, and references therein, for further discussion of this point). Finally, as discussed in the following section and the next chapter, it is not immediately obvious whether the Jacobi limit can be extended to non-static potentials, such as those experienced by satellites in non-circular orbits.

3.2.3 Comparison with Numerical Mass-loss Rates

These reservations aside, how well does the tidal limit approximation describe mass loss in simulations? The expressions derived above can be extended to non-circular and decaying orbits in a naive way, by taking the potential and angular velocity in equation (3.7) to be the instantaneous values characterising the motion of the centre of mass of the satellite. Since this expression was derived for a spherical potential, the axisymmetry of the disk must also be accounted for. The most straightforward way to do this is to average over the asphericity of the potential due to the disk component, setting:

$$\frac{\mathrm{d}^2\Phi}{\mathrm{d}r^2} = \frac{\mathrm{d}^2\Phi_{\mathrm{sph}}}{\mathrm{d}r^2} = \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{-GM(<r)}{r^2}\right) . \tag{3.8}$$

where Φ_{sph} is the potential produced by a spherically symmetric distribution with the same mass $M(<r)$ interior to r. This will be very close to the radial gradient of the actual force on the satellite when it is far from the disk, or when it is in the plane of the disk. Only when the satellite is close to the disk, but on an inclined orbit, will the true gradient differ substantially from this value, and in practice tidal shocking should dominate the physics of the mass loss in these cases, as discussed in the next chapter.

Given these conventions, figure 3.2 shows the bound mass determined by VW in simulations considered in chapter 2, compared with the bound mass estimated from the tidal-limit approximation, if all the material beyond the satellite's instantaneous tidal limit is considered to be immediately and permanently unbound. (The Coulomb

logarithm used to calculate these orbits has been held constant, even though the mass of the satellite decreases; the effects of a changing logarithm will be considered in the next chapter). We see that while this prescription for mass loss reproduces the sharp drop in the bound mass at each pericentric passage, the overall rate of mass loss is overpredicted; on the first pericentric passage, for instance, the satellite generally loses only a quarter to a third of the mass expected from the tidal-limit approximation. Furthermore, this result still holds even if a different definition of the overdensity used; the dotted curves show the same results for a constant overdensity $\eta = 2$. In this case the agreement on the first pericentric passage is improved, but mass loss is still overpredicted in general.

We can get some indication of the cause of this overprediction by examining the dependence on orbital properties. We see that the estimate considered here is more accurate for more circular orbits. For radial orbits, the mass-loss rate is predicted to be much higher than what is seen in the simulations. This is understandable, given that these systems only spend a short time near the pericentre of their orbit, and then rapidly return to regions with smaller tidal fields. Unfortunately, typical subhalo orbits seen in cosmological simulations are fairly radial; Ghigna *et al.* (1998) find an average apo-to-pericentre ratio of 6:1, for instance. Thus it is precisely these cases which the analytic model should describe best. In the next section, I will consider a mass-loss model which does not assume circular, unchanging orbits, in an effort to improve the accuracy of the model.

3.3 The Impulse Approximation

On the radial orbits typical of cosmological simulations, satellites in a spherical potential pass only very briefly through the pericentre of the orbit, as mentioned above. During the pericentric passage, material within a satellite will experience

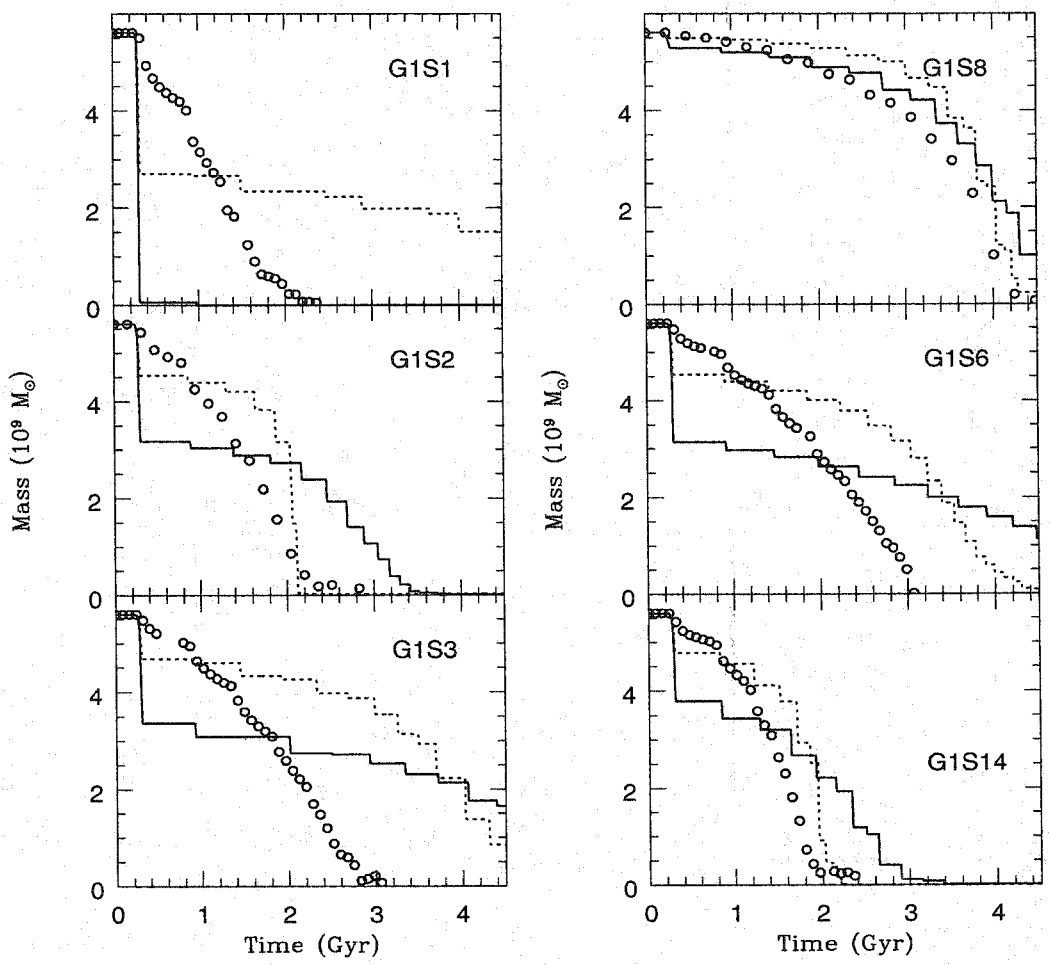

Figure 3.2 Mass-loss rates predicted by truncating the satellite at its instantaneous tidal limit. The solid curves show the analytic prediction for the overdensity described in the text, while the dotted curves show the same results for a constant overdensity $\eta = 2$. The points are the numerical results of VW.

a rapidly fluctuating tidal force, which accelerates individual particles within the satellite with respect to its centre of mass. Before and after pericentric passage, on the other hand, tidal forces on the satellite will be minimal. If one assumes that the motion of particles within the satellite is negligible during the course of the acceleration, an assumption commonly called the *impulse approximation*, then the

energy change for a particle of unit mass, ΔV, can be written:

$$\Delta E = \mathbf{V} \cdot \Delta \mathbf{V} + \frac{1}{2}\Delta V^2. \tag{3.9}$$

Between pericentric passages, ΔV will normally be uncorrelated, such that $\mathbf{V} \cdot \Delta \mathbf{V} = 0$. Thus, only when $\Delta \mathbf{V}$ is comparable to the escape velocity of the system over a single pericentric passage will the first term in equation (3.9) be important, unbinding the particle completely as the satellite passes through the pericentre of its orbit. The contribution from the second term, on the other hand, will add up from one orbit to the next, producing systematic heating of the satellite. In the limit of brief encounters, Allen and Richstone (1988) get a partial estimate of impulsive mass loss by ignoring the heating term, and determining how many particles are directly unbound by the first term alone. Equating $\Delta \mathbf{V}$ with the circular velocity at radius r, they obtain a condition on the outer radius of the satellite:

$$\frac{Gm}{r} \lesssim (\tau_{\rm enc} \Delta F_{\rm tid})^2 \tag{3.10}$$

similar to the tidal limit derived above (cf. eq. 3.6). For a brief encounter, the right-hand side of this equation will be small, so the tidal limit will be larger than the Jacobi limit, and mass loss will be reduced correspondingly.

This approach provides one possible method for determining mass loss on simple, radial orbits in a spherical potential, where mass loss occurs mainly at the pericentre of the orbit (HN). In particular, equation (3.10) assumes that before and after the encounter, the satellite is isolated, in the sense that tidal forces on the satellite are negligible. In a typical galactic potential, however, a satellite may experience several shocks in rapid succession as it passes through the disk and through the pericentre of its orbit[1] . To determine mass loss in these more general cases requires a method which can account for continuously changing tidal fields. Equation (3.10) also depends

[1]Dynamical friction also causes the orbital period to change, making it less convenient to determine the timing of these shocks.

on the validity of the impulse approximation, and neglects the long-term effects of the second term in equation (3.9). I will address these points when discussing tidal heating in chapter 4.

3.4 A General Model for Mass Loss

The tidal limit is a steady-state solution to the problem of mass loss on an orbit with a fixed pericentre, but typical satellite orbits do not spend enough time near pericentre to achieve this steady state. In the method of Allen & Richstone (1988), the predicted mass-loss rate therefore depends not only on the strength of the tidal field at the pericentre of the orbit, but also on the duration of the passage through pericentre. The relationship between the tidal limits derived in either case (equations (3.10) and (3.3) or (3.6)) suggests a more general description of tidal stripping that interpolates between these two cases. One way of deriving a more general model is to make the ansatz that a system initially out of equilibrium, in the sense that it extends beyond its tidal limit, will evolve transiently towards a steady state, where it is stripped to its tidal limit, over a timescale equal to the orbital period. If one supposes that transient behaviour, which determines the evolution from the initial to the final state, is simply an exponential decay of the difference between the two states, then as a satellite evolves from the initial state $X_0 = X(t = 0)$ to the final state $X_1 = X(t = \infty)$, its state at any time will be:

$$X(t) = X_0 + (X_1 - X_0)(1 - \exp(-t/t_{\rm char})), \tag{3.11}$$

with $t_{\rm char} \simeq t_{\rm orb}$. Thus for over short timestep Δt, we can write:

$$\delta X = X(t + \Delta t) - X(t) \simeq -\Delta X(\Delta t/t_{\rm orb}), \tag{3.12}$$

where $\Delta X \equiv X_1(t) - X(t)$ is the difference between the actual state of the system and the state it would achieve if it remained in that tidal field indefinitely.

Of course this general argument does not specify whether X should be the mass of the system, or its density, or its outer radius, or some other property such as the energy of the system, nor is there any guarantee that the timescale for the transient evolution is exactly $t_{\rm orb}$, as several related times, such as the dynamical time of the satellite or the internal orbital period at its outer edge or at the half-mass radius characterise the system equally well. I will show in this chapter and the next that taking X to be the mass of the system, and $t_{\rm char}$ to be the instantaneous orbital period of the system, yields results in excellent agreement with the simulations. I note however that this solution may not be unique, and the relative accuracy of different schemes is somewhat obscured by the freedom one has in choosing the Coulomb logarithms and the heating coefficient described in the next chapter. Further analysis is required to establish the theoretical justification for this mass-loss model; in practice, however, it is easy to implement and accurately represents the simulations, so I will use it as a general description of mass loss.

In detail then, the method is as follows:

- In each timestep, one calculates the mass loss ΔM predicted by the tidal stripping approximation, that is the total mass the system would lose if it remained indefinitely on a circular orbit in a tidal field of that strength.
- Assuming that this mass is lost as described above, the corresponding change over a short interval of time Δt is:

$$\delta M = \Delta M \exp(\Delta t/t_{\rm orb}) \simeq \Delta M(\Delta t/t_{\rm orb}). \tag{3.13}$$

Thus in this model, a satellite on a circular orbit will lose all the mass outside its tidal radius in one orbital period, while a satellite on a stable, eccentric orbit will lose a constant fraction of this mass, roughly proportional to the time spent close to pericentre, at each pericentric passage. In this sense, the model reproduces the results derived above in these two limiting cases, while for a realistic orbit, decaying due to dynamical friction, the mass loss will vary over the course of a single orbital

period, peaking as the satellite passes through pericentre, and also changing from one orbital period to the next, as the radius of the pericentre decays.

Finally I note that while this prescription predicts the total mass of the satellite as a function of time, it does not specify its size. It is not clear how the density profile of the system evolves as the satellite loses mass. From energy distribution arguments, mass will preferentially be stripped off the outside of the satellite, but its actual size at any time will also depend on tidal heating, and therefore I postpone a discussion of this point to chapter 4.

3.5 Comparison with Simulations

Finally, how well does this model reproduce the behaviour seen in simulations? Figure 3.3 shows the bound mass for VW orbits G1S1–G1S3, G1S6, G1S8 and G1S14. The prediction of the total mass loss up to the first pericentric passage is now greatly improved (compare the solid lines with figure 3.2). For G1S2, G1S8, and G1S14, the overall mass-loss rate is also fairly accurate, while for G1S1, G1S6, and G1S14, it is somewhat too slow. This may indicate that the characteristic timescale for mass loss, taken to be the instantaneous orbital period in this case, is in fact overestimated, but choosing a shorter time does not necessarily improve the estimate, as shown by the dashed lines on the same figure. Thus, it appears that some process beyond the passive tidal stripping described in this chapter accelerates mass loss, and furthermore that this process is related to passage through the disk. Not only does the satellite lose mass suddenly when it crosses the disk; we can also see that its mass loss rate (the slope of the curve) decreases slowly over time, after each disk crossing.

The simple model proposed has no memory; a satellite moving through the disk into a low-density region with a weak tidal field will forget its encounter and behave subsequently as if its internal state had never changed. In reality, rapid encounters

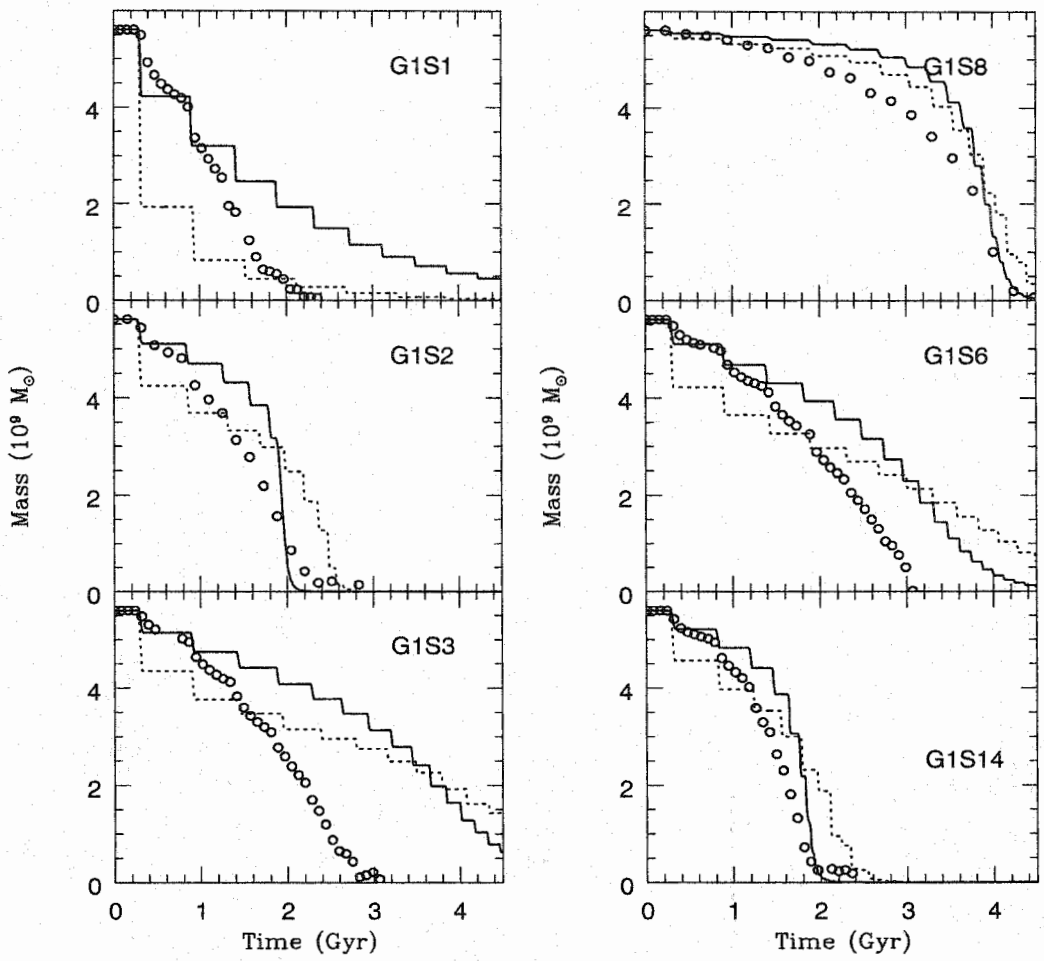

Figure 3.3 As figure 3.2, for the general mass-loss model, in which the mass lost in each timestep is scaled to the instantaneous orbital period of the system (see text).

will permanently change the energy distribution within a satellite, making it more susceptible to subsequent mass loss (the effect comes from the second term in equation (3.9)). I refer to this process as tidal heating, and examine it in detail in the next chapter.

3.6 Summary

In this chapter, I have discussed the need to include mass loss in studies of satellite dynamics, in order to correctly predict orbital decay times, to study the effects of satellites on disks, and to explore other problems such as the distribution and properties of tidally stripped material, some of which may account for the stellar halo of the Galaxy. I describe the two extremes in which mass loss has been modeled analytically, the tidal stripping approximation, a steady state solution that applies to systems on circular orbits, and the impulse approximation, which describes very rapid changes in otherwise isolated systems. Since real satellites are neither isolated nor in a steady state, I construct a general model which interpolates between the two cases, reproducing the expected behaviour in either limit. I show that this model offers a good description of the mass loss seen in simulations, for orbits in a smooth potential, but that disk crossings in a realistic galactic potential will accelerate mass loss, in ways not accounted for by tidal stripping. This 'tidal heating' will be discussed in the next chapter.

Chapter 4

Tidal Heating

In the previous chapter, we saw that rapid gravitational shocks, such as those induced by the passage of a satellite through the plane of the disk, can produce changes in its mass not accounted for by simple tidal mass loss. In this chapter I present a model of tidal heating, based on the theoretical work of Gnedin, Ostriker and collaborators, which accounts for this effect. I derive the increase in energy per unit mass produced by tidal forces in the outer part of the satellite, and show how this change in energy modifies the phase-space distribution of satellite material. These changes can be included in the model of mass loss described in chapter 3 by using an 'effective density', which includes the effects of heating, when calculating the tidal limit of a system. I show that this model does an excellent job of matching simulation results, for a reasonable choice of the one free parameter, the heating coefficient $\epsilon_{\rm h}$. Since it is often of interest to know how the physical size of the satellite changes as well, however, I also propose an approximate description of the structural changes produced by heating, based on the simulation results of HN. While the proposed behaviour matches the simulations to first order, several higher-order effects, some transient and some systematic, are clearly visible in the numerical results. The structural changes observed in these simulations require further study, but the model proposed here does at least give a reasonable indication of how the peak circular velocity will change with mass loss, for instance. Finally, I summarise the results of the first part of this thesis, describing the full analytic model of satellite evolution that has been presented in chapters 2–4.

While a slowly varying tidal field produces steady mass loss through tidal stripping, as discussed in the previous chapter, rapid tidal shocks, such as those produced by the passage through a thin, dense disk, will preferentially rearrange the internal

structure of a satellite, leaving it more prone to subsequent mass loss. This 'memory' effect is not accounted for in the theory of tidal mass loss introduced in the previous chapter, where mass loss in any timestep is more or less independent of the previous history of a satellite of given mean density. It is my goal in this chapter to account for the irreversible heating of a satellite produced by tidal shocks, in a way that complements the work in the last chapter on tidal stripping. It is worth noting that at a rigorous level these two processes form part of a continuum; in both cases, the application of a tidal field, varying in time and space, to the initial distribution function of the satellite, produces direct changes and subsequent readjustment of the distribution, leading in particular to a smaller total mass of material being bound. Tidal stripping from the steady-state or impulse approximation models can be distinguished from tidal heating on the basis of the timescale involved; tidal shocks occur over timescales short compared to the internal dynamical time of the satellite, impulsive tidal stripping on timescales short compared to the orbital period of the satellite, and steady-state tidal stripping on even longer timescales. In this regard, it is the existence of thin, dense, galactic disks, with associated crossing times that are very short for typical satellite orbits, that motivates a separate treatment of tidal heating, the most rapid of these disturbances.

4.1 Theory

4.1.1 The First and Second-order Heating Terms

The basic theory of tidal shocking was developed initially by Ostriker, Spitzer & Chevalier (1972) and is summarised in Spitzer (1987). It has recently been studied extensively by Gnedin, Ostriker and collaborators in a series of papers (Kundić & Ostriker 1995; Gnedin & Ostriker 1997, 1999; Gnedin, Hernquist & Ostriker 1999). Spitzer (1958) first recognised that the effect of brief, abrupt changes in the potential

of the background system can be modeled by taking them to be instantaneous, with respect to the orbits of particles within the satellite. In general, a fluctuating tidal acceleration $\mathbf{A}(x,t)$ produces the velocity change

$$\Delta\mathbf{v} = \int \mathbf{A}(x(t),t)dt \tag{4.1}$$

on a particle whose position x varies during the course of the acceleration. In the impulse approximation, the change in position is neglected, so the net acceleration on a particle initially at position x_0 is simply $\mathbf{T}(x_0)dt$ over the interval dt, where $\mathbf{T}(x_0)$ is the tidal acceleration at position x_0, that is the difference in gravitational acceleration $\mathbf{g}$ between the centre of the satellite and x_0:

$$\mathbf{T}(x_0) = \mathbf{g}(x_0) - \mathbf{g}(0). \tag{4.2}$$

If the satellite is small compared with the characteristic scale of the background potential (that is, the scale over which it varies appreciably), as it must be for the tidal approximations developed in the previous chapter to be valid, then we can write

$$\mathbf{T}(\mathbf{x_0}) = \mathbf{x}_0 \cdot \nabla\mathbf{g}|_{bfx=0} = g_{a,b}x_b\mathbf{e}_a, \tag{4.3}$$

where $g_{a,b} = \partial g_a/\partial x_b$, $\mathbf{e}_a$ is the unit vector in the x_a-direction, and repeated indices indicate summation over the three Cartesian coordinates. Over a single rapid shock (say a passage through pericentre, or through the plane of the disk), a particle has its energy per unit mass change by

$$\begin{aligned} \Delta E &= \frac{1}{2}\left((\mathbf{v}+\Delta\mathbf{v})^2 - \mathbf{v}^2\right) = \mathbf{v}\cdot\Delta\mathbf{v} + \frac{1}{2}\Delta\mathbf{v}^2 \\ \text{where } \Delta\mathbf{v} &= \int \mathbf{x}(t)\cdot\nabla\mathbf{g}\ dt. \end{aligned} \tag{4.4}$$

If the velocity distribution at any point within the satellite is symmetric about $\mathbf{v} = 0$, then the first term will vanish for a brief shock where $\mathbf{T}$ has a roughly constant direction. Thus only the second term remains:

$$\Delta E = \frac{1}{2}\int \mathbf{x}(t)\cdot\nabla\mathbf{g}\ dt\ \cdot \int \mathbf{x}(t)\cdot\nabla\mathbf{g}\ dt \tag{4.5}$$

In the case of a small, spherical satellite, averaging over a shell of radius r gives:

$$< \Delta E >= \frac{1}{2} < (\Delta v)^2 >= \frac{1}{6} \left(\frac{d\mathbf{g}}{dr} \right)^2 r^2 \Delta t^2 \tag{4.6}$$

for a brief shock of duration Δt. This is the first moment of the change in energy for particles at radius r, that is the change in the average energy. Kundić and Ostriker (1995) also calculate the higher moments of the change in energy. The second moment, that is the change in the dispersion of energies at radius r, is:

$$< (\Delta E)^2 >= \frac{1}{3} < v^2 (\Delta v)^2 > + \frac{1}{4} < (\Delta v)^4 > \tag{4.7}$$

this can be comparable to the first order term in its effects, while the third and higher order terms are negligible.

4.1.2 Adiabatic Corrections

It is also possible to account for some of the limitations of the impulse approximation within this same framework. The expressions for the heating terms, equations (4.6) and (4.7), were derived assuming that the particles are stationary for the duration of the shock. In fact, the internal orbital times of particles within the satellite can often be similar to the shock timescales. The orbital motion of the particles then damps out the net tidal acceleration over the course of the shock. In the limit where the internal orbital times are short compared to the shock timescale, the system responds adiabatically to shocking, and it has no effect on the satellite. To account for this damping effect, one can introduce an adiabatic correction A, which describes the ratio of the energy change to the change predicted by the impulse approximation. This correction will depend on the ratio of the shock timescale t_s to the orbital time of a particle at radius r within the satellite $t_{\rm orb}(r)$, $x \equiv t_s/t_{\rm orb}(r)$, called the *adiabatic parameter* (Spitzer 1987). Gnedin and Ostriker (1997, 1999) find that the first and second-order adiabatic corrections are well approximated by:

$$A_1 \;=\; \Delta E/\Delta E_{\rm impulse} = (1+x^2)^{\gamma_1}, \tag{4.8}$$

$$A_2 \;=\; (\Delta E)^2/(\Delta E_{\rm impulse})^2 = (1+x^2)^{\gamma_2}, \tag{4.9}$$

with the coefficients $\gamma_1 \simeq 3/2$ for slow shocks (those where $x \simeq 4$ at the half-mass radius of the satellite) and $\gamma_1 \simeq 3/2$ for fast shocks ($x(r_{\rm hm}) \simeq 1$.

4.2 A General Model for Tidal Heating

While the formalism of Gnedin, Ostriker and collaborators provides a detailed description of tidal heating, and has been verified by comparison with numerical simulations, it requires a few modifications before it can be used in the analytic model of satellite evolution proposed here. Specifically, the analytic model is general, so the form of the potential and the location of pericentric passage are not known in advance. Thus the effect of heating on mass loss must be broken down into a series of the changes, produced step by step as the satellite moves through the potential. Furthermore, it would be prohibitively expensive in computational terms to keep track of the energy distribution within each satellite, so heating must be related to global changes in satellite properties. Consequently, the model requires a scaling relation to relate mass loss to the change in internal energy of the satellite. I will describe these two changes in detail below.

4.2.1 The Discrete Heating Calculation

First, consider the effect of heating described by equation (4.5) above. If we divide the shock into a series of n discrete time steps of equal length Δt, then the work done is:

$$W_{\rm tid}(t_n) = \frac{1}{2}\,\Delta t^2 \left[\sum_{i=0}^{n-1} \mathbf{T}_{\rm tid}(t_i) \cdot \sum_{j=0}^{n-1} \mathbf{T}_{\rm tid}(t_j)\right]. \tag{4.10}$$

In going from t_n to t_{n+1}, the energy change in a single timestep is therefore:

$$\Delta W_{\rm tid}(t_n \rightarrow t_{n+1}) = \frac{1}{2}\,\Delta t^2 \mathbf{T}_{\rm tid}(t_n) \cdot \left[2\sum_{i=0}^{n-1} \mathbf{T}_{\rm tid}(t_i) + \mathbf{T}_{\rm tid}(t_n)\right]. \tag{4.11}$$

Finally, using equation (4.3), we can rewrite this as:

$$\Delta W_{\rm tid}(t_n \rightarrow t_{n+1}) = \frac{1}{6}\, r^2\, \Delta t^2 \left[2\, g_{a,b}(t_n) \sum_{i=0}^{n-1} g_{a,b}(t_i) + g_{a,b}(t_n) g_{a,b}(t_n) \right] \qquad (4.12)$$

with eighteen terms from the summation over a and b, where we have used the fact that

$$\langle x_i x_j \rangle = \frac{1}{3} r^2 \delta_{ij}$$

averaged over a sphere[1] . This discrete expression will allow the code to track the energy of the satellite as it evolves in its orbit.

4.2.2 Corrections to the Heating Rate

There are two further corrections required of equation (4.12). First, the adiabatic correction mentioned above will reduce the change in energy over the course of a shock, if its duration is long compared to the internal dynamical time of the satellite. This can be accounted for by using equation (4.8) to reduce the energy change above, replacing $\mathbf{T}$ with $\mathbf{T}(1 + x_{\rm hm}^2)^{\gamma_1/2}$. This makes the approximation that the orbital period at the half-mass radius characterises the response of the whole system over the course of the shock, which should be the case provided the shock is brief and the mass of the system does not change radically over this time. More importantly, however, the system will also readjust itself between subsequent shocks. Thus, the sum in equation (4.12) should only be calculated over the duration of a single, coherent shock of duration comparable to or less than $t_{\rm orb}(r_{\rm hm})$; otherwise the increased mass loss due to heating will include some of the effect of the steady tidal field, already accounted for in the tidal stripping calculation described in chapter 3. Thus shocks must be identified in a general way, as they occur in the orbit. I do this using the same ratio of timescales that appears in the adiabatic correction; shocks produce a coherent effect if the characteristic timescale of the shock is short compared to the

[1]This is only strictly true if the shock is rapid, so that $x_i(t_0) \simeq x_i(t_j) \simeq x_i(t_n)$ for all $t_j < t_n$.

orbital times of material in the satellite. Specifically, the satellite is heated when $t_{\rm shock} < t_{\rm orb,sat}$, where $t_{\rm shock} \equiv (t_{\rm sh,d}^{-1} + t_{\rm sh,b}^{-1})^{-1}$ is the harmonic mean of the disk and bulge shock times $t_{\rm sh,d} = z/V_{z,\rm sat}$ and $t_{\rm sh,b} = r/V_{\rm sat}$, and $t_{\rm orb,sat} = 2\pi r_{\rm h}/V_{\rm c}(r_{\rm h})$ is the orbital period of the satellite at its half mass radius. This criterion corresponds to that considered by Gnedin and Ostriker (1999) for short shocks, such as the disk shocks which dominate in the VW simulations considered here. The corresponding adiabatic correction has an exponent $\gamma_1 = 5/2$. In potentials without a disk, the values for long shocks may be more appropriate. This is the case, for instance, with the simulations of HN.

Second, heating also leads to a change in the internal velocity dispersion of the satellite, as discussed by Kundić & Ostriker (1995). Both the average energy gain and the increase in the dispersion will drive mass loss. In keeping with the simplicity of the analytic approach, I will only compute the first-order change in the energy distribution, and account for the higher-order effects through the introduction of a heating coefficient, $\epsilon_{\rm h}$, that I will adjust to yield reasonable overall matches to the numerical mass-loss rates:

$$\Delta E = \epsilon_{\rm h} \Delta E_1 \tag{4.13}$$

where

$$\Delta E_1 = A_1(x) \Delta E_{1,\rm imp} = A_1(x) \Delta W_{\rm tid} \ .$$

Kundić & Ostriker (1995) estimate that the second-order heating term has an effect comparable to or greater than that of the first-order term, so $\epsilon_{\rm h}$ should be greater than 2. From the disruption timescale arguments in Gnedin & Ostriker (1997), for instance, one might expect that $\epsilon_{\rm h} \simeq 7/3$. The value of $\epsilon_{\rm h}$ used in practice, however, will also depend on the shocking criterion and the adiabatic parameters discussed previously.

4.2.3 Relating Heating and Mass Loss

Having described the energy change produced by heating in a way that can be implemented step by step in the orbital calculation, it remains to relate this energy change to mass loss. Mass loss is difficult to model in physical coordinates, because changes in the bound mass are really the result of changes in the energy distribution of the satellite. Whether these changes correspond to a spatial rearrangement of the satellite, (that is a change in the potential energies of the particles) or simply a change in its velocity distribution (that is a change in the kinetic energies of the particles) will depend on the timescale for shocking, and the dynamical timescale of the satellite. I will derive the relation between heating and mass loss in the first case, that is assuming heating causes the satellite to expand instantaneously, since this is the easiest case to understand physically. I will then show however, that this relation is general, and more or less independent of the spatial evolution of the material in the satellite. The actual behaviour of the satellite is quite different from a simple expansion, as explained in section 4.4, so it is important to have a description of mass loss that does not depend strongly on these structural changes.

Instantaneous heating produces an energy change $\Delta W(r)$ at radius r given by equation (4.12). This work will initially increase the kinetic energy of the particle, but through expansion, some of this kinetic energy will be converted to potential energy. Assuming the system returns to virial equilibrium, the total change in potential energy will be $\Delta U = 2\Delta W$ (Gnedin & Ostriker 1999). If we suppose that this additional energy causes the satellite to expand, and further more that it does so smoothly, with no shell-crossings, that is no net rearrangement of its density profile, then this energy input will cause a change in radius $\Delta r = \Delta W r^2$, since

$$\Delta(E(r)) \propto \Delta\left(\frac{-GM}{r}\right) \propto -\frac{\Delta r}{r^2} \tag{4.14}$$

if $M(< r)$ remains constant. The corresponding change in density is:

$$\Delta\bar{\rho}_r = -\frac{3\Delta r}{r}\bar{\rho}_r = \frac{9\Delta W(r)}{2\pi G r^2} = \frac{9}{2\pi G}\left(\frac{d\mathbf{g}}{dr}\right)^2 \Delta t^2 \tag{4.15}$$

Thus, if we ignore the adiabatic correction, the radial dependence has dropped out of the change in density. As explained previously, tidal stripping eventually removes from the satellite all the material below some mean density threshold. If the satellite has expanded due to heating, all the material that was originally within $\Delta\rho$ of this threshold will be removed as well, that is heating will drive additional mass loss. Furthermore, because $\Delta\rho$ is independent of r, this will be equivalent to stripping the original satellite density profile with a higher density threshold for the external tidal field. This leads to the following convenient method for including the effects of heating in mass loss. At each step, $\Delta\rho_{\rm eff}$ is calculated as before, and added to some total change in density $d\rho_{\rm eff}$, recorded over the duration of the shock. The density threshold used to determine mass loss is increased by this total change. The algorithm for mass loss described in the last chapter will then strip off a fraction of the material beyond this higher threshold. Thus both heating and mass loss can be included in each step of the orbital evolution. I note finally that the cancellation of radial dependence that occurs in the change of density produced by heating is no accident; heating, that is instantaneous changes in the tidal field, scales with radius in the same way as the long-term changes that produce steady mass loss, because the two are ultimately the same phenomenon, differing only in timescale. I treat them separately here only because the other characteristic timescale in the system, the dynamical time of the satellite, distinguishes between them.

4.3 Comparison with Simulations

4.3.1 The VW Simulations

By accounting for tidal stripping and tidal heating, the analytic model should now be able to reproduce the behaviour seen in the numerical simulations. Figures 4.1, 4.2 and 4.3 show the overall comparison, for instance, with the fifteen orbits of VW. Figures 4.1 and 4.2 show the effect of inclination on the orbital evolution, for the satellite models S1 and S2, which differ mainly in concentration. Figure 4.3 shows more or less radial orbits, as well as orbits for the more massive satellite S3. The Coulomb logarithms have roughly the same fiducial values used previously, $\ln \Lambda_h = 2.4$ and $\ln \Lambda_d = 0.4$, at the start of each run, but increase as the satellite loses mass, as discussed in chapter 2. The heating parameter has been adjusted to match the numerical mass loss rates; the value used here, $\epsilon_h = 3$, provides a good match and is in the range expected from the results of Gnedin & Ostriker (1999).

We see that the overall accuracy of the analytic predictions is greatly improved by considering tidal heating. Now the analytic model produces an excellent match for orbits out of the plane of the disk. For orbits in the plane of the disk, there are some systematic variations between the numerical and analytic results. In particular, the analytic model appears to overestimate dynamical friction, and to underestimate mass loss slightly. This presumably stems from the approximations made in deriving expressions for these phenomena, which break down when the scale of the satellite is comparable to the scale of the system in which the satellite orbits. While further adjustment of the model parameters would yield a better fit to the numerical results for the in-plane orbits, these cases are clearly unrepresentative in the general cosmological situation, so I will use the values appropriate for inclined orbits in the general model.

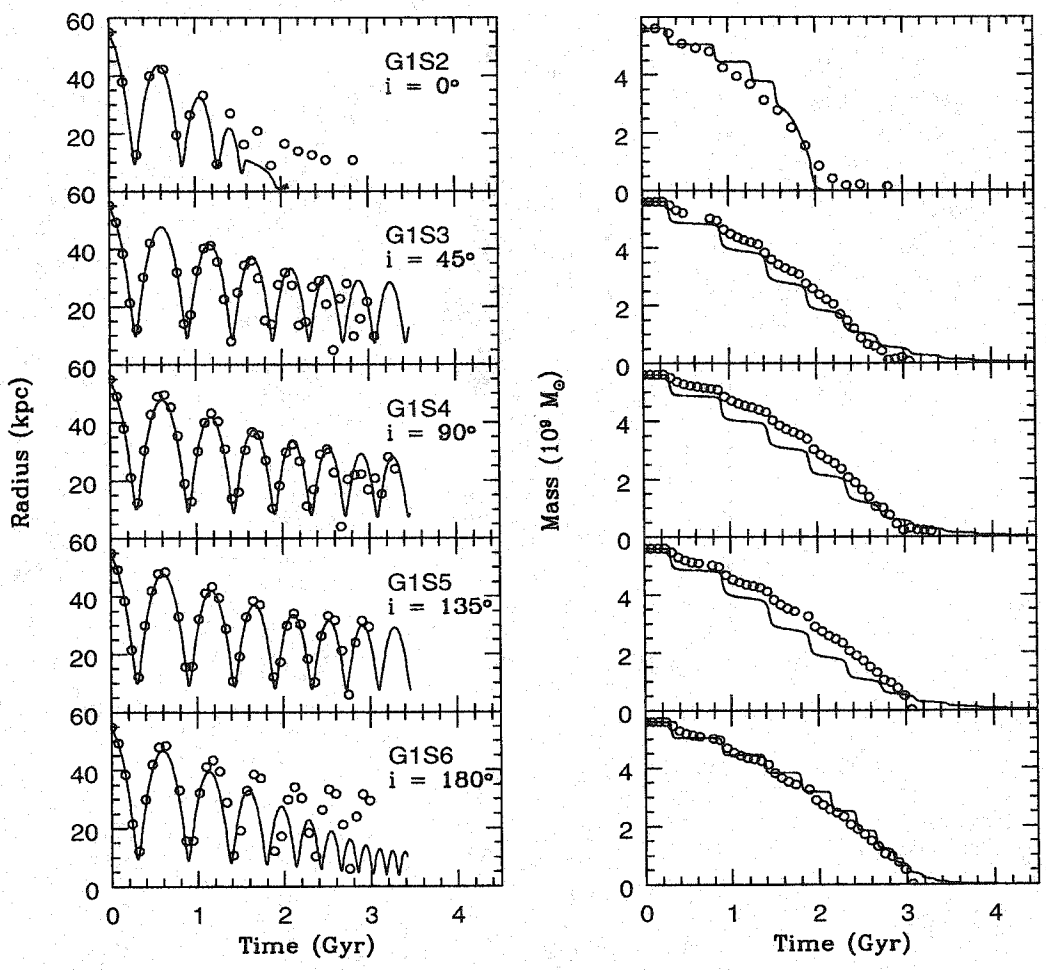

Figure 4.1 The orbital decay and mass loss predicted by the full analytic model, compared with the simulations of VW.

4.3.2 The HN Simulations

Although the main goal of this work is to reproduce realistic simulations such as those of VW, simpler simulations allow independent estimates of the overall accuracy of the subcomponents of the model. Figure 4.4 shows a comparison between the analytic predictions and the simulations of HN, which follow the evolution of a satellite in a static potential, and therefore remove the complicating effects of dynamical fiction, and of the disk. The only free parameter in this case is therefore the overall efficiency of heating, ϵ_h. Here I have used $\epsilon_h = 3.0$, the same value required to

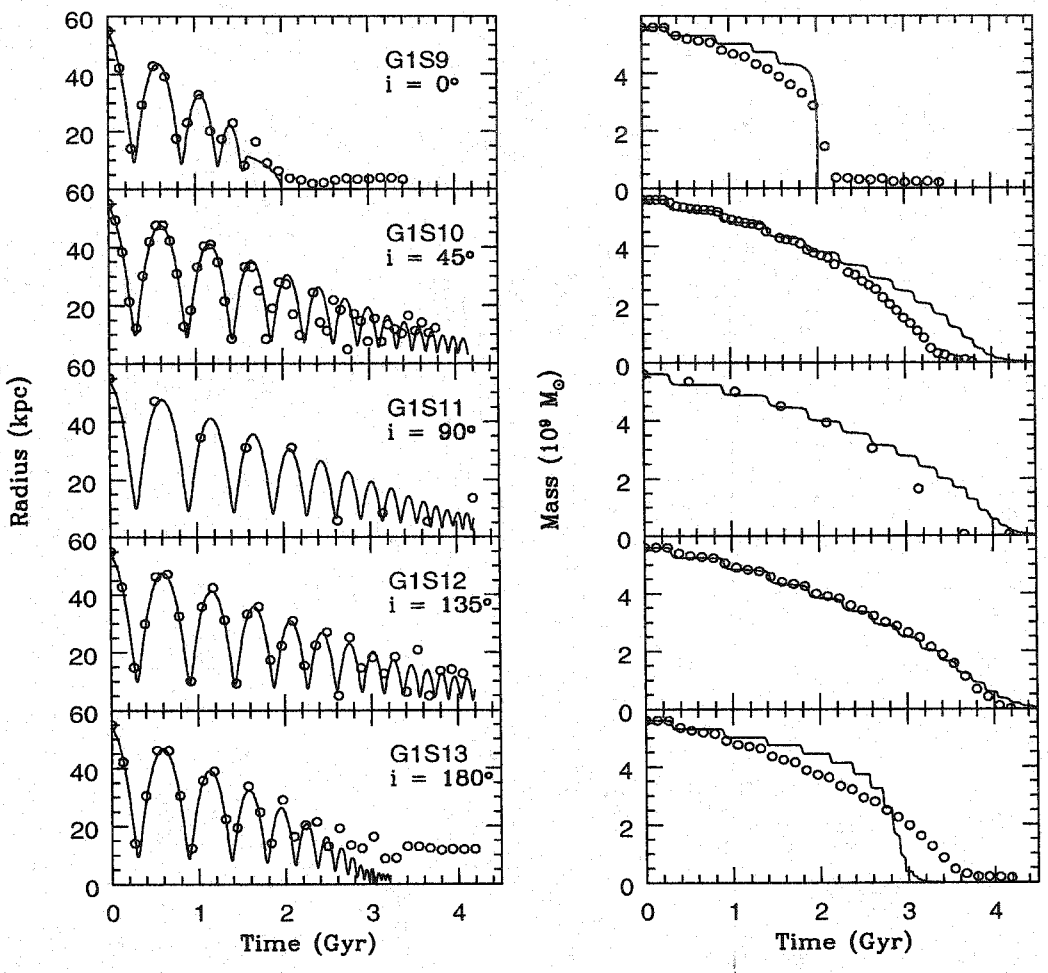

Figure 4.2 As figure 4.1, for the more concentrated satellite model S2.

match the VW simulations. We see that for this value heating parameter, the overall match to the HN simulations is excellent for all but the most extreme of these orbits, HN300:15. As noted in chapter 2, this orbit, which has an apo-to-pericentre ratio of 20:1, also shows some loss of energy and orbital decay, which may account for some of the increased mass-loss rate seen in the simulation. There are also some slight discrepancies between the numerical and analytic results for the more circular orbits HN 100:60 and 100:30, which start out at smaller apocentres, and lose mass very quickly as a result. These minor differences may be due to one of the several different approximations made in deriving heating and mass loss.

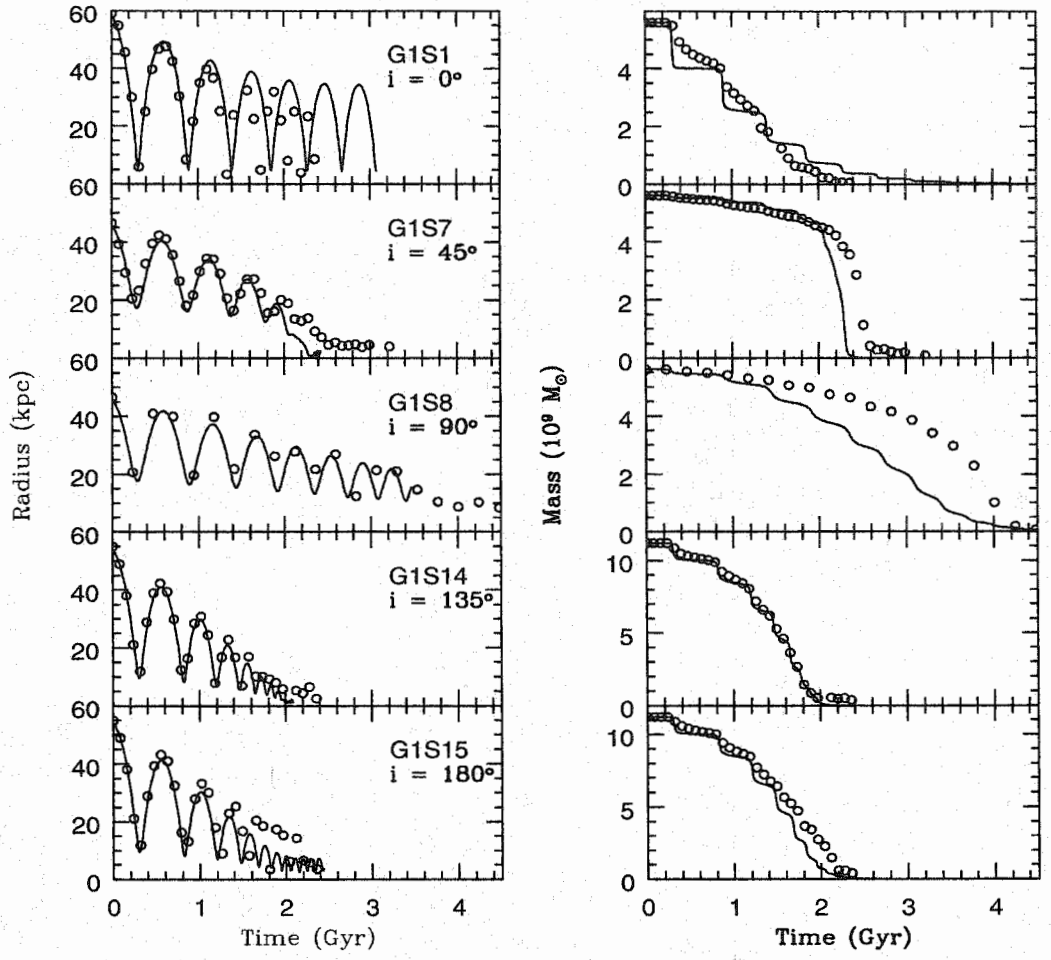

Figure 4.3 As figure 4.1, for the orbits varying in circularity, and orbits for the more massive satellite S3.

Even with these minor discrepancies, the overall estimates of infall and mass-loss rates are still remarkably reliable. For all the orbits considered here, the estimated mass and apocentre are remain within 20% of the numerical results until the satellite has lost 80–90% of its mass. Thus, the analytic model should be able to predict infall and disruption times that are accurate to within 20%, and also provide a reasonable estimate of the mass lost throughout an orbit, in order to determine how much satellite material winds up in tidal streams, for instance.

It is particularly compelling that the same heating and stripping model describes

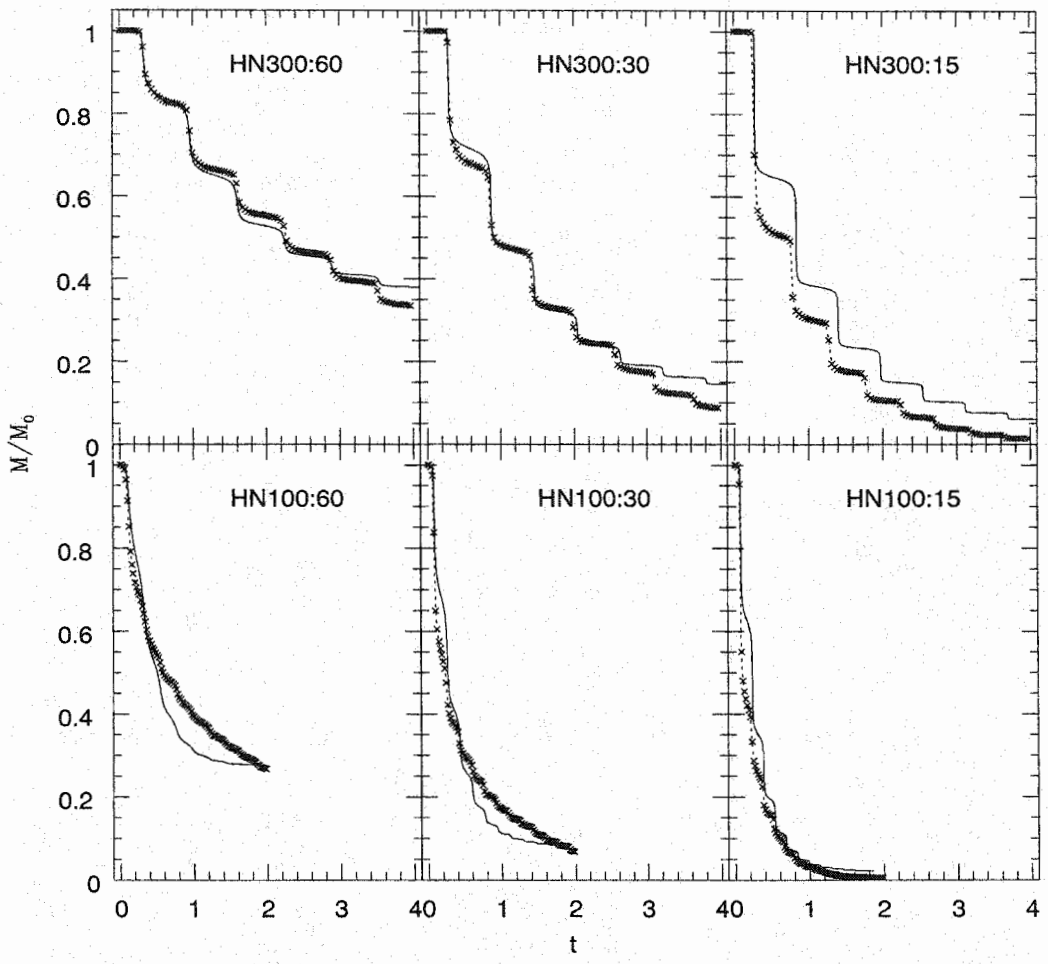

Figure 4.4 Analytic mass-loss predictions, compared with the numerical results of HN.

simulations of different satellite models in two very different potentials, with a single value of the heating coefficient. This suggests that this value is both correct and generally valid. Finally, I note that the satellites considered in the two sets of simulations are both massive and orbiting close to the centre of the potential, where tidal effects are particularly strong. They have also been followed for many orbital periods. The accuracy of the analytic model is expected to be even better in the more common case of a low-mass satellite far out in the halo, which orbits only a few times between entering within the virial radius of the main system and the present day.

4.4 Structural Changes

So far I have not discussed how the spatial distribution of material within the satellite changes, as it is heated and loses mass. The rough details of this process have been determined from many simulations, including Aguilar & White (1986), Oh, Lin & Aarseth (1995), Gnedin & Ostriker (1999), Johnston *et al.* (1999), and HN, although it is still hard to describe rigorously in simple analytic terms. First, the passage of a satellite through the disk or the pericentre of its orbit results in spatial distortion of the material that remains bound. Once it has passed through this region, unbound material escapes from the satellite, while the remaining bound material expands slightly, until the mean density of the satellite has more or less returned to its original value. Detailed simulations of this process (Gnedin & Ostriker 1999) reveal that the satellite actually undergoes several oscillations after a shock, before returning to virial equilibrium. The timescale for these oscillations is comparable to the dynamical time of the satellite. The return to equilibrium is the result of a balance between the expansion caused by the energy injected by the tidal field, and the removal of material beyond a certain radius, although the details of the process on realistic orbits remain obscure. Nonetheless, if the mean density of the satellite returns to its previous value, then the change in the outer radius, or any radius containing a fixed fraction of the satellite's mass, should vary as $r \propto M^{1/3}$. In particular, this implies that the peak circular velocity of the satellite will decrease as

$$V_{\rm p} = \sqrt{\frac{GM_{\rm p}}{r_{\rm p}}} \propto \sqrt{\frac{M_{\rm p}}{M^{1/3}}} \propto M^{1/3} \tag{4.16}$$

(where $M_{\rm p}$ is the mass within the peak radius) since this type of expansion preserves the form of the density profile, and thus the ratio $M/M_{\rm p}$.

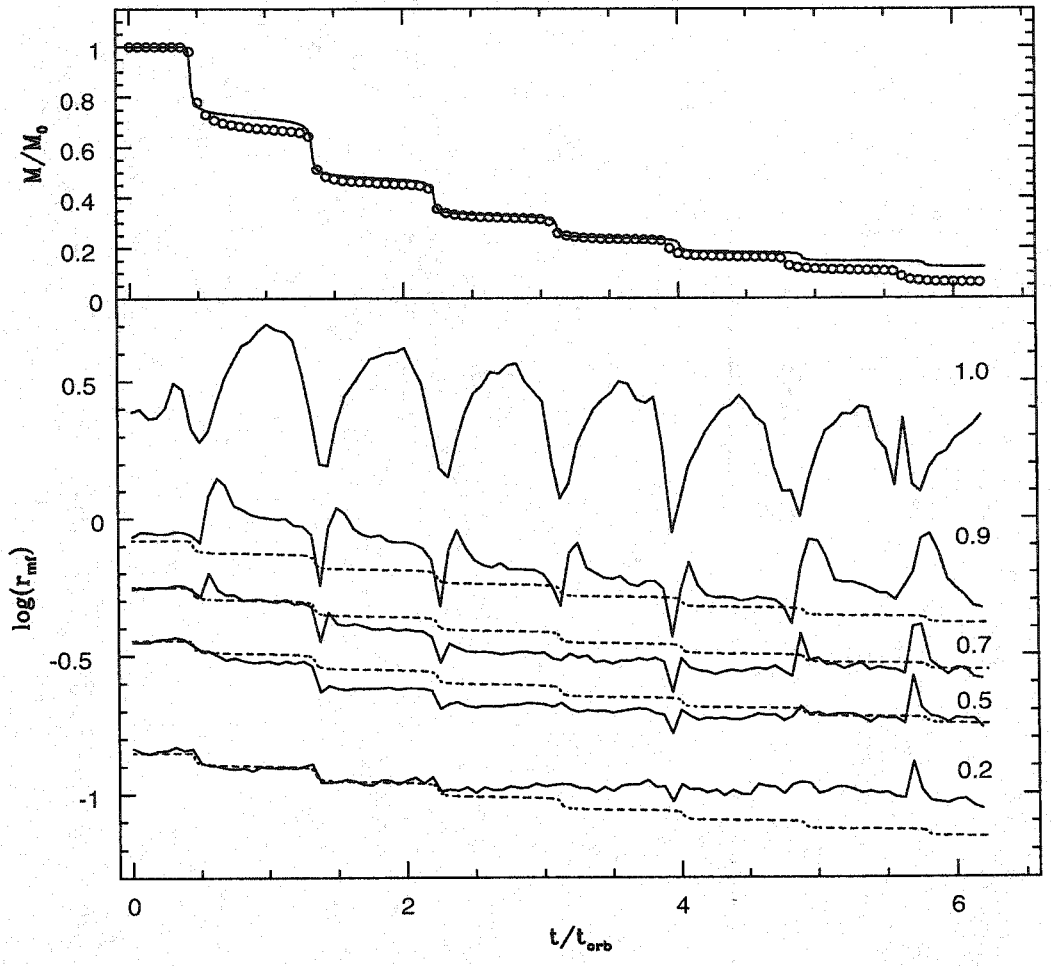

Figure 4.5 The mass-loss history for HN300:30 (top panel), and the evolution of the fractional mass radii (bottom panel). The smooth curves show the predicted behaviour in the analytic model, if the mean density within any given radius remains constant.

4.5 Comparison with Simulations

Figure 4.5 shows the evolution of several fractional mass radii, that is radii containing fixed fractions of the bound particles at any given time, for the orbit HN300:30, along with the mass-loss history shown previously. The curves in the bottom panel are labeled by the corresponding mass fraction. We see that the radius of the bound material decreases sharply as the satellite loses material at the pericentre of its orbit,

but that the radii containing fixed fractions of the bound mass then expand (at in the case of the outer radii, containing 70–100% of the mass). The overshoot and first oscillation is also visible, although the simulations are sufficiently noisy that it is hard to make out in the inner regions. This process has been studied at higher resolution by Gnedin and Ostriker (1999), and is shown in more detail in their figure 3. We also see that the timescale for the response increases with radius, as expected.

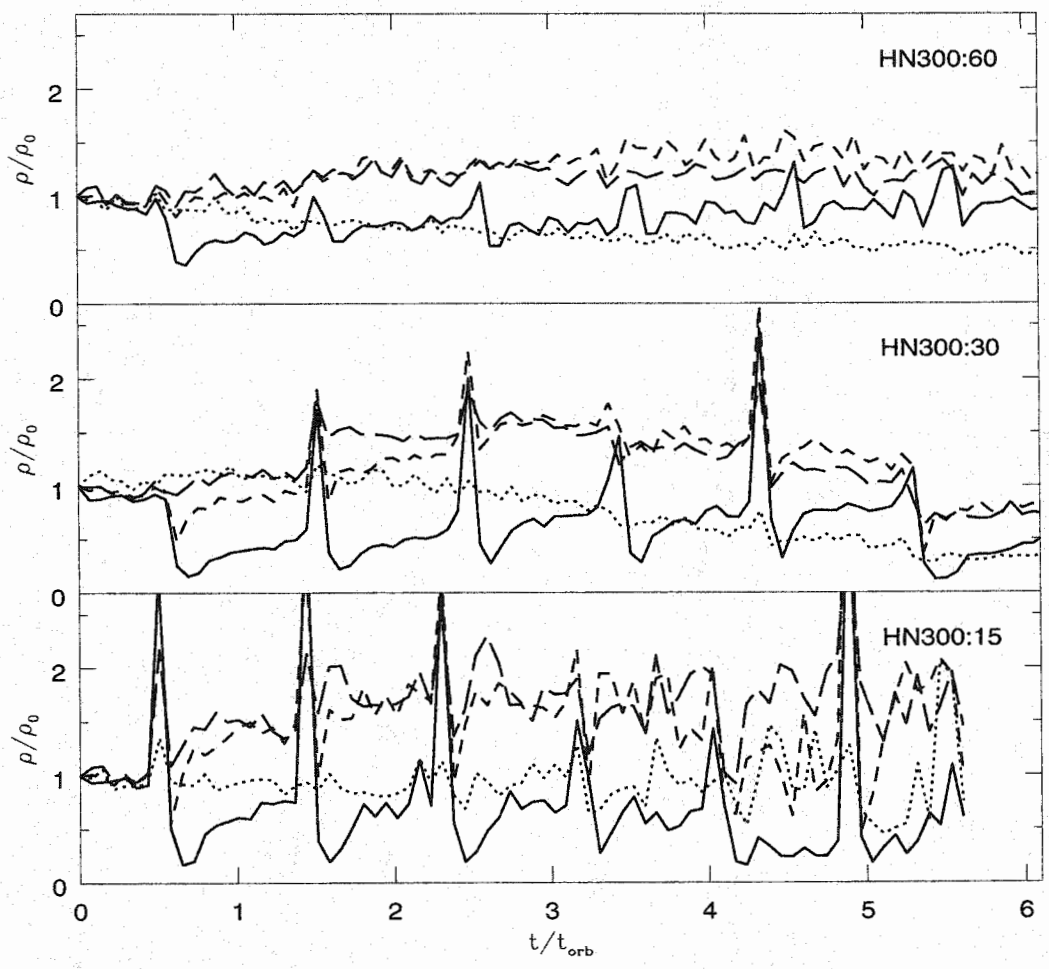

Figure 4.6 The mean density within several fractional mass radii, vs. time, for three of the HN simulations. The solid, short-dashed, long-dashed and dotted lines correspond to radii containing 90%, 70%, 50%, and 20% of the mass respectively.

The smooth curves in the bottom panel correspond to the analytic prediction,

assuming the density within any given fractional mass radius remains constant. While this estimate does not capture the full details of the evolution of the satellite's density profile, it does provide an estimate of its average behaviour. The actual density within any fractional radius, shown in figure 4.6, does not change by more than a factor of 1.5 over the first 5 orbital periods for orbits HN300:60 and HN300:30, and even for the extremely radial orbit HN300:60, the change is only a factor of 2, so prediction for the radial behaviour should accurate to within 15% on typical orbits (or to within 30% in more extreme cases). The figures also give some indication of the systematic change in the density profile with respect to this approximation. While the inner and outer radii become smaller than this estimate (solid and dotted lines), the middle radii grow systematically larger (short and long-dashed lines), producing a density profile with a flatter central core and a sharper outer cutoff. These changes may be dependent on the initial concentration and density profile of the satellite, however, and they require further investigation.

In summary, looking at satellite behaviour in simulations, we see truncation and possibly compression at each pericentric passage, followed by an expansion and oscillation around the old equilibrium density. While the details of this process require further study, the approximation that the satellite returns to its original density provides an estimate of the fractional mass radii of the satellite that is accurate to within 15% for orbits of intermediate circularity. In particular, this approximation allows one to estimate how the circular velocity of a satellite will change as it loses mass. Assuming ρ remains constant, then $V_c \propto M^{1/3}$. Given the systematic trend towards flatter density profiles mentioned above, the actual peak circular velocity may drop slightly faster than this as the satellite loses mass, but for the orbits considered the resulting error in V_c is only 8–15%. I will examine the consequences of the dependence of the peak circular velocity on mass loss further in chapter 5.

4.6 Summary of Part I

In the first part of this thesis, I have developed an analytic description of the evolution of substructure in dark matter halos, which will serve as the basis for a semi-analytic model of galaxy formation. In this description, a halo is represented by a smooth background potential, including a dark component and possibly several baryonic components, and a separate set of distinct subhalos. The latter correspond to self-bound lumps of material within the halo, the former cores of smaller halos which have merged with the larger system, and possibly the sites where dwarf galaxies may have formed. The evolution of the smooth background components is described in part II; in this first part of the thesis I have considered only the evolution of the subhalos. The latter can be inferred from dynamical arguments, and by examining numerical simulations of interactions between small companions and galaxies. The main processes that determine this evolution are described in three separate chapters; dynamical friction, the net transfer of orbital energy from the subhalo to the halo through many small collisions with background particles, tidal mass loss, and tidal heating, two related processes that result from the spatial extent of the satellite.

I account for dynamical friction by using Chandrasekhar's formula, with local values for the density and velocity dispersion, and appropriately adjusted values of the Coulomb logarithm. I determine the tidal mass-loss rate by assuming that a satellite loses the material outside its steady-state tidal limit over the course of one orbital period, an assumption that relates the impulse approximation to more common descriptions of mass loss. Finally, I include tidal heating by measuring the extent to which it accelerates mass loss, in a simple model based on Fokker-Planck studies of globular cluster disruption. The resulting description of satellite evolution depends on three main parameters, two Coulomb logarithms, $\ln\Lambda_{\rm h}$ and $\ln\Lambda_{\rm d}$, which scale dynamical friction from the halo and disk respectively, and a heating coefficient $\epsilon_{\rm h}$, which scales tidal heating. I find that a single set of parameter values, close to

those expected from theory, suffices to match the results of more than 20 different simulations of satellite evolution by two different groups. While the orbital evolution of a satellite should not depend on its radius to first order, a prediction for the size of the satellite is useful for estimating its peak circular velocity, and for comparing analytic predictions with numerical simulations or observations. Without predicting the detailed evolution of the satellite density profile, I show that assuming that the density within any fractional-mass radius remains constant gives an estimate of the evolution of these radii which is accurate to 15–30%.

I note that in general, the treatment of satellite dynamics outlined here will break down for very large satellites, those which are comparable to the main system in mass or in size. I will show in chapter 8 that mergers of this magnitude only occur early on in the history of most systems, however, and from dynamical friction arguments, they are expected to evolve very quickly thereafter. At later times, the largest satellites in most systems are similar to those considered in the previous three chapters, and the vast majority of satellites in any given system are much smaller.

Overall, the analytic description thus predicts the position, velocity, mass, and size of most satellites at each point in their orbit to an estimated accuracy of 20%, for typical cases, and does so using only a small number of calculations. As such, it provides a useful, interpretive point of comparison for numerical simulations. More importantly, however, it makes the problem of modelling the formation of galaxy halos through hierarchical merging computationally tractable. In Part II of this thesis, I will describe a full semi-analytic model of galaxy formation based on this preliminary work, and apply it to some of the dynamical problems mentioned in the Introduction, such as the properties of satellite galaxies, the formation of stellar halos, and the survival of galactic disks in hierarchical cosmologies.

Chapter 5

Constructing Merger Histories for Dark Matter Halos

In hierarchical CDM cosmologies, galaxies, groups, and clusters form within merging halos of dark matter. Analytic expressions for the statistics governing halo formation rates can be derived using the method of Press and Schechter. Press-Schechter statistics have been used recently to develop algorithms for constructing random, but statistically representative sets of merger histories for individual halos. These merger histories form the basis for semi-analytic models of galaxy formation. In this chapter, I describe several algorithms for generating merger trees, and the implementation of one of these algorithms, which is used in this thesis. I demonstrate the method's accuracy, and discuss some of the basic properties of the resulting merger histories.

As outlined in the introduction to this thesis, the formation, evolution and clustering of galaxies is currently understood in terms of hierarchical models of structure formation. In these models, small fluctuations present in the early universe grow through gravitational instability, eventually causing certain regions of the universe to cease their initial expansion and recollapse, forming dense, gravitationally bound objects, called 'halos'. The spectrum of initial fluctuations, that is their average amplitude as a function of spatial scale, determines the order of collapse (the relative phases of the initial fluctuations are assumed to be random). In cold dark matter (CDM) cosmologies, the amplitude of fluctuations is largest on small scales, so the smallest structures collapse first, and most large structures form later, from material that is already inhomogeneous on small scales. Thus, as halos form, they incorporate or merge with 'subhalos', smaller halos which have collapsed previously within the same volume. These subhalos may be disrupted in the merger process, or they may survive as distinct substructure within the larger halos, if they are sufficiently dense.

Until matter reaches very high densities, it consists of a uniform mix of baryons (or other dissipational material), and dissipationless dark matter. At high densities and within a certain temperature range, the interaction rates between matter particles become appreciable, and these interactions convert kinetic energy into radiation. As a result, the dissipational component within halos in a certain mass range loses energy and collapses further. This process is halted by the partial conservation of angular momentum, which spins up material as it falls in, and by the injection of more energy into the system, from star formation and related processes. Hierarchical models thus predict a characteristic structure to the universe, consisting of vast regions of dark matter, only slightly denser than the average background density, linking dense dark matter halos which are perpetually merging and growing, and which in turn contain central condensations of baryons, in the form of the stars and gas we see in galaxies.

While many details of galaxy formation are highly uncertain, hierarchical models do make certain basic statistical predictions about the formation and evolution of dark halos, which shed light on this complex process. Press and Schechter (1974), for instance, followed a simple argument to produce an estimate of how many halos of a given mass should have collapsed by a given epoch. Using statistical methods, this approach was more recently extended to predict the merger rates between halos, and other related bivariate statistics (Bower 1991; Bond *et al.* 1991; Lacey & Cole 1993). These statistics place constraints on the formation of a typical halo, specifying its average growth rate as a function of redshift, for instance. Several groups have used extended Press-Schechter (EPS) statistics to construct merger histories which are individually random, but statistically representative as an ensemble. These 'merger histories' or 'merger trees' are the basis for most semi-analytic models of galaxy formation, including the one described in this second part of this thesis. In this chapter, I will describe methods for generating merger trees, test the accuracy of my own implementation of one of these methods, and comment on the various characteristic ages of subhalos that can be defined within a merger tree. In the next chapter, I will

then develop a model for the assembly of galaxy halos, based on merger trees, and compare the basic predictions of this model to the results of high-resolution numerical simulations. First, however, I will introduce the two cosmological models considered throughout the rest of this thesis.

5.1 Cosmological Models[1]

Simple cosmological models can be described in terms of a few fundamental parameters, specifically the relative expansion rate of the universe H, often parameterised by $h \equiv H/(100\,\mathrm{km\,s^{-1}\,Mpc^{-1}})$, and the present-day density of the universe in units of the critical density required for recollapse, $\Omega \equiv \rho/\rho_\mathrm{c}$ or more precisely the density of each of its constituent forms of matter and energy. These different components include radiation, normal matter, and 'dark' or 'vacuum' energy, corresponding to the cosmological constant. Of these, only the latter two are important at late times, contributing Ω_m and Ω_Λ to the density, respectively. The contribution to the matter density is further divided into the baryonic contribution Ω_b and the dark matter contribution Ω_x, although the global dynamics of the universe depend only on the total matter density.

The cosmological parameters are normally measured at the present day, indicated by a subscript '0' – thus H_0, Ω_0 and so forth – but I will omit the subscript where the meaning is clear. In a given cosmology, the parameters at any other epoch can be determined from equations describing the overall evolution of the universe with time. It is convenient to describe the expansion of the universe in terms of a changing *scale factor*:

$$a(t) \equiv R(t)/R_0, \tag{5.1}$$

where R_0 is the size of some region at the present day, and $R(t)$ is its size at an

[1]The general reader is referred to the preface for an explanation of the terms and units used in this section.

earlier time t. Light from distant objects, seen at that earlier time, is shifted to redder wavelengths by the expansion of the universe. The scale factor is simply related to the observed redshift of objects at that time by:

$$a(t) = (1 + z(t))^{-1}. \tag{5.2}$$

In terms of the cosmological parameters and the redshift, the expansion rate of the universe at any time is

$$H(z) = H_0 \left[(1 - \Omega_{\rm tot})(1+z)^2 + \Omega_{\rm m}(1+z)^3 + \Omega_\Lambda\right]^{1/2} , \tag{5.3}$$

where $\Omega_{\rm tot}$ is the total density from all contributions. The term $(1 - \Omega_{\rm tot})$, sometimes denoted Ω_k, describes the overall curvature of the universe; in particular, $\Omega_k = 0$ corresponds to a flat universe.

In principle, the values of the cosmological parameters can be determined by observation. In practice, however, their values are currently uncertain by factors of up to 2. A full discussion of the uncertainties in the observed values is beyond the scope of this work; the reader is referred to Krauss (2001) or Balbi (2001) for a discussion of recent determinations. In this work, I will consider the simplest cosmological model, in which $\Omega_{\rm tot} = \Omega_{\rm m} = 1$ and $h = 0.5$, as well as a model with $h = 0.7$, $\Omega_{\rm m} = 0.3$ and a cosmological constant contributing $\Omega_\Lambda = 0.7$. The former model was previously thought to describe the universe, and is still referred to as 'standard' CDM (SCDM). In this model, the merger rate between halos remains high at recent times, so it offers the strictest test of the disruptive effects of merging on present-day galaxies. On the other hand, the latter model, referred to as Lambda-CDM (ΛCDM), is strongly favored by current observations (Krauss 2001). Separate observational constraints, namely the abundance of deuterium, helium and other light species, fix the value of $\Omega_{\rm b}$ in either of these models to be approximately 0.025 h^{-2}, or 10% and 5% of the critical density for SCDM and LCDM respectively.

The formation of structure in either of these models depends one or two other parameters. Structure forms from an initial spectrum of fluctuations whose shape is

Table 5.1 Summary of the Cosmological Models

Model	$\Omega_{\rm tot}$	$\Omega_{\rm m}$	Ω_Λ	$\Omega_{\rm b}$	h	Γ	σ_8
SCDM	1.0	1.0	0.0	.052	.5	.50	.70
ΛCDM	1.0	0.3	0.7	.052	.7	.21	.90

predicted by CDM models, and can be characterised by a curvature scale, Γ, and an amplitude, σ, measured on some fiducial scale. In simple CDM models with $\Omega_{\rm tot} = 1$ and $\Omega_{\rm b} \simeq 0$, Γ is in fact related to the other cosmological parameters by $\Gamma = \Omega_{\rm m} h$, so the initial spectrum is specified by the amplitude σ alone. As explained in the next section, the amplitude can be measured at the present day on any mass scale where it is sufficiently small; conventionally the fiducial scale of $8h^{-1}$Mpc is used, and the amplitude is labeled σ_8. The value of σ_8 can be determined observationally by comparison with the amplitude of fluctuations in the microwave background (Efstathiou *et al.* 1992), or with the abundance of massive clusters (White, Efstathiou, & Frenk 1993; Viana & Liddle 1996; Eke, Cole & Frenk 1996). The SCDM model used here uses the former normalisation, to allow direct comparison with numerical simulations by Moore *et al.* (1999), while the LCDM model uses the latter normalisation, which has been adopted by other numerical studies (Colberg *et al.* 1998). Table 5.1 summarises the full set of parameters describing the SCDM and LCDM model, while figure 5.1 shows the spectrum of fluctuations for either model, as a function of mass or of spatial scale. We see that the main difference between the power spectrum in the two models on galaxy scales is in its amplitude, rather than its shape.

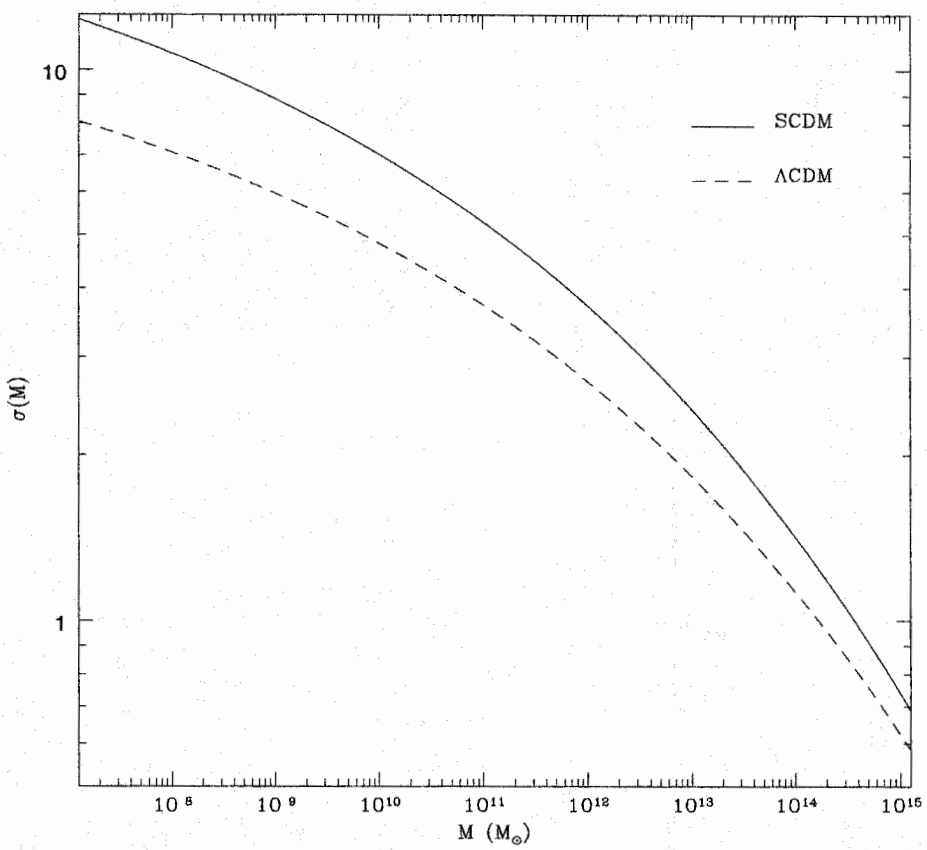

Figure 5.1 The mean amplitude of fluctuations in the early universe, $\sigma(M)$, extrapolated linearly to the present day, as a function of mass.

5.2 Gravitational Instability and Press-Schechter Statistics

The small fluctuations in the density field, whose amplitude at early times is shown in figure 5.1, will evolve through gravitational instability, producing much denser structures at late times. While the growth of structure through gravitational instability is generally complex, the evolution of an isolated fluctuation in an SCDM universe can be described in particularly simple terms, for as long as its mean density is comparable to that of the universe as a whole. Following the derivation in Padmanabhan (1993, PB hereafter, p. 273 ff.), for instance, let us consider the evolution of a 'top-hat' fluctuation, that is a spherical region of radius r, which has a uniform

density $(1+\delta)$ times that of the background, where $\delta \ll 1$ initially. Provided the perturbation is small compared to size of the universe, its evolution can be described in Newtonian terms. In particular, individual shells of material within the perturbation will satisfy energy conservation:

$$\frac{1}{2}\frac{dr}{dt}^2 - \frac{GM}{r} = E \tag{5.4}$$

where r is the radius of the shell, M is the total mass of the system and E is its total energy (PB eq. 8.8). For bound shells, $E < 0$ and this equation has the parametric solution:

$$r = a(1 - cos\theta), \quad t = B(\theta - \sin\theta); \quad A^3 = GMB^2 \tag{5.5}$$

(PB eq. 8.15).

This solution can then be used to predict the density of the fluctuation at any time, $\bar{\rho}(r,t)$ by considering the behaviour of its outermost shell. It is interesting to compare this with the background density of the universe at the same time, $\rho_{\rm bg}$. In the SCDM model,

$$\frac{\bar{\rho}(r,t)}{\rho_{\rm bg}} = (1+\delta) = \frac{9}{2}\frac{(\theta - \sin\theta)^2}{(1-\cos\theta)^3}. \tag{5.6}$$

(PB eq. 8.24–8.25). For small values of θ, corresponding to the early stages of collapse when $\delta << 1$, this gives

$$\delta = \frac{3}{20}\frac{6t}{B}^{2/3} \propto a(t), \tag{5.7}$$

where a is the scale factor introduced in the previous section. Thus, the evolution of small fluctuations will initially be *linear* in the scale factor. Linear evolution is particularly convenient mathematically, as it preserves the shape of the initial spectrum of density fluctuations (figure 5.1), changing only the amplitude as the universe expands.[2]

It is also instructive to compare the full, non-linear evolution of a fluctuation with the linear approximation. The solution described by equation (5.5) will reach

[2]In fact, the spectrum shown in figure 5.1 is actually the spectrum as it would have appeared at early times, extrapolated to the present day by linear evolution.

a maximum radius, called the *turn-around radius*, when $\theta = \pi$. At this point the density within this radius relative to the background density is simply:

$$\frac{\bar{\rho}(r,t)}{\rho_{\rm bg}} = \frac{9}{2}\frac{(\pi)^2}{(2)^3} \simeq 5.6\ . \tag{5.8}$$

The linear calculation, on the other hand, gives $\bar{\rho}(r,t)/\rho_{\rm bg} = 2.063$, or $\delta = 1.063$ at this time. After turn-around, the solution will collapse back to $r = 0$ at $\theta = 2\pi$. Long before it reaches this point, however, any real perturbation will cease to follow the idealised mathematical solution, as shells of material cross and small potential fluctuations within the perturbation are amplified. If the system reaches virial equilibrium in this process (known as violent relaxation), then conservation of energy predicts that its final radius will be half the turn-around radius, so the final density will be 8 times the density at turn-around. Assuming virialisation occurs at roughly $\theta = 2\pi$, then in terms of the background density at this time, $\rho_{\rm bg,vir}$, the density of the virialised object will be

$$\bar{\rho}(r,t) = \frac{9\pi^2}{2}\rho_{\rm bg,ta} = \frac{9\pi^2}{2}((2)^{2/3})^3\rho_{\rm bg,vir} = 18\pi^2\rho_{\rm bg,vir} \simeq 178\rho_{\rm bg,vir}\ , \tag{5.9}$$

whereas the linear result at the same time is: $\bar{\rho}(r,t)/\rho_{\rm bg} = 2.686$, or $\delta \equiv \delta_{\rm c} = 1.686$. While the preceding derivation is only valid in SCDM, it is possible to carry out a similar calculation in other cosmologies such as LCDM (e.g. Viana & Liddle 1996). In this case, both the critical overdensity $\delta_{\rm c}$ and the density ratio in equation (5.9), defined more generally as $\Delta_{\rm c} \equiv \rho_{\rm vir}/\rho_{\rm c}$, will take on slightly different values; $\delta_{\rm c} \simeq \delta_{\rm c,SCDM}\Omega_{\rm m}^{0.0055}$ in LCDM models with $\Omega_{\rm tot} = 1$ (Peebles 1980), while figure 5.2 compares $\Delta_{\rm c}(z)$ for the SCDM and LCDM models.

5.3 Press-Schechter Theory

The results of the previous section suggest an analytic approach to predicting the statistical properties of collapsed structures. As long as the density of a region of

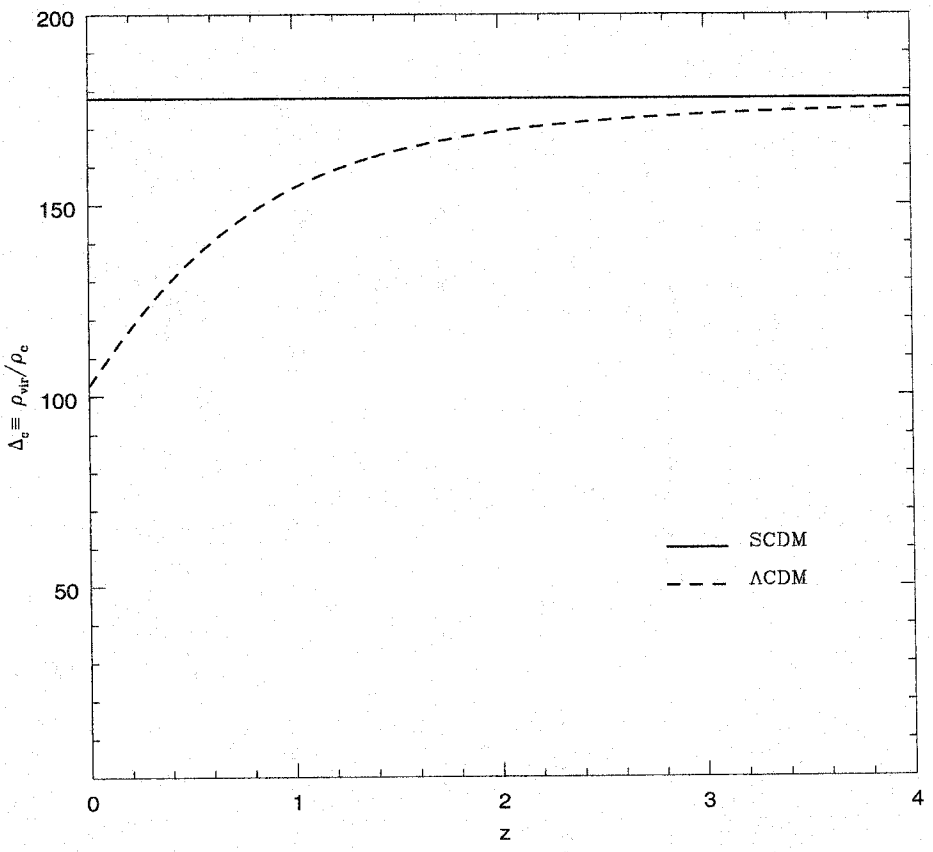

Figure 5.2 Δ_c, the ratio of the virial density of a collapsed top-hat perturbation, to the critical density at that redshift.

space is close to the mean value, the density contrast will grow at a universal linear rate $a(t)$. Thus, if the distribution of density fluctuations is known at some early time (say at the epoch of reionisation, when the shape of the fluctuation spectrum can be determined directly from the CMB), the distribution at some later time will simply be a scaled version of the original, wherever the density contrast δ is small. Only in regions where δ is large will non-linear evolution have taken over; for SCDM the calculations above suggest that regions which would have reached $\delta = 1.063$ by linear evolution alone will be turning around, while regions which would have reached $\delta_c = 1.686$ will actually have recollapsed. In particular, if the spherically averaged density contrast δ about some point in the initial density field is known as a function

of radius, then setting $a(t)\delta(r) = \delta(r)/(1+z) = \delta_c(z) = 1.686$ will provide an estimate of the region around this point which has collapsed and virialised by redshift z, as well as the mass of the corresponding collapsed object:

$$M(z) = \frac{4\pi}{3} r(z)^3 \delta_c(z) \rho_c(z). \tag{5.10}$$

Conversely, if a volume V around a point, which contains a mass M in the initial density distribution, is overdense by more than $\delta_c(z)$ when evolved linearly up to the epoch z, then its contents should reside in a collapsed structure of at least that mass, by that redshift.

Based on this reasoning, Press and Schechter (1974) suggested a method for predicting the number of collapsed objects in the universe as a function of mass and redshift. Writing the fraction of the initial density distribution that has collapsed by redshift z, when averaged over a volume containing an average mass M, as:

$$f(M, z) = \int_{\delta_c}^{\infty} p(\delta) d\delta, \tag{5.11}$$

where $p(\delta)$ is the probability of a point in the initial density field having an overdensity δ, they estimated that the number of collapsed objects of mass M at redshift z was simply:

$$n(M, z) = \frac{1}{M} \frac{d}{dM} f(M, z). \tag{5.12}$$

This estimate is not quite correct, since it assumes that f increases monotonically for decreasing M, or equivalently that if material is within an object that exceeds the critical density threshold on some mass scale M, the volume around it will also exceed this threshold when smoothed on any smaller mass scale. In fact, collapsed regions will contain a mixture of overdense and underdense material, introducing a correcting factor of roughly $1/2$ into equation (5.12) [3] . Assuming the initial spectrum of density fluctuations is Gaussian, with the amplitude $\sigma(M, z) = D(z)\sigma_0(M)$ (where

[3] A more rigorous derivation of the PS result, such as that of Bower (1991) or Bond *et al.* (1991), confirms that the factor is exactly $1/2$.

$\sigma_0(M)$ is the amplitude extrapolated linearly to the present day, and $D(z)$, the linear growth factor, accounts for the redshift dependence), then the predicted number of objects is:

$$n(M,z) = \frac{1}{M}\,\mathrm{erfc}\left[\frac{D(z)\sigma_0(M)}{\delta_c}\right]\frac{D(z)}{\delta_c}\frac{d}{dM}\sigma_0(M), \tag{5.13}$$

where erfc is the complimentary error function:

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty \exp^{-s^2}\,ds \tag{5.14}$$

Figure 5.3 shows $n(M)$ at $z = 0$ for the two cosmologies considered here, over the mass range corresponding to galaxies, groups and clusters.

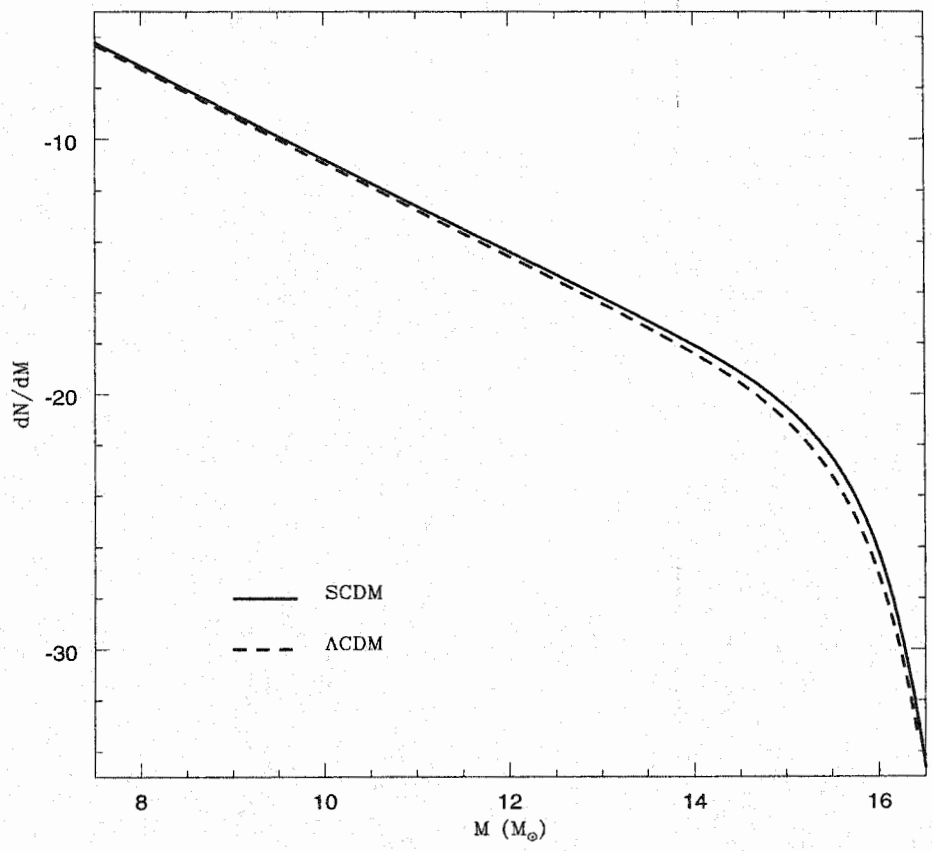

Figure 5.3 The Press-Schechter prediction for the number of collapsed halos at z = 0.

The Press-Schechter prediction has been tested extensively against numerical simulations (Efstathiou *et al.* 1988; Carlberg & Couchman 1989; Lacey & Cole 1994;

Gelb & Bertschinger 1994; Ma 1996; Tozzi & Governato 1998; Gross *et al.* 1998; Tormen 1998). Overall it shows reasonably good agreement, given the simplifying assumptions made in its derivation, although the relative number of low-mass and high-mass halos are somewhat over-predicted and under-predicted respectively (Jain & Bertschinger 1994; Tormen 1998; Gross *et al.* 1998; Governato *et al.* 1999). This may be due to a systematic variation in δ_c with mass; a model which accounts for the non-spherical geometry of collapsing halos provides an improved fit to simulation results (e.g. Sheth & Tormen 1999; Sheth, Mo & Tormen 2001; Jenkins *et al.* 2001). Unfortunately, this model cannot be incorporated into the merger tree methods described below, which require δ_c to be independent of mass (van den Bosch 2001). Furthermore, in the merger tree code described below, the Press-Schechter method is only used to determine relative rates of halo formation, over a small, sub-galactic mass range, which should be less sensitive to this effect. Thus, I will use equation (5.13) as the basis for generating merger trees.

With this caveat, the Press-Schechter prediction provides a simple and relatively accurate way of estimating the outcome of the highly non-linear process of structure formation, and as a result it has been extended and generalised by several authors. In particular, both Bower (1991) and Bond *et al.* (1991) laid the groundwork for determining the statistical properties of individual regions, as well as putting the theory on much sounder theoretical footing. This allowed Lacey & Cole (1993, LC hereafter) to formulate merger probabilities and various other bivariate statistics. Thus, for example, while the Press-Schechter formula gives the probability that a region has collapsed on a mass scale M by a redshift z, thereby providing an estimate of the number of such regions, Lacey and Cole (1993) calculated the probability that a region would collapse on a mass scale M_1 by a redshift z_1, given that it had *already* collapsed on a mass scale M_2 by the earlier redshift z_2, thereby providing estimates of the instantaneous merger rate between halos, and several average properties of their individual formation histories. These results, known as extended Press-Schechter (or

EPS) statistics, form the basis a whole class of semi-analytic galaxy formation models described below.

5.4 Constructing Merger Histories

Using the results of Lacey & Cole (1993), it is possible to calculate the probability that material that is in collapsed structure, or halo, of mass M_2 at an epoch z_2 was in a halo with a smaller mass between M_1 and $M_1 + dM_1$ at a later epoch z_1. This probability is simply:

$$f_{S_1}(S_1, \omega_1 | S_2, \omega_2)\, dS_1 = \frac{1}{(2\pi)^1/2} \frac{\Delta\omega}{\Delta S^{3/2}} \exp\left[-\frac{\Delta\omega^2}{2\Delta S}\right] dS_2, \qquad (5.15)$$

in terms of the dimensionless variables $S = \sigma_0^2(M)$ and $\omega = \delta_c(z)$ (LC eq. 2.15). The conditional probability $f_{S_2}(S_2, \omega_2 | S_1, \omega_1)$ can also be calculated, and gives the probability of a halo of mass M_1 merging to form an object of mass M_2 over some redshift range (LC eq. 16). The original halo may steadily accrete structureless material to make up the mass difference $dM = M_2 - M_1$, or it may merge with other halos, in which case the mass of the halo (that is the mass of the connected overdense region) will increase instantaneously at each merger. Thus, taking the limit of equation (5.15) as z_2 approaches z_1, for some fixed mass difference dM, gives the probability that the original halo merged with another single halo of mass dM. Using a set of these probabilities, calculated for different values of M_1, dM, and z, one can then generate random mass accretion histories for halos which start out with a given initial mass. In some cases, the result is even analytic. Thus, for instance, by counting the number of progenitors of halos with some final mass M_2 at z_2, and selecting those that have a given fraction r of the final mass at some earlier epoch z_1, one can estimate the probability that a halo 'formed' by this epoch, that is the probability that its main

progenitor had attained the mass rM_2 by z_1:

$$P_r(z_1) = \int_{S_2}^{S_r} \frac{M_2}{M_1} f_{S_1}(S_1, \omega_1 | S_2, \omega_2)\, dS_1, \tag{5.16}$$

where $S_r = S(rM)$ (LC eq. 2.26). This expression is only valid for $r > 0.5$, since otherwise there may be more than one progenitor with more than this mass, and even then it is not fully rigorous, as explained in LC, due to the limitations in EPS statistics.

While this approach of following mergers forward in time yields several useful results, it is generally more interesting to be able to reconstruct a halo's complete mass-accretion history, given a specified final state. In its fullest form, this 'merger history' describes how the material in a halo at the final epoch is distributed in smaller progenitor halos at every earlier epoch. It is sometimes also called a merger 'tree', by analogy with a tree structure whose trunk corresponds to the final halo, and whose branchings represent the division of this material into distinct progenitors at earlier times (see figure 5.4).

The problem of generating merger trees is much harder to resolve algorithmically. The fundamental difficulty arises from the fact that while binary merger probabilities of type (5.15) are well defined and can be evaluated from analytic expressions, they cannot be constrained to give a specific final mass for an object undergoing many successive mergers (Kauffmann & White 1993; Somerville & Kolatt 1999). Conversely, starting with a halo of a known mass at some final time, there is no closed-form analytic expression to predict the joint probability that its contents will be split between two or more progenitors of specified masses at an earlier time. Thus, merger tree algorithms must either violate mass conservation, or use inaccurate merger probabilities. Early attempts to construct merger histories that resulted in objects of a specific mass at the present day (Cole 1991; Kauffmann & White 1993; Cole *et al.* 1994), for instance, violated mass conservation by using discrete mass steps in the merger history. This limited their mass resolution, as well as their overall efficiency

and accuracy (see Somerville & Kolatt (1999, SK hereafter) for a discussion of these problems).

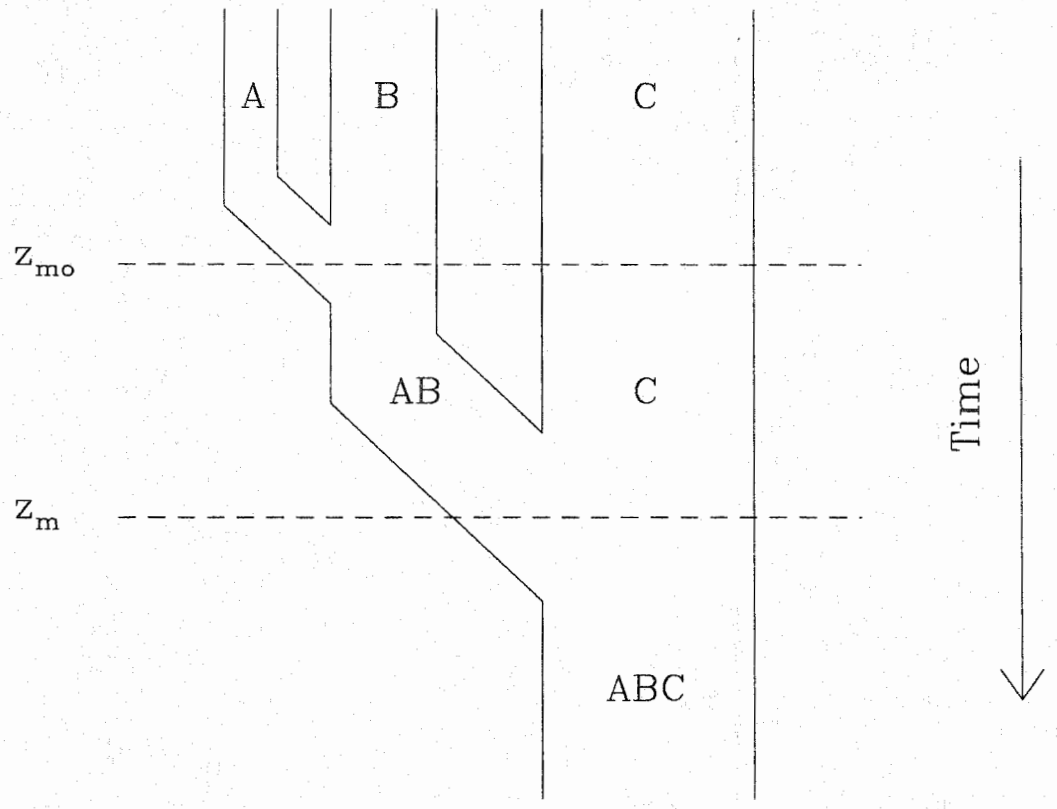

Figure 5.4 A sample merger tree, in which halo '**A**' merges with halo '**B**', and the combined system **AB** then merges with **C**. The merger epoch z_m and the original merger epoch z_{mo} for halo **A** are indicated on the plot.

An alternate approach, which has since become standard in semi-analytic models, was first proposed by SK. In this method, the merging history of a halo with a specified mass at the final epoch is traced backwards in time. To calculate accurately the probability of the halo dissolving into one or more progenitors at any given timestep, the stepsize is kept so short that the probability of a multiple merger occurring is small. Thus, most of the mergers in the tree are binary mergers, whose probability distribution is given by the closed-form analytic expression (5.15). Once the progenitors of a given halo are determined, the merger history of each progenitor

in the next timestep is then calculated in the same fashion. While the calculation does have a limiting mass resolution, in the sense that branches with less than some limiting mass are considered smooth accretion, and traced no further, halo masses can vary continuously above this limit. Furthermore, having reduced most of the merger calculations to a series of simple, binary mergers, the method is very fast computationally. Adjusting the scaling of the timestep to optimise the overall performance and accuracy, it is easy to generate merger trees with dynamic ranges of 10^4–10^5 in mass, with thousands of individual branches and timesteps, in a few minutes of CPU time. In the next section, I describe an implementation of this approach, designed to generate merger histories for galaxy, group and cluster halos.

5.5 Implementation

To establish the mass accretion histories of a representative sample of galaxy-sized halos, to be used in determining the evolution of halo substructure, I developed an implementation of the Somerville & Kolatt algorithm, as described in SK. This algorithm starts with a halo of a given mass at some final redshift, and traces its accretion history back until none of its progenitors exceed the mass resolution, or until some limiting redshift is reached. To model the formation of a large disk galaxy comparable to the Milky Way, I chose a final mass of $1.6 \times 10^{12}\,\mathrm{M}_\odot$ at the present day, consistent with recent estimates of the total mass of the Milky Way's halo (Zaritsky 1998). The substructure observed within the halo of the Milky Way, namely its retinue of dwarf satellites, extends down to estimated masses 10^4–10^5 times smaller than the total mass of the Milky Way's halo, so the mass resolution in the merger tree was chosen to be $5 \times 10^7\,\mathrm{M}_\odot$, or roughly 3×10^{-4} times the mass of the whole tree. The limiting redshift required to achieve convergence in the properties of the trees depends on the mass resolution; for the resolution used here, a limiting redshift

of around 20–25 was found to be sufficient, so a redshift of 30 was chosen to ensure convergence.

An important element of the method of SK is the choice of timestep; it has to be short enough to guarantee that halos have a single progenitor in most timesteps, yet long enough that the calculation is reasonably fast. SK pick a fiducial timestep based on the probability that a halo of mass M_0 will have more than one progenitor over the resolution limit M_1, in order to satisfy the first constraint:

$$\Delta\omega_0 = \delta_c(t_2) - \delta_c(t_1) \leq \sqrt{\left.\frac{dS}{dM}\right|_{M_0} M_1} \tag{5.17}$$

In order to improve the efficiency of the calculation, however, they suggest the scaling:

$$\Delta\omega = (a \log(M_0/M_1) + b)\Delta\omega_0, \tag{5.18}$$

with a = 0.3 and b = 0.8. In the merger trees considered here, the mass resolution is higher than that considered by SK, so more conservative values of $a = 0.15$ and $b = 0.02$ were used, based on the tests described below.

Overall, the merger trees used here recorded approximately 3000 mergers with the main trunk, over about as many timesteps. The length of each timestep was approximately 0.1 Gyr (corresponding to $\Delta z \simeq 0.01$) at the present day, decreasing to 0.01 Gyr ($\Delta z \simeq 1$) at z = 30. The number of progenitors chosen per step was $\simeq 1.2$, in keeping with the requirement that it be close to 1. Each merger tree took on the order of 15 minutes to generate on a desktop system, but much of this overhead came from the dynamical calculations in the branches, described in the next chapter. The basic calculation took only a third as long. Recording all the mergers with the main trunk of the tree produced approximately 100 kb of data per tree.

To test the timestep scaling, and the overall accuracy of the code, I performed several tests. The average mass spectrum of progenitors of a halo at a given epoch, for instance, can be calculated analytically from EPS theory:

$$n(M)dM_1 = \frac{M_0}{M_1} f_{s_1}(S_1, \omega_1 | S_2, \omega_2) \frac{dS}{dM} dM_1 \tag{5.19}$$

In figure 5.5, I show the spectrum of progenitors averaged over 120 SCDM trees, at four different epochs (histograms), compared with this prediction. Figure 5.6 shows similar results for LCDM. The abscissa is labeled in units of the mass resolution M_1. We see that overall, the agreement is excellent for masses more than 10 times the mass resolution. The spectrum close to the resolution limit is slightly lower than the EPS estimate at the lowest redshift, and slightly higher at the highest redshift. SK found similar results, however, suggesting that this is an intrinsic limitation of the algorithm. Nonetheless, the agreement is generally excellent, particularly when compared with other algorithms (SK).

5.6 Characteristic Ages within the Merger Tree

The merger tree records only the mass of progenitors of a final halo, at each of a series of earlier epochs. All other properties of these progenitors must be inferred from their own merger histories within the tree. Numerical results, notably NFW (1996, 1997), but also more recently Bullock *et al.* (2001b) and Eke, Navarro & Steinmetz (2001) suggest that the properties of dark matter halos depend principally on their mass and the epoch at which they formed. In particular, older halos are expected to be more centrally concentrated, because their cores are denser, having formed at a time when the mean density of the universe was greater. Thus it is interesting to establish a rigorous definition for the formation epoch of a given halo in the merger tree. Several different definitions have been proposed. The first was introduced in the previous section; for a halo of mass M at a given epoch, there is a unique epoch when its main progenitor first had a mass of more than fM, where f is some fraction greater than a half, which can be taken as the formation epoch of the system. (If f is less than a half, it may no longer be clear which progenitor is the 'main' progenitor, and the analytic calculation is no longer valid, as explained in section 5.4.)

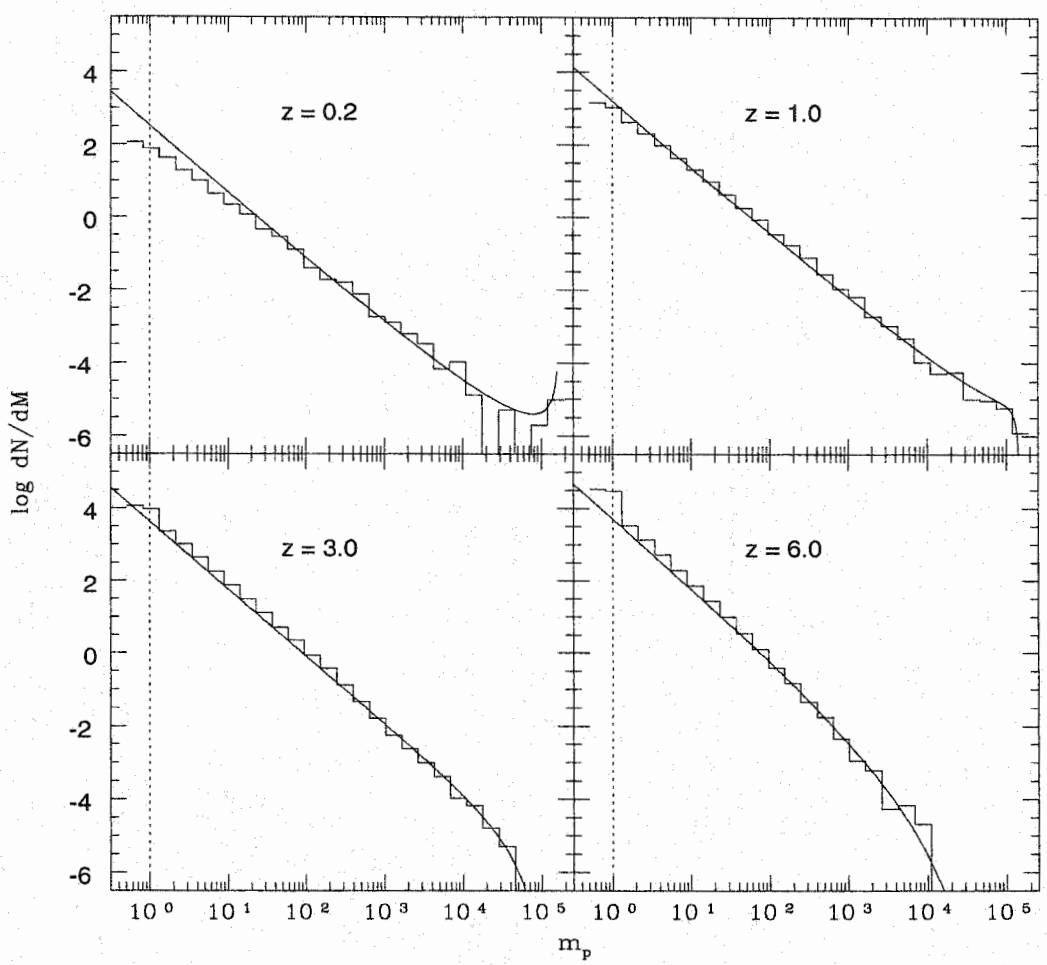

Figure 5.5 Distributions of progenitors at four different epochs in the SCDM merger trees. The smooth curves are the EPS prediction, while the histograms are the average results for 120 trees. Mass is measured in units of the resolution limit.

Alternately, one can use the related definition, that the formation epoch is the epoch when the average particle, in the halo at the present day, was in a progenitor of mass fM. This epoch can be determined by solving the equation:

$$\mathrm{erfc}\left[\frac{\Delta\omega}{\sqrt{2\Delta S}}\right] = \frac{1}{2}, \tag{5.20}$$

where $\Delta S = \sigma_0^2(fM) - \sigma_0^2(M)$. It was this formation epoch that was found to correlate with concentration by NFW (1996, 1997). It is also possible to define the formation epoch in terms of the merger history of a halo, as the epoch when the mass of a halo

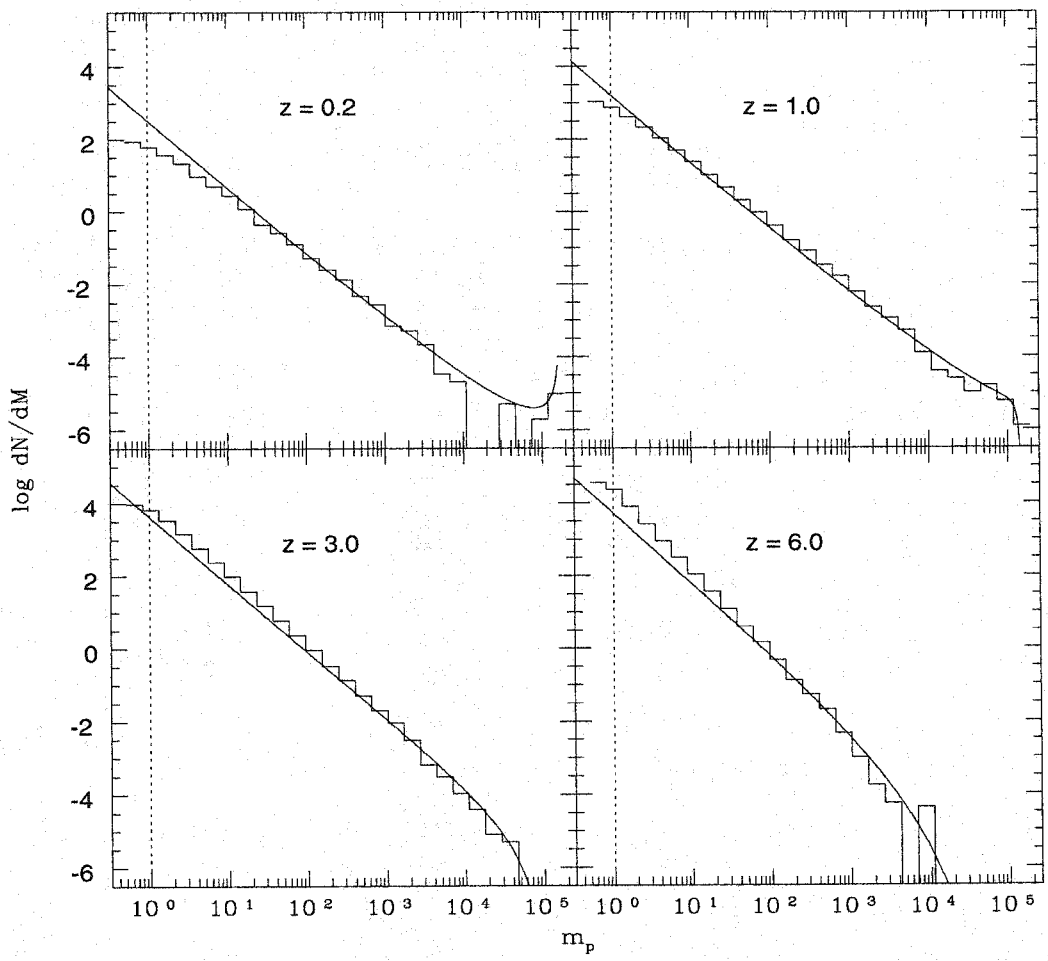

Figure 5.6 As 5.5, for the LCDM trees

last increased by some fraction. An analogous definition will be used to assess disk disruption rates in chapter 8.

Two characteristic epochs used in the next chapter depend on a halo's fate within the merger tree, rather than on its previous mass accretion history. Figure 5.4 shows a schematic section of a merger tree, in which halo **A** merges with halo **B**, before the combined system **AB** merges with **C** to form **ABC**. While EPS theory considers **AB** a single, collapsed object, it may contain substructure, just as the halo of the Milky Way does. In particular, the smaller halo **A** may not have been fully assimilated before the combined system **AB** merges with halo **C**, so that this later merger adds a

group of two distinct subhalos, **A** and **B**, to halo **ABC**, rather than a single system, **AB**. To characterise the (sub)halo **A** in this situation, I will refer to the epoch when **A** first merges with **B** as the *original merger epoch* for halo **A**, denoted $z_{\rm mo}$, and the epoch when **A** and **B** both merge with the main halo **C** as simply the *merger epoch* for halo **A**, denoted $z_{\rm m}$. These two definitions will be used in the next chapter. The bottom panel of figure 5.7 shows the distribution of $z_{\rm m}$ for halos in a typical merger tree (the discrete appearance of this distribution will be explained in the next chapter), while the top panel shows the distribution of $z_{\rm mo}$. Clearly the original merger epoch is much earlier for most of the halos in the tree – typically $z_{\rm m} \simeq 1$–3 while $z_{\rm mo} \simeq 5$–10.

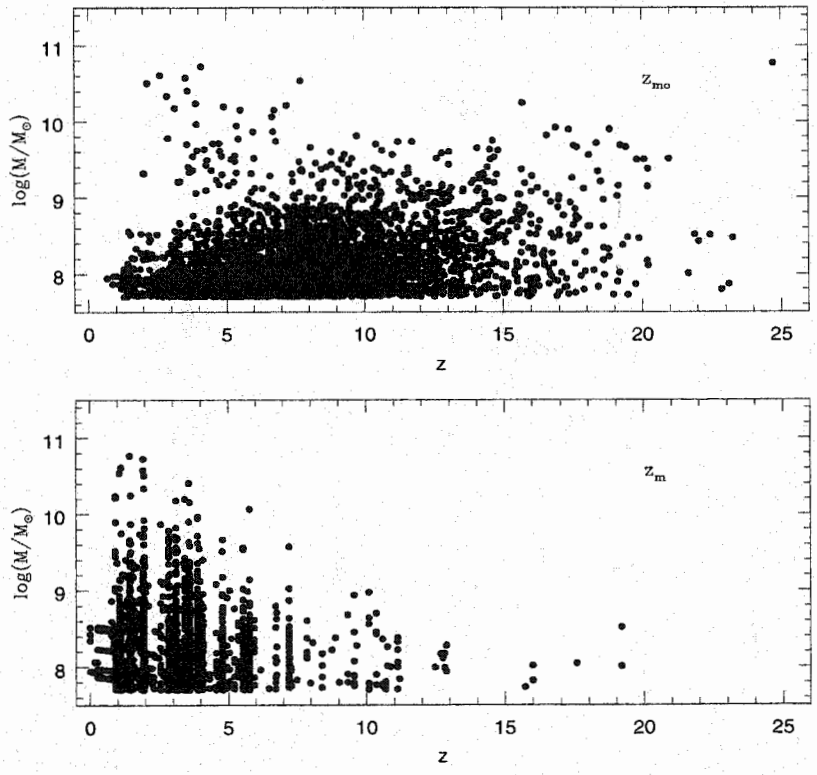

Figure 5.7 The characteristic epochs $z_{\rm m}$ and $z_{\rm mo}$ defined in the text, for the subbranches in a single merger tree.

5.7 Summary

In this chapter, I have outlined the statistical approach to structure formation first suggested by Press and Schechter. This approach has been extended considerably by many authors over the past fifteen years, and is now the basis of Monte Carlo techniques for generating representative mass-accretion histories, or merger trees, for sets of dark halos. I have also presented an implementation of one of these methods, based on the algorithm of Somerville and Kolatt. This merger-tree code can produce mass accretion histories resolving thousands of subhalos, in only a few minutes of CPU time on a desktop machine. The resulting merger histories show excellent agreement with the predictions of EPS theory. Finally, I have discussed how to assign characteristic ages to individual halos within a merger tree.

The PS description of structure formation, while mathematically straightforward, is rather conceptually abstract. In particular, the merger-tree code developed in this chapter contains no spatial information about the halos at any given epoch. While this information was unnecessary in previous semi-analytic models, which averaged over the spatial structure within individual halos, it is clearly required for any model which hopes to resolve the contents and structure of individual halos. The move from merger trees to a full model of galaxy formation is problematic, and will be discussed separately in the next chapter.

Chapter 6

From Merger Trees to Galaxy Formation

In this chapter, I discuss how to relate the Press-Schechter description of halo growth to the process of galaxy formation. Moving from one to the other requires a careful treatment of surviving substructure within each side-branch joining the main trunk of a merger tree. I explain how this substructure can be accounted for by using a simplified version of the dynamical models developed in part I. I also introduce a set of basic rules to specify the evolution of the central baryonic components within the main halo. Together, these components constitute a semi-analytic model of galaxy formation, or more specifically the formation and dynamical evolution of galaxy halos. I compare some basic predictions of this model with numerical simulations of groups and clusters. Overall, the model matches the numerical results very accurately, producing similar distributions of subhalo mass, peak velocity and distance from the centre of the system. The predictions for the stellar contents of a galaxy halo and for disk survival will be discussed separately in two subsequent chapters.

The previous chapter started out by considering the dynamics of dissipationless material during the initial stages of gravitational collapse, when the growth of structure is linear, or only weakly non-linear. In this regime, gravitational dynamics are sufficiently simple that it is possible to derive analytic estimates of the numbers, total masses and overall growth rates of halos as they enter the non-linear regime. Even in a purely dissipationless model, the internal dynamics of a realistic system, once shells of material cross after their initial collapse, is much more complicated, and cannot be determined analytically. Unfortunately, it is precisely in this high-density regime that galaxy formation occurs, permitting stringent observational tests of theories of structure formation. In this chapter, I will try to make the connection between the EPS merger histories discussed in chapter 5, and the process by which galaxies, groups

and clusters form, developing a semi-analytic (SA) model of galaxy formation based on the evolution of individual halos.

A basic outline of this SA model is as follows. The model considers the formation and evolution of a central galaxy within the dark matter halo corresponding to the main trunk of the merger tree. The structure of this halo-galaxy system at any epoch is taken to be smooth and simple, the dark component having a cosmological density profile, and the baryonic components constituting the central galaxy, a disk and a spheroid, having density profiles similar to the stellar and gaseous components of the Milky Way. The total mass of the dark component is determined by the merger tree, while the other components grow following various simple laws. In particular, the model does not derive a growth rate for the disk, but assumes that it grows roughly in proportion to the dark halo, while the spheroid is built up from stellar material stripped from dwarf galaxies, as explained below and in chapter 7.

The interpretation of branches in the merger tree is more complex. Each represents a merger with another halo, which adds its mass to the total mass of the main system. The core of this halo may be sufficiently dense to withstand tidal disruption, however, producing a distinct subhalo within the main system. The subsequent evolution of this subhalo can be described using the analytic methods developed in part I. This approach suffices if each branch is small, and contains only a single dense core. If at some point the main trunk of the tree experiences a major merger with another halo of comparable mass, this other halo may contain its own substructure. A full model of halo growth should in principle examine the history of each side branch in as much detail as the main trunk. This would be very expensive computationally, however, and would require a more detailed treatment of mergers.[1] The detailed evolution of the main galaxy is also generally of more interest, as it dominates the system and can be related to local observations of our own Galaxy. Thus in practice, the model

[1] As mentioned in chapter 4, the analytic model will also break down for mergers of similar-mass components.

determines the substructure within each branch, at the moment it merges with the main trunk, using a simpler method based on its previous merger history. I will start by explaining this process of 'pruning' in the next section, and then I will return to the evolution of the main trunk in section 2.

6.1 Pruning Merger Trees

The problem of accounting for branchings in a merger tree is best understood in terms of the physical example of cluster substructure. Many real clusters show substructure in their spatial or velocity distributions (e.g. Virgo (Petrosian *et al.* 1998), Coma (Gurzadyan & Mazure 2001, and references therein), Fornax (Drinkwater, Gregg & Colless 2001), and more distant clusters (Plionis 2001)). This substructure indicates that they include several distinct groups which have not yet dissolved completely into the main system. In the hierarchical, CDM picture, clusters have assembled recently through the merger of several large halos, each corresponding to a distinct group of galaxies. Each of these groups has many members, which collectively show correlations in their positions and velocities. If one were to construct a merger history for the whole cluster to predict its total galaxy population, it would be necessary to count mergers with large halos as contributing many small dense subhalos containing individual galaxies, rather than a single massive object. The same situation occurs on smaller scales, due to the scale-free properties of dark matter. If the progenitor of a Milky Way-like system merges with a large halo, this halo should contain its own substructure, some of which will survive in the halo of the main system, just as galaxies from groups survive within clusters. To determine the substructure of the main halo, it is thus necessary to account for the substructure of all its branches.

Substructure will not survive indefinitely within a halo; we saw in part I that

dynamical friction will cause its orbit to decay into the centre of the halo, where it is disrupted. While it is not feasible to include the full analytic calculation for each subhalo in each side-branch of the tree, it is possible to estimate how long it will take a subhalo to fall in and be disrupted. For an orbit of circularity ϵ, this timescale is given by:

$$\tau_{\rm infall} = 1.2 \frac{J_{\rm c} r_{\rm c}}{G M_{\rm sat} \ln(M_{\rm halo}/M_{\rm sat})} \epsilon^{0.4}, \qquad (6.1)$$

where $J_{\rm c}$ and $r_{\rm c}$ are the angular momentum and radius of a circular orbit of the same energy, and $\epsilon \equiv J/J_{\rm c}$ is the circularity of the orbit. (Colpi *et al.* 1999). These authors further scale this time by a factor 'e' to account for the reduction in dynamical friction due to mass loss. When pruning the merger trees, I will set this parameter to 1, which gives good agreement with numerical results, as shown below.

This estimate can be used to determine whether a given subhalo survives within its sub-branch, up until the time when this branch merges with a larger branch, or with the main trunk of the tree. Considering the situation shown schematically in the top left-hand corner of figure 6.1, if the infall time estimated using equation (6.1) (where ϵ, the circularity, is picked at random from a cosmologically representative distribution) is longer than the time between $z_{\rm m}$ and $z_{\rm mo}$, then the whole branch will have merged with the main trunk before the subhalo disrupts within the subsystem, and therefore the subhalo should be passed on to the main trunk as a distinct object, and the mass of its parent branch reduced accordingly. If the infall time is shorter than the time between $z_{\rm mo}$ and $z_{\rm m}$, then the subhalo will have been absorbed by its parent, and thus the branch can be treated as a single, structureless object. These two possibilities are shown schematically in figure 6.1. For large branches, this process of 'pruning' (or more precisely 'grafting') produces bursts of mergers with the main trunk at a single epoch, corresponding to the epoch when the branch itself merges. This explains the discrete distribution of merger epochs seen in figure 5.7 in the previous chapter. It is worth pointing out that this is not an artefact of some approximation

in the merger tree. If one could determine the distribution of epochs when individual galaxies had joined a large cluster, they would show similar discreteness, as whole groups of galaxies merged with the cluster at once.

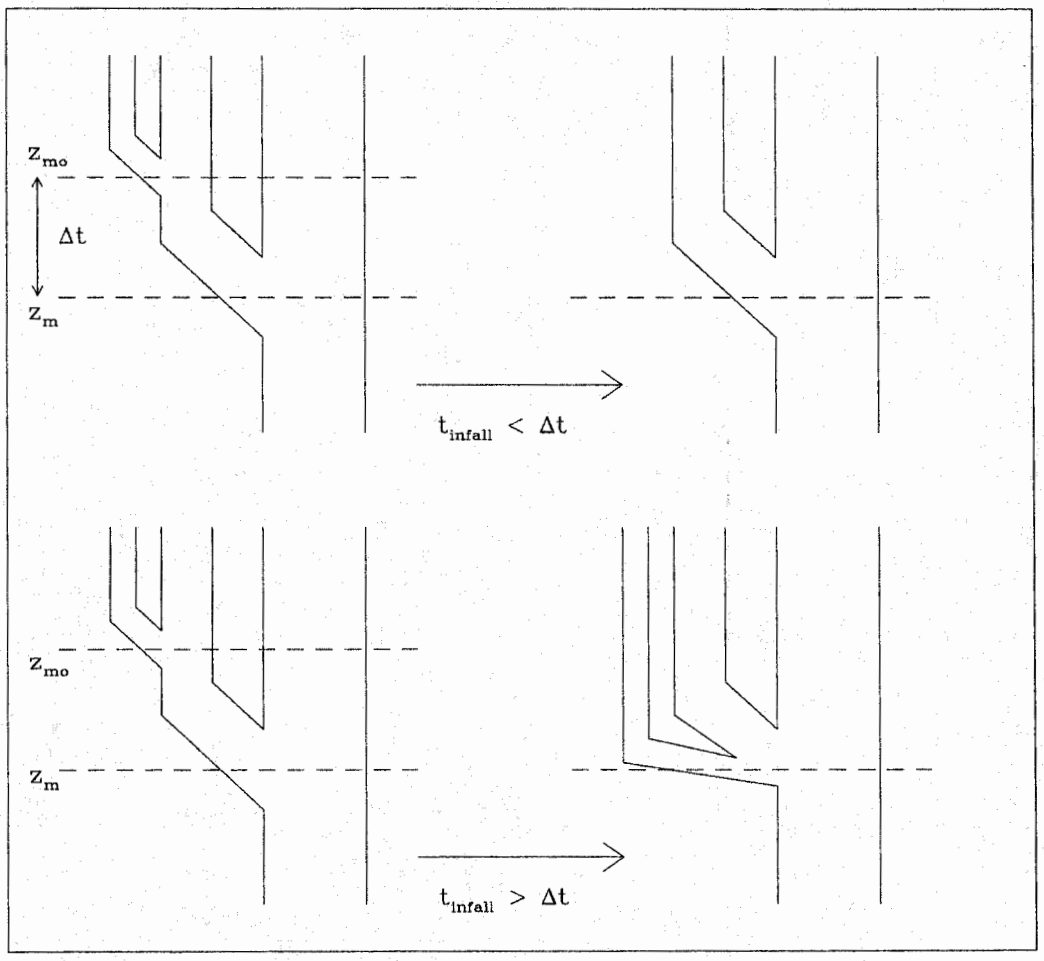

Figure 6.1 A schematic diagram indicating how substructure in a branch of the merger tree is either incorporated into the branch or counted as a separate object, at the epoch of the final merger with the main trunk, depending on its infall time within the branch.

Given this method for dealing with substructure within branches, a merger tree can be reduced to a series of individual mergers with the main trunk, occurring in sporadic bursts. The cumulative mass function of these merging objects, averaged

over 120 SCDM merger trees, is shown in figure 6.2. As the halo grows, it 'sweeps up' more and more material from surrounding space, including smaller halos. On small scales, the cumulative mass function therefore resembles the cumulative mass function of the universe (cf. figure 5.3), with a slope of -1 on galactic scales, while close to the present-day mass of the halo, mass conservation truncates this distribution. The resulting mass distribution is well-described by the integral of a Schechter function:

$$n(>M) = \int_{M}^{M_{\rm vir}} \left(\frac{M}{M^*}\right)^{\alpha} \exp(M/M^*) \frac{dM}{M^*}, \tag{6.2}$$

with $\alpha = -1.96$ and $M^* = 0.16\, M_{\rm vir}$, where $M_{\rm vir}$ is the mass of the whole tree. The rate at which the halo merges with other halos increases as it grows, so that most of its substructure at any time has been acquired fairly recently. The thick line in the top panel of figure 6.2 shows the spectrum for the SCDM trees, counting all mergers back to the initial redshift of the tree, $z = 30$, while the thin lines show the spectrum for mergers which have occurred prior to $z = 0.5$, $z = 1$, $z = 3$, and $z = 6$. The bottom panel shows similar results for LCDM. We see that in both cases, most subhalos merge with the main tree between redshifts of 1 and 3.

In a pure CDM cosmology, substructure should continue down to very small scales. The distribution in figure 6.3 is truncated at $5 \times 10^7\,\mathrm{M}_\odot$, as explained in the previous chapter, both for computational reasons, and since this corresponds to the estimated mass of the smallest dwarf galaxies in the Local Group. Given that the pruning operations described above may depend on the shape of the mass spectrum over some range in mass, it is important to test the convergence of the mass distribution as a function of limiting resolution. Figure 6.4 shows the mass function calculated for sets of trees with limiting resolutions of $2.5 \times 10^7\,\mathrm{M}_\odot$, $5 \times 10^7\,\mathrm{M}_\odot$, $1 \times 10^8\,\mathrm{M}_\odot$, and $2 \times 10^8\,\mathrm{M}_\odot$. We see that the normalisation of the mass distribution converges at masses greater than the resolution limit, for a resolution limit of $5 \times 10^7\,\mathrm{M}_\odot$ or less, and that the slope below $10^{-2} M_{\rm vir}$ appears to be constant and more or less independent of mass resolution.

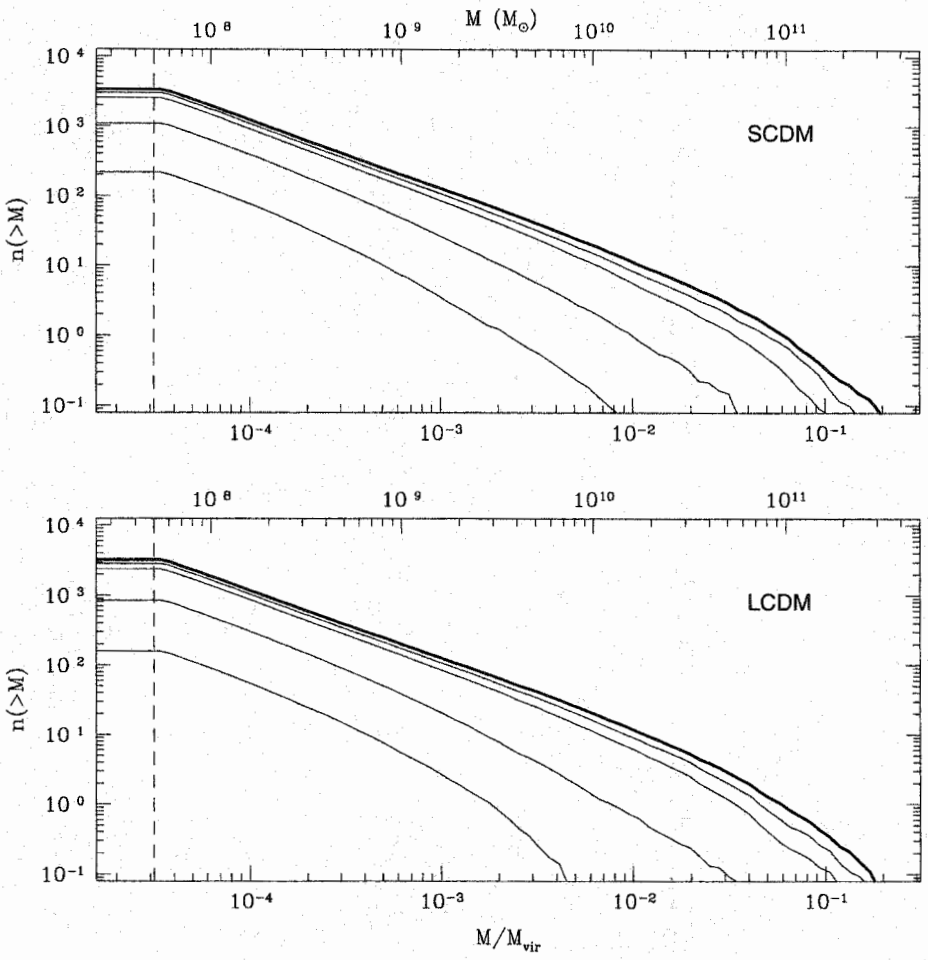

Figure 6.2 The cumulative distribution of subhalos merging with the main trunk, shown for several different redshift ranges. The ranges are $z = 6$–30, 3–30, 1–30, 0.5–30 and 0–30 from top to bottom. Most subhalos merge between $z = 3$ and $z = 1$.

Finally, to clarify the effect of including substructure from side-branches as discussed above, in figure 6.5 I show the cumulative mass spectrum of halos merging with the main trunk, calculated assuming each branch corresponds to a single, structureless object (dotted curves). The dashed curves in figure 6.5 show the effect of counting all subbranches separately, while the solid curves show the result of pruning the merger tree according to infall times. We see that the first spectrum is flatter overall, and extends to larger masses; it is roughly described by a Schechter function

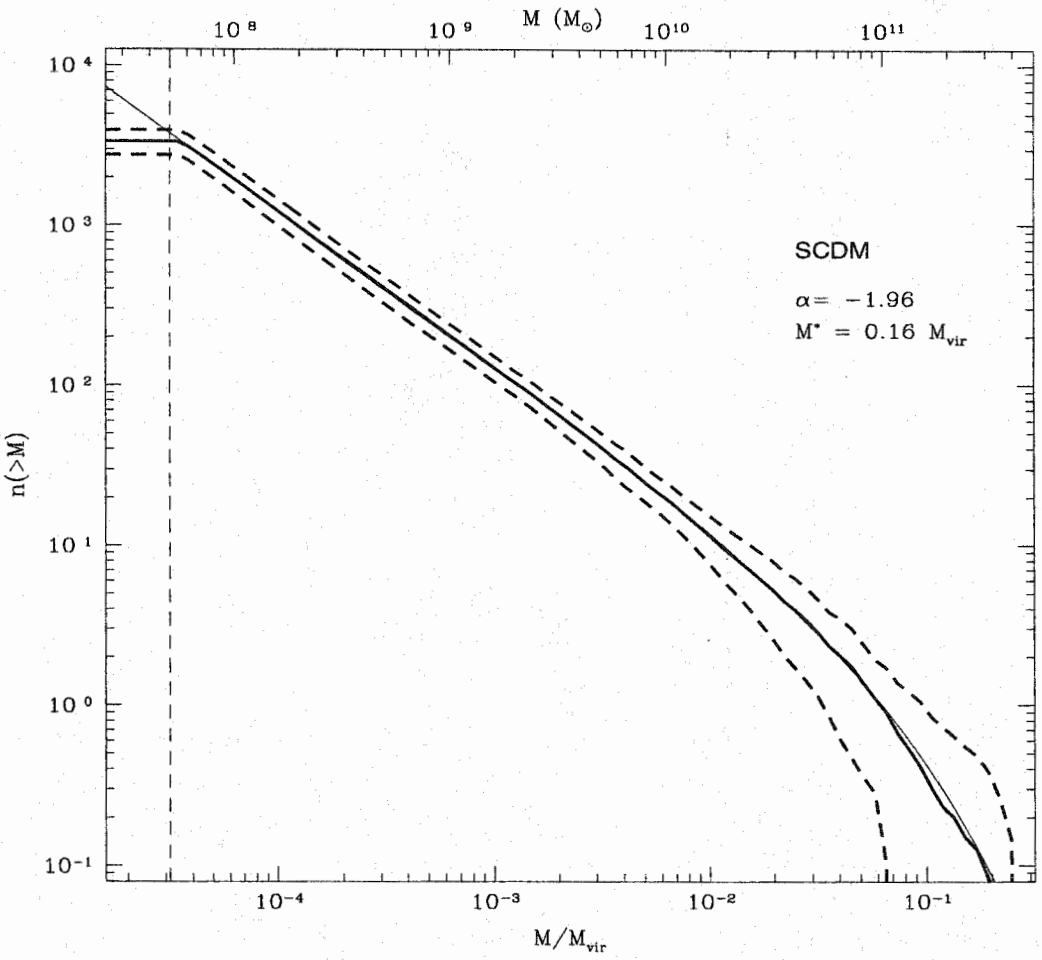

Figure 6.3 The cumulative distribution of all subhalos merging with the main trunk of the merger tree. The thick solid line is the average from 120 merger trees. while the dashed lines show the cosmological (1-σ) variance. The line solid line is an integrated Schechter-function fit with the slope and mass listed.

with $\alpha = -1.5$ and $M^* = 0.085\,M_{\rm vir}$. The other two cases are steeper, with slopes close to $\alpha = -2$. I will show in section 6.4 that the slope of the first distribution is inconsistent with numerical results for group and cluster halos, confirming that substructure in the side-branches does have to be accounted for in models based on merger trees. First, however, having explored the general properties of the merging halos, it remains to determine their detailed structure, and the structure of the

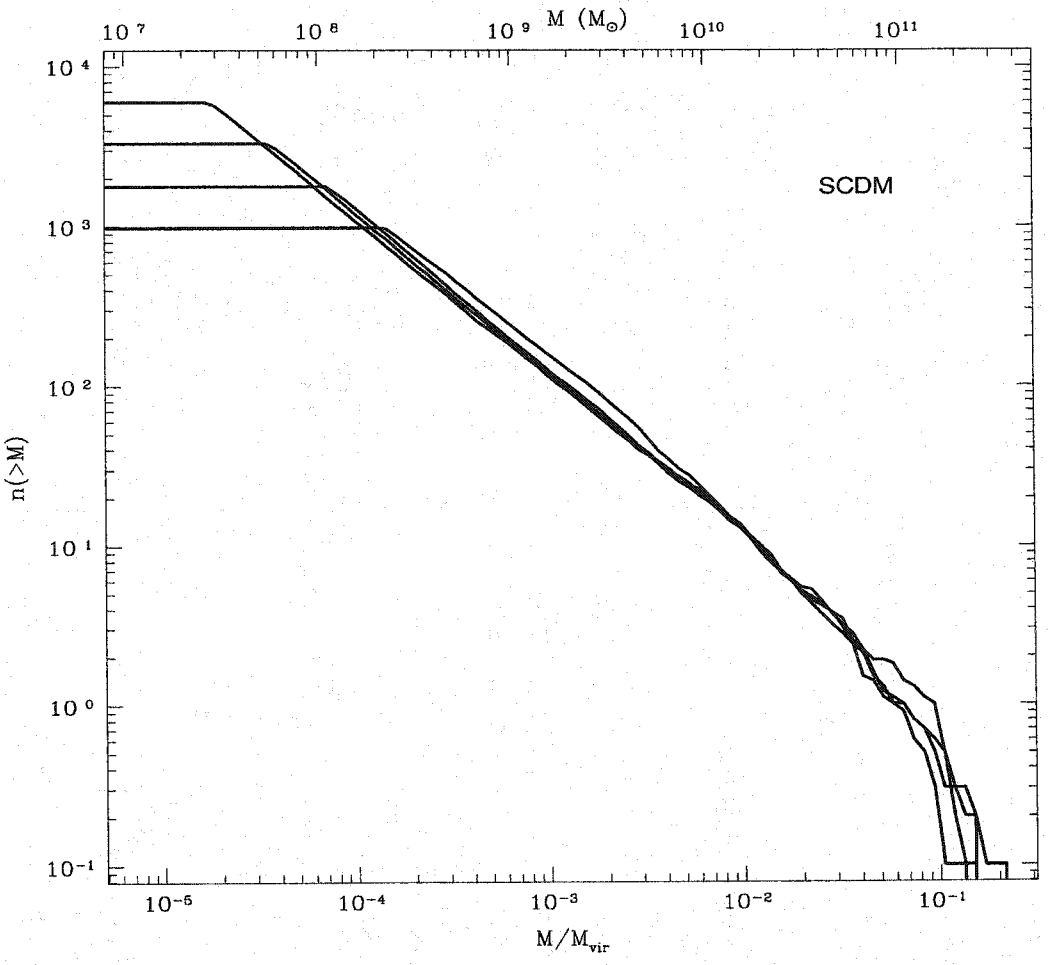

Figure 6.4 The cumulative distribution of all subhalos merging with the main trunk, calculated using mass resolution limits of $2.5 \times 10^7\,M_\odot$, $5 \times 10^7\,M_\odot$, $1 \times 10^8\,M_\odot$, and $2 \times 10^8\,M_\odot$. The distribution converges rapidly for masses over the resolution limit (the variation at the high-mass end is statistical).

potential they fall into. I will describe this part of the model in the next section.

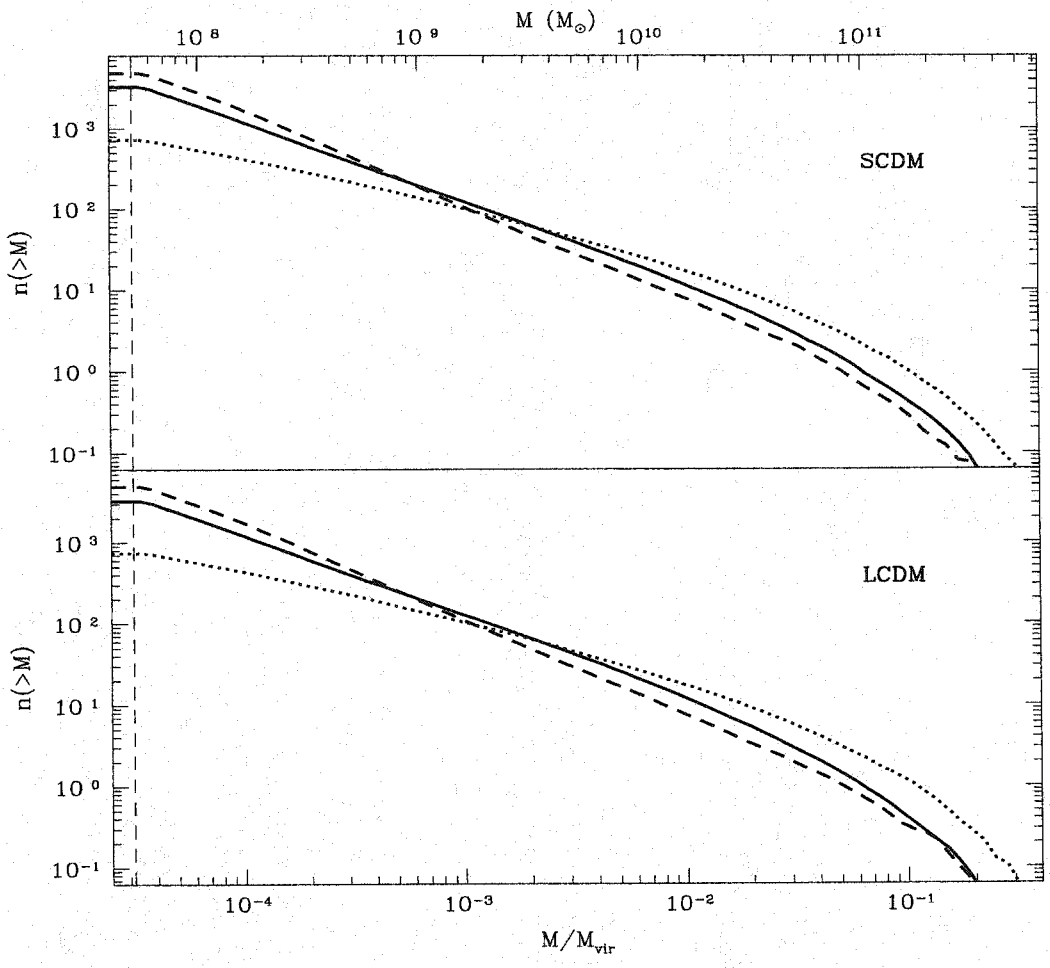

Figure 6.5 The effect of substructure in the side branches. The curves show the effect of counting each branch as a single object (dotted curves), of counting each sub-branch as a single object (dashed curves), or of 'pruning' according to subhalo infall times (solid curves) on the cumulative distribution of subhalo masses.

6.2 A Model of Galaxy Formation

6.2.1 Evolution of the Dominant Galaxy

The merger tree code described in chapter 5, together with the pruning method described above, provides a list of all mergers with the main trunk between the initial

and final redshift, as well as the mass in the trunk at each step[2] . This data provides the basis for a semi-analytic description of galaxy, group or cluster formation. In this section, I will describe the components of the model designed specifically to model galaxy formation, or more precisely the formation of a single large galaxy, with a number of small satellites in orbit around it. The model can easily be generalised to larger groups or clusters, by changing the mass of the merger tree and the form of the central potential; this is not the main purpose of this work, however, so I will not discuss these extensions in what follows.

Within the main halo (the 'trunk' of the merger tree), a large disk galaxy forms through dissipation of the energy in the halo gas, its collapse, and subsequent star formation. These processes are complex, and it is beyond the scope of this work to attempt to model them. Instead, they are simply assumed to produce a galaxy with a basic structure similar to the observed structure of the Milky Way, but structural parameters that depend on its evolutionary history. The two components of the galaxy model are a thin, exponential disk with an isothermal vertical density profile, and a spheroidal component, including both the bulge and the corona, (or stellar halo), with a Hernquist density profile. These components are similar in form to those used in the VW simulations, but change in scale as the main halo acquires material.

Since the merger trees trace the halo's mass accretion history back to an epoch when it has only a small fraction of its final mass, the central components must also grow during this accretion process. As the main halo merges with other halos, gas should cool and collapse into its centre, to add to the mass of the galactic disk. In the simplest model, the disk mass would increase instantaneously and in proportion whenever the halo mass increased, so that one mass tracks the other. In a more sophisticated model, gas infall would be further delayed by the cooling time. The cooling

[2]This 'pruning' is actually included in the code, and accounts for much of the computational effort of generating the trees.

time is more complicated to calculate, however, depending on the mass, density and metallicity of the system. As an approximation to this more realistic situation, I will consider a model where halo gas is added to the disk over the infall time, which is longer than the cooling time for galaxy-sized objects (White & Rees 1978). As shown in figure 6.6, the effect of this delay is fairly minor.

The overall efficiency with which the gas content of the halo is added to the disk is a free parameter; I set it to 0.5 for the SCDM model and 0.8 for the LCDM model, which, when multiplied by the baryon fraction of the universe in either model, gives a disk mass of approximately $5.6 \times 10^{10}\,\mathrm{M}_{\odot}$ at the present day, in keeping with recent mass models for the Milky Way (Dehnen & Binney 1998). Major mergers may disrupt the disk and convert it into a spheroidal component; I will discuss this process in detail in chapter 8.

Finally, the (radial) scale length of the disk, r_{d}, grows as $M_{\mathrm{d}}^{(\alpha-2)/\alpha} = M_{\mathrm{d}}^{1/3}$. This will produce a Tully-Fisher relation of slope $\alpha = 3.0$, if one assumes that the total luminosity of the disk is proportional to its mass, and that the peak circular velocity of the disk is proportional to $\sqrt{M_{\mathrm{d}}/r_{\mathrm{d}}}$. While the peak circular velocity of the disk actually depends on the total mass distribution in the inner part of the halo, the tightness of the observed Tully-Fisher relation suggests that the latter proportionality is maintained in practice, in most real galaxies. The scale height of the disk is taken to be 1/5 the radial scale length at all times, mainly due to the computational requirement that the form of the potential remain the same throughout the run.

The other component of the potential, the spheroid, grows in quite a different way. I assume that this component forms entirely from the debris of merging, following two different processes. The first is tidal disruption of satellite galaxies, which I describe in detail in the next chapter. The stellar contents of any satellite which is stripped beyond some point, or which falls too deep into the potential, are added to the mass of the spheroid in the corresponding timestep. Furthermore, if the disk experiences a sufficiently violent encounter, it is assumed to be entirely disrupted, and its whole

mass is added to the spheroid[3] . This process is also described in more detail in chapter 8. When the mass of the spheroid changes, its scale length is assumed to follow a Fish (or constant central surface brightness) law: $r_s \propto M^{1/2}$.

Figure 6.6 shows the mass accretion histories of the different components in a typical (SCDM) merger tree. The disk grows in proportion to the main halo, with a slight lag, as shown in the top panel (dashed line), while the spheroid grows due to satellite and disk disruption, as shown in the second panel. The third panel shows the growth of the bulge, while the bottom panel shows the resulting disk-to-spheroid mass ratio versus time.

6.2.2 Evolution of the Subhalos

Having determined the basic properties of the dominant galaxy within the main halo, the evolution of substructure must now be taken into account. Each merger in the pruned merger tree corresponds to a single object crossing the virial radius of the main halo. If many halos merge with the main halo at the same time, this corresponds to a single group merging with the system, as explained above. For each group or independent subhalo, orbit is assigned, passing through a random point at one virial radius from the centre of the main halo. The radial velocity is taken to be negative and equal to the infall velocity of the main halo, that is simply its circular velocity at the virial radius, $\sqrt{GM/r_{\rm vir}}$, the velocity a particle falling in from rest at the turn-around radius would have at this point. Numerical simulations (Tormen 1997) have shown that this is in fact close to mean radial velocity for satellites first crossing the virial radius. A random circularity is chosen from the distribution measured by Ghigna *et al.* (1998), and the orbit is given the corresponding tangential velocity, with a random orientation in the tangent plane.

Coherent groups of objects should have correlated initial positions and velocities,

[3]Throughout the run, both the disk and spheroid are actually given minimum seed masses, to prevent numerical instabilities in the code, but these are small enough to be negligible.

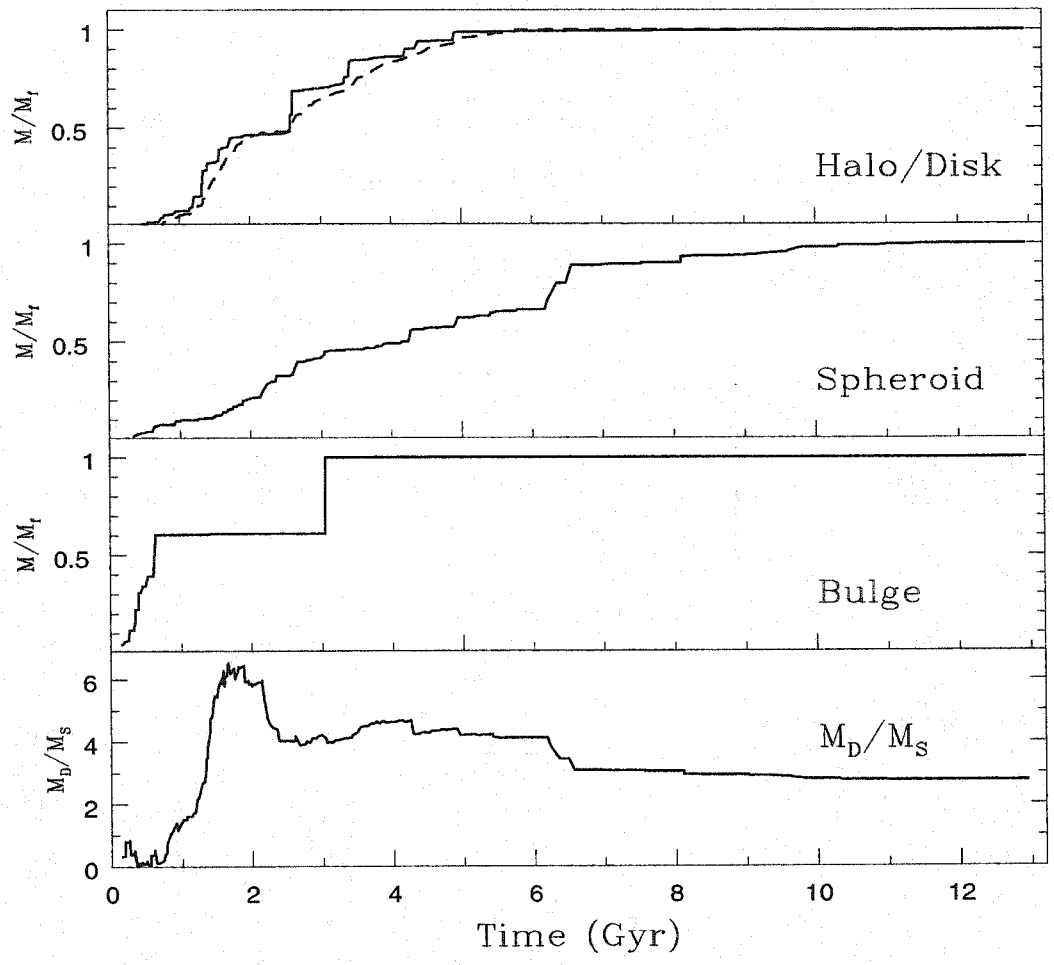

Figure 6.6 The mass of the halo or disk components (top panel, solid and dashed lines), the spheroidal component (second panel) and the bulge component (third panel) vs. time, normalised to their final mass, in an individual SCDM tree. The bottom panel shows the resulting disk-to-spheroid mass ratio vs. time.

with some scatter related to the structure of the group before it merges with the main halo. To reflect this, a single orbit is chosen for each group, and the individual members are then assigned positions and velocities which deviate from this mean by small random offsets. The offset in position is chosen from a uniform distribution of width $\Delta R = r_{\rm vir,g}$, while the offset in velocity is chosen from a Gaussian distribution of width $\Delta V = V_{\rm vir,g}$, where $r_{\rm vir,g}$ and $V_{\rm vir,g}$ are the virial radius and circular velocity

of the whole group, respectively.

Finally, the initial structure of each individual subhalo must be specified. Halos should be dominated by dark matter in all but their central regions, particularly on small mass scales. Thus each halo is represented by the analytic profile:

$$\rho(r) = \frac{\rho_0 r_s^3}{r_s^{1.5}(r + r_s)^{1.5}} \quad , \tag{6.3}$$

found by Moore *et al.* (1998) to describe the general profile of halos seen in high-resolution simulations. There is continuing controversy over the exact inner slope of this profile, but in practice these differences have very little effect on the results described below. In particular, a Moore profile and an NFW profile, which has a flatter inner slope, are almost identical outside the central region if they have the same outer radius and peak circular velocity radius.

Given a mass and a mean density set by the virial density estimated in chapter 5, this profile has a single free parameter, the ratio of the outer radius to the scale radius, called the concentration. While there is no analytic prediction for the concentration of halos with this exact profile, work by Bullock *et al.* (2001b), NFW (1996, 1997) and Eke *et al.* (2001, ENS hereafter) has suggested relations between concentration, mass and redshift for the similar NFW profile, which can be related to Moore concentrations if the prediction is written in terms of the ratio of the virial radius to the peak radius. The predicted concentrations from ENS, the most recent and extensive of these studies, are shown in figure 6.7. There are several caveats in using these predictions, however. First, they were derived by studying isolated halos, rather than subhalos; second the halos studied were more massive than those considered here, and third, there was considerable scatter in the relations, even under these restricted conditions.

Ultimately, the concentration specifies the central density of a halo; for subhalos in clusters, this density may be strongly affected by tidal heating and mass loss. I will show below that the concentration relations of ENS give reasonably good results,

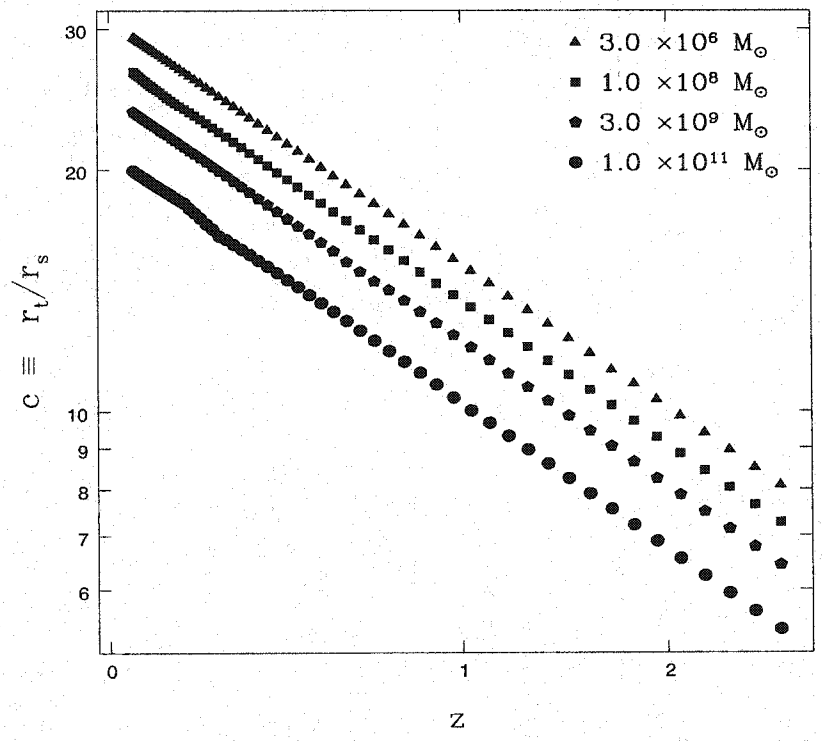

Figure 6.7 The concentration of dark matter halos of different mass (four sets of points), predicted by ENS, as a function of formation epoch.

provided that the epoch used to determine the halo's mean density is the merger redshift $z_{\rm m}$, typically 1–3 for most halos, rather than the earlier redshift $z_{\rm mo}$, which is typically 5–10. This result is somewhat surprising, as one might expect the density of subhalos to reflect their own formation conditions, rather than the density of the larger system into which they had fallen subsequently. Whether this relatively low density is the result of heating and mass loss in the sub-branches, or whether the analytic concentrations determined by Bullock *et al.* (2001b) and ENS are no longer valid beyond redshift $z \geq 2$, will require further investigation. In any case, the system used here is shown to agree reasonably with the numerical results in the next section.

6.3 Comparison with Numerical Simulations

So far, most of the processes I have described involve only dissipationless dynamics. As a result, it is possible to evaluate the accuracy of the model up to this point by direct comparison with numerical simulations. A few high-resolution simulations of individual halos do exist; the current list includes one massive SCDM cluster, the 'Coma' cluster, one smaller SCDM cluster, the 'Virgo' cluster, and a pair of SCDM galaxy halos, the 'Local Group' halos, all by Quinn and collaborators (Moore *et al.* 1999), as well as two LCDM clusters, one by Navarro *et al.* (Taylor & Navarro 2001) and one by Springel *et al.* (2000). These simulations are self-consistent and based on very simple assumptions, so they offer far more robust predictions of dark substructure than the semi-analytic model, which excels instead in speed and flexibility. Some aspects of the simulations are subject to resolution limits or other numerical effects, and here the comparison with the semi-analytic model may be particularly interesting. In any case, it is clearly important to test how well the SA model reproduces these simulation results, before applying it to other problems.

Since halo properties are predicted to show a certain amount of cosmological scatter, it is important to compare the SA results against as many simulations as possible. To do this, the masses, lengths and velocities in each simulation have to be scaled to the same fiducial values used in the merger trees. The simulations do not contain a central disk or baryonic component, so I ran a set of trees without these components present. The smoothed density profiles of the main halos in each simulation have already been determined to be close to the form given by equation (6.3), and are the basis for the profile used in the SA trees. Instead, I will focus on the properties of the substructure.

Substructure in the simulations was identified by the authors using the group finder SKID, which assigns particles to the nearest local minimum in the potential,

and determines masses, radii and velocities for the resulting groups. These measurements can be affected by numerical noise (especially in low-mass halos) and the position of the subhalo within the main halo (see chapter 2 for a discussion of some of these uncertainties, as well as the SKID website (http:// www-hpcc.astro.washington.edu / tools / skid.html) for a description of SKID, and Klypin *et al.* (1999a) for a comparison of group-finding algorithms). Of the group properties produced by SKID, I took the total mass and the peak circular velocity to be the most reliable, as they are less subject to random fluctuations due to single particles, and the effects of the background density field, than other properties such as the outer velocity.

Figure 6.8 shows the cumulative mass function for halo substructure averaged over 120 SCDM trees (thick solid line), compared with the results from the SCDM simulations mentioned above (thin dotted lines), scaled by the ratio of the mass of the main system (the mass within its virial radius) to the fiducial mass $1.6 \times 10^{12}\,\mathrm{M}_{\odot}$. Two of the dotted lines correspond to the Virgo simulation at two different epochs, two correspond to the two main halos in the Local Group simulations, and one (the steepest) is from Coma. The thick dashed lines indicate the cosmological variance from tree to tree in the SA model. We see that the slope, normalisation and variation of the SA results match the numerical results quite well, although there is a marginally significant offset of about 20% in the normalisation. Only the Coma simulation shows a slightly different distribution, with a similar normalisation around a mass ratio of 10^{-3}, but a slightly steeper slope. Whether this is due to its much larger mass ($2.4 \times 10^{15}\,\mathrm{M}_{\odot}$), a different set of parameters for SKID, or simply represents cosmological variance is unclear. In any case, this discrepancy exists within the simulation results, and needs to be clarified by further numerical work.

Figure 6.9 shows a similar comparison of the distribution of peak circular velocities in the SA merger trees and the simulations. Line styles are as in figure 6.8. The numerical results have been normalised to the mass used in the SA trees by assuming that the subhalo velocities within a halo of mass M_{vir} scale as $V \propto M_{\mathrm{vir}}^{1/3}$. This

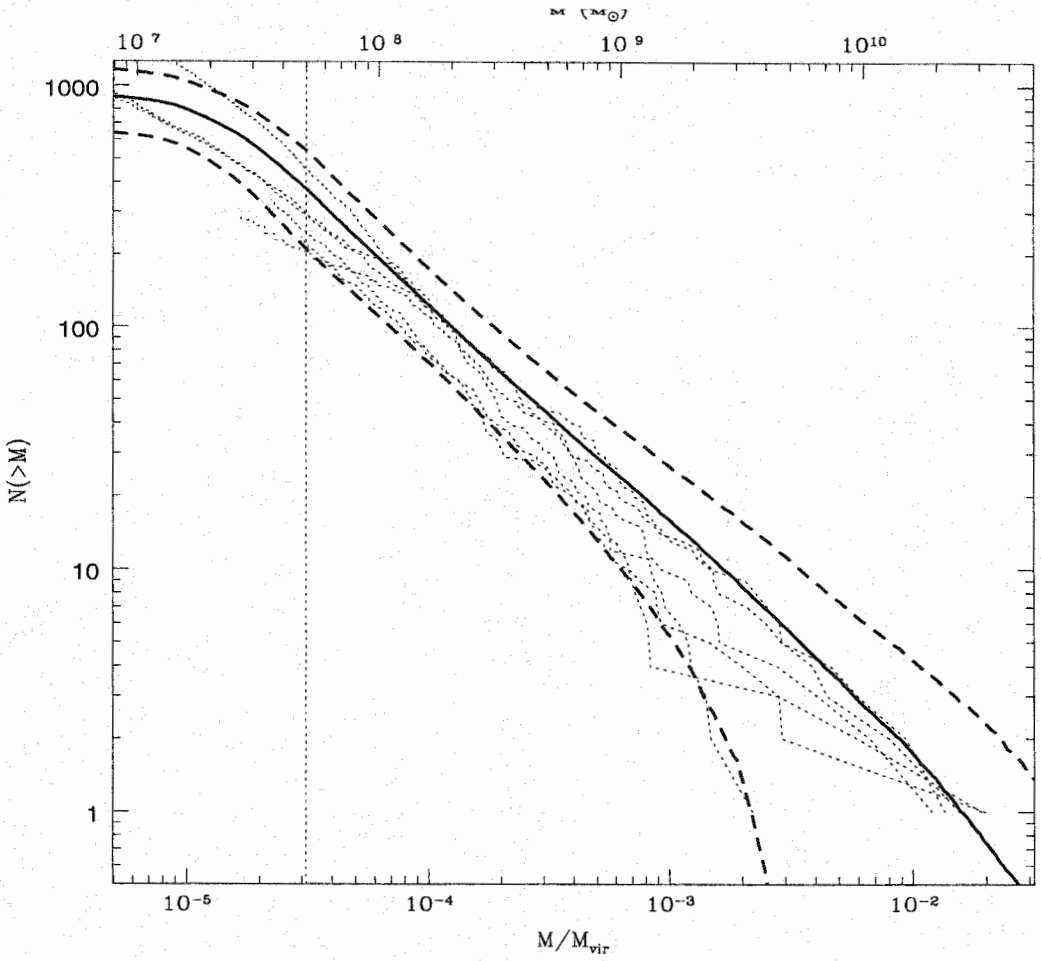

Figure 6.8 Comparison of the cumulative mass function of surviving halos within the virial radius at late times, in the SA model (thick solid line, with dashed lines indicating the 1-σ variance), and several numerical simulations (thin dotted lines). Both sets of results have been normalised to the total mass within the virial radius. The top axis gives the equivalent mass in a galaxy-size halo. The vertical line indicates the resolution limit of the SA model.

normalisation produces the smallest scatter in the numerical results; normalising by the peak or outer circular velocity of the main halo (e.g. Moore *et al.* 1999) produces a slightly larger offset between the numerical simulations of galaxy halos and those of cluster halos, because the concentration of the background density profile is slightly

different.

The SA peak circular velocities are determined by the initial concentration of the subhalos, and their subsequent structural evolution within the main halo. As discussed in chapter 4, as subhalos lose mass, their peak velocity decreases as $m^{1/3}$. The initial concentrations are those predicted by the ENS formula, for a formation redshift corresponding to the epoch $z_{\rm m}$ when the subhalo falls into the main system. The SA results match the distributions from the Local Group simulations (lower dotted curves) reasonably well, up to velocities of 50 $\rm km\,s^{-1}$; the cluster distributions appear to be somewhat steeper. Shifting the SA velocities up by 20% produces a reasonable match to both sets of results. This could indicate that subhalos in simulations are slightly more concentrated than predicted by the SA model, possibly because the halos retain some memory of their initial formation epoch (although using the original merger epoch, $z_{\rm mo}$, to determine the concentration produces halos that are much too concentrated). It is also possible that stripping and heating modify the subhalo profiles in a more complicated way than suggested in chapter 4, or that the ENS concentrations are incorrect on the small mass scales considered here. Finally, given the small number of simulations available, it is unclear whether the difference between the clusters and the galaxy halos is significant, in which case the SA prediction may be quite accurate. Investigating these different possibilities requires further numerical work; in the interim, I will assume a 20% uncertainty in the peak velocities predicted by the SA model.

It is also interesting to compare the spatial distribution of substructure in the numerical and SA halos. Figure 6.10 shows the number density of halos within some fraction of the virial radius, normalised to the number density within the virial radius. The jagged solid line is the average semi-analytic result, while the lines with error bars are the simulation results. The solid line with points and error bars is for all subhalos with $M/M_{\rm vir} > 10^{-4}$ in the Coma simulation, while the dotted line with points and error bars is for all subhalos with $M/M_{\rm vir} > 3 \times 10^{-4}$ in the Virgo simulation. The

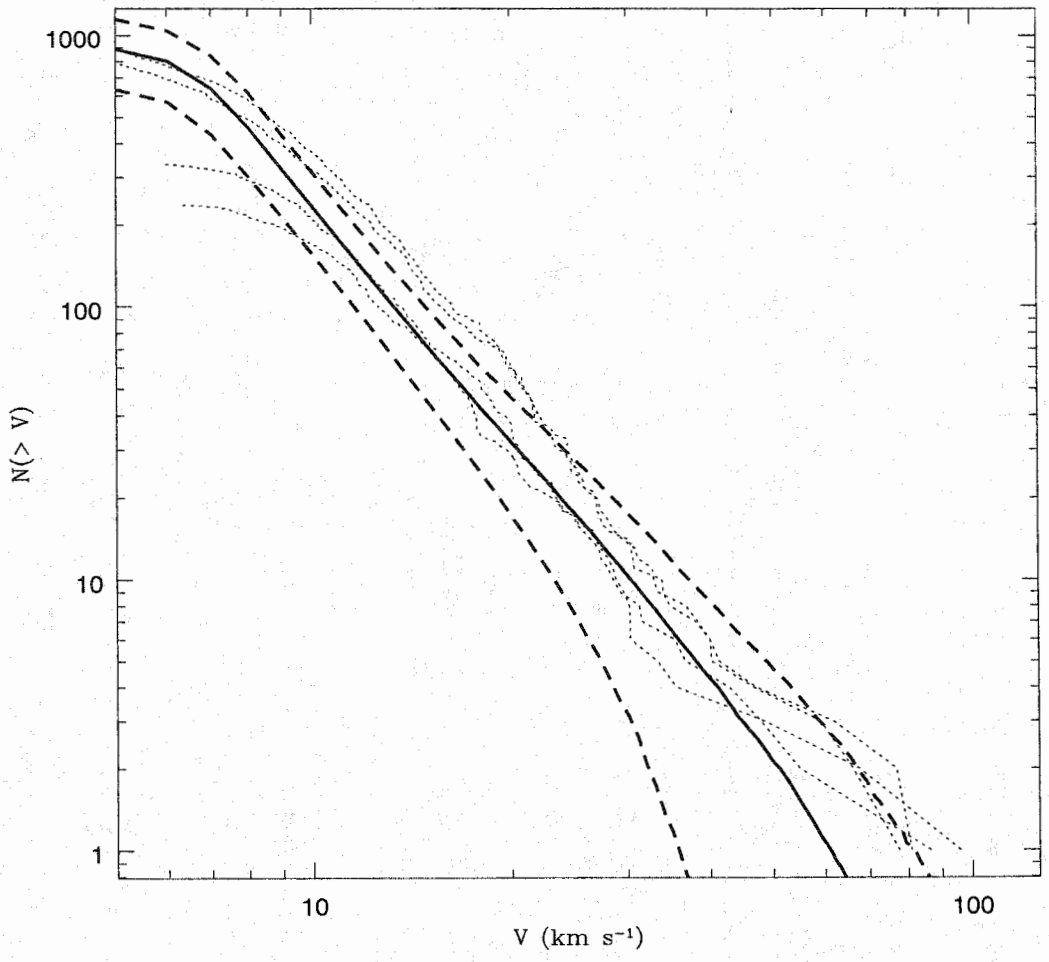

Figure 6.9 As figure 6.8, but comparing the distribution of peak velocities. Velocities for the simulations have been normalised by the mass of the system relative to the SA model, to the 1/3-power.

SA distribution is for $M/M_{\rm vir} > 10^{-4}$. The error bars are Poissonian for the two simulations. We see that the SA distribution is broadly consistent with the numerical results, although they are sufficiently noisy that detailed comparison is difficult. Both SA and numerical density profiles are much shallower than the density profile of the smoothly distributed background material (dashed line).

One notable difference between the SA model and the simulations considered here is the presence of the central disk. The disk has a complicated effect on satellites,

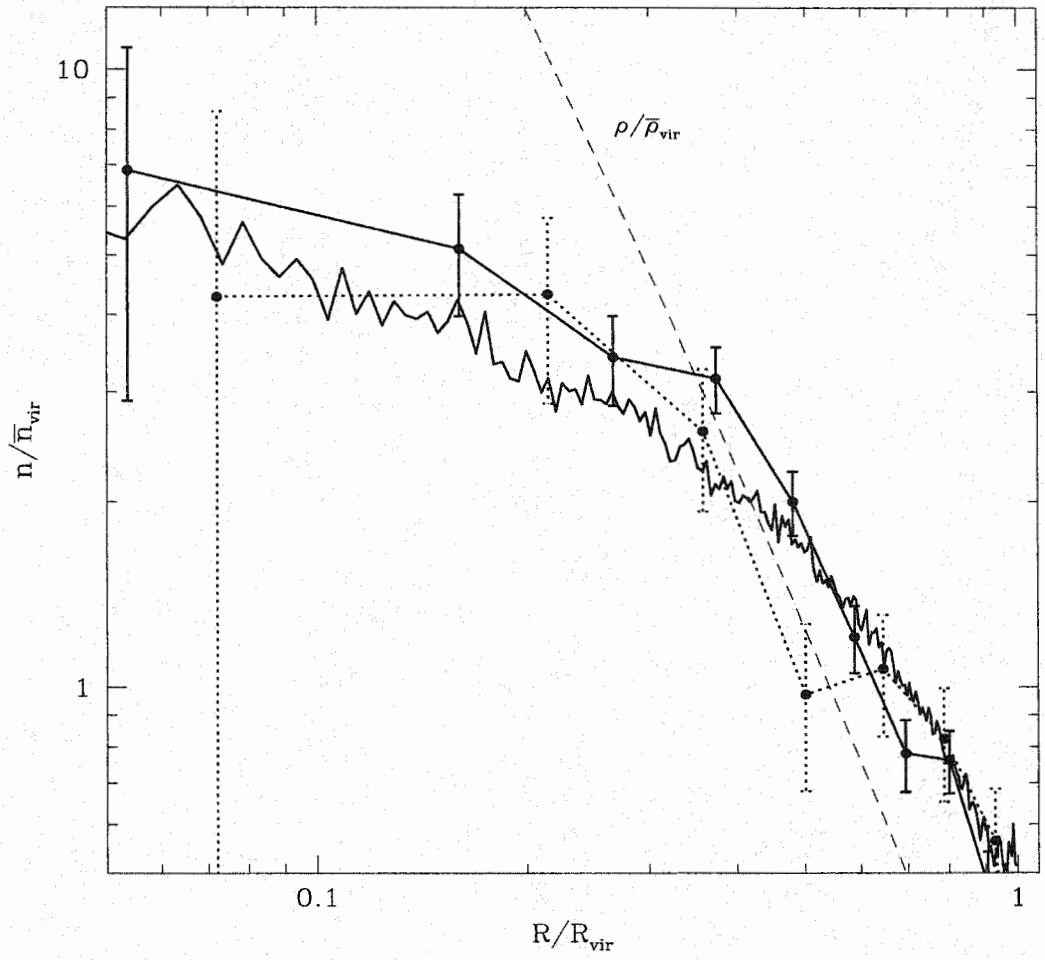

Figure 6.10 A comparison of the number density of subhalos within the SA halos (solid line) and the numerical halos (solid and dotted lines with points and error bars), normalised to the number density within $R_{\rm vir}$. The solid line with points and error bars is for all subhalos with $M/M_{\rm vir} > 10^{-4}$ in the Coma simulation, while the dotted line with points and error bars is for all subhalos with $M/M_{\rm vir} > 3 \times 10^{-4}$ in the Virgo simulation. The SA distribution is for $M/M_{\rm vir} > 10^{-4}$. The dashed line indicates the background density (Moore) profile, normalised to the mean density within the virial radius.

increasing their mass loss rates, but also decreasing their orbital decay rates due to dynamical friction as a result. Figure 6.11 shows cumulative mass functions in models

with and without the disk. The curves are for all satellites within the virial radius, 160 kpc, 80 kpc, 40 kpc, 20 kpc and 10 kpc (from top to bottom). We see that the disk disrupts massive satellites within 40 kpc; at larger distances or for small masses, the effect on the cumulative mass function is minimal.

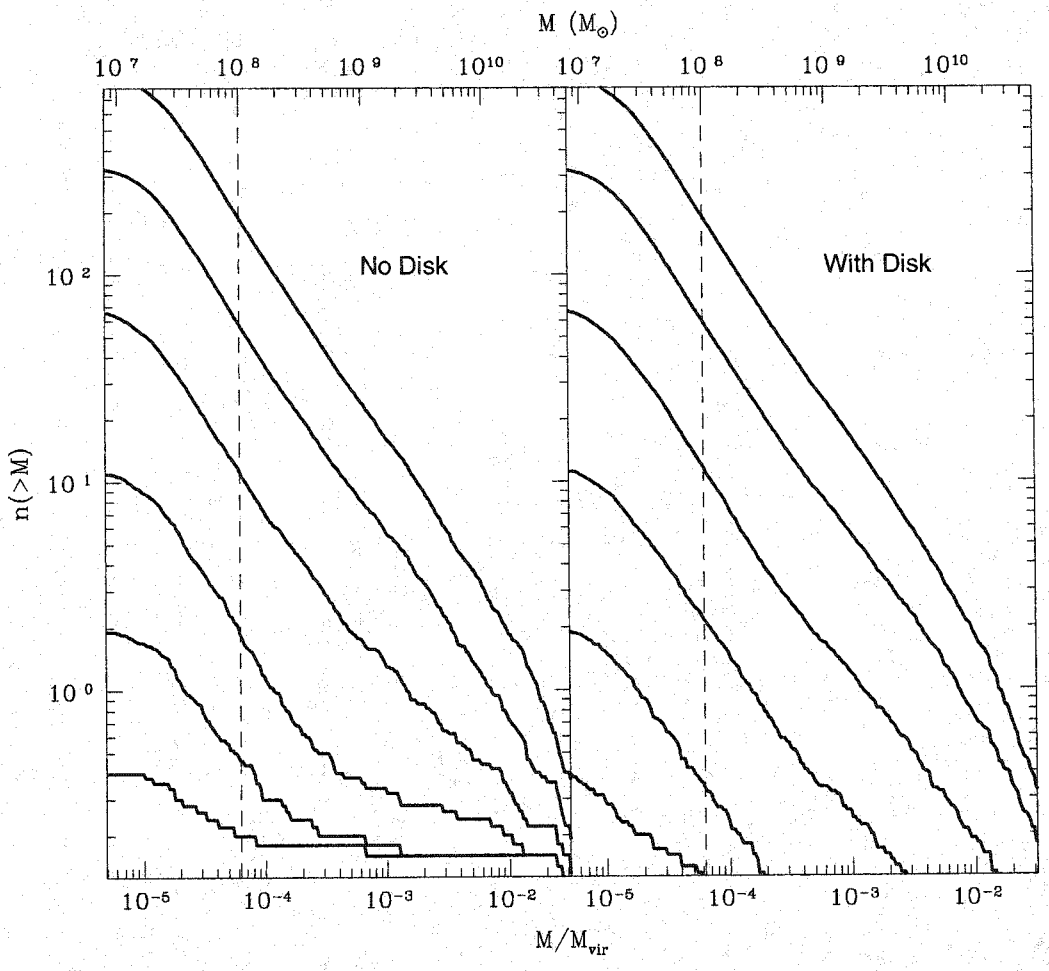

Figure 6.11 The effect of the disk on satellites, as a function of mass and distance from the centre of the main halo. The different mass functions are for satellites within the virial radius, 160 kpc, 80 kpc, 40 kpc, 20 kpc and 10 kpc, from top to bottom.

Finally, it is worth making some basic remarks about how substructure builds up within a halo. From EPS theory, the total mass of the halo is expected to grow as

$t^{2/3}$ or $(1+z)^{-1}$ in an SCDM cosmology, so much of the mass of the halo is assembled fairly early on. The size of the system goes as $(1+z)^{2/3}$, so the halo will also be smaller at early times. Subhalos crossing the virial radius at some epoch will fall to the centre, pass through it, move out towards their initial turnaround radius, and then return again over roughly one Hubble time (that is approximately the age of the universe at the time they first fell in). Many subhalos will also be disrupted, in the first or subsequent passages through the pericentre of their orbit. The halo seen at any epoch will thus contain well-mixed material from early merger epochs, distributed fairly evenly in its core, as well as shells of material falling in or moving back outwards, according to their dynamical 'age'. Figure 6.12, for instance, shows the position, velocity, remaining mass and original mass for the satellites in a typical (SCDM) merger tree at $z = 0$, as a function of the time elapsed since they entered the main halo, $\Delta t_{\rm m}$. Figure 6.13 shows the total number of satellites vs. time in a number of LCDM halos, in three mass ranges: all satellites (solid curves), those with $M > 5 \times 10^8\,{\rm M}_\odot$ (dotted curves), and those with $M > 5 \times 10^9\,{\rm M}_\odot$ (dashed curves). The numbers increase for the first few Gyr as the halo forms, and then decrease slowly thereafter as infall slows down and satellites are stripped and disrupted. This general structure will also be obscured by correlations between the positions and velocities of individual subhalos within larger groups, however, as discussed in the previous section. In general, the radial distribution of subhalos from a given merger era is expected to remain roughly constant in spatial scale, but to decrease in normalisation as halos are disrupted. These general features provide interesting observational tests of the SA model, using high-redshift clusters.

6.4 Summary

In this chapter, I have discussed how to apply the rather abstract models of merging used to generate merger trees to the more concrete problem of galaxy formation. This requires three main additions to the merger tree code described in the previous chapter: a mechanism for 'pruning' the tree, to reduce it to a series of mergers with the main trunk, a model for the evolution of the main system, and the dynamical model for subhalo evolution developed in part I.

Adding these components to the merger tree code produces a full model of galaxy formation, in which the main structural features of a galaxy and its halo, as well as the masses, peak velocities and orbits of its subhalos, are specified at all times. The behaviour of the central components is based on empirical recipes, designed to reproduce the features of observed galaxies. The evolution of substructure is a prediction of the model, however, and can be compared directly with numerical simulations of halo formation. This comparison shows that the SA model reproduces numerical results quite accurately, matching the distributions of masses, peak velocities and positions within the halo to within 20% or better. Some minor discrepancies in the peak velocity distributions and the number of objects in the centre of the system may be artefacts of the SA model or the numerical simulations, but the overall agreement between the two is good. In the next two chapters, I will use the SA model of galaxy formation to address the two main questions raised in the introduction, namely how the stellar contents of the halo could be produced by hierarchical structure formation, and how the thin disk could have survived this process.

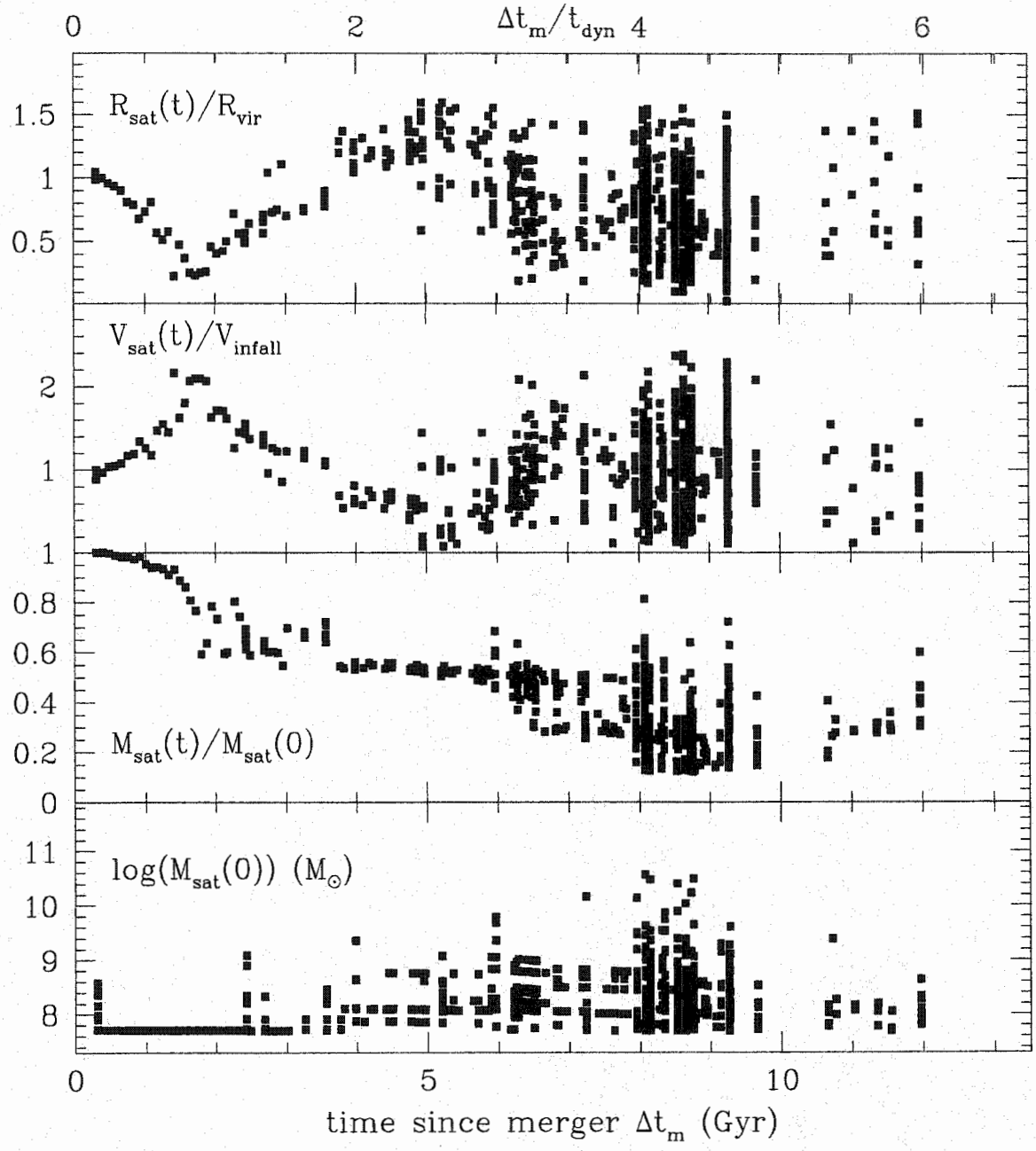

Figure 6.12 The position, velocity and remaining mass, normalised to their original values (top three panels), as well as the original total mass (bottom), of the satellites in a typical (SCDM) merger tree at $z = 0$, as a function of the time elapsed since they entered the main halo, Δt_m. The top axis shows Δt_m as a function of the dynamical time.

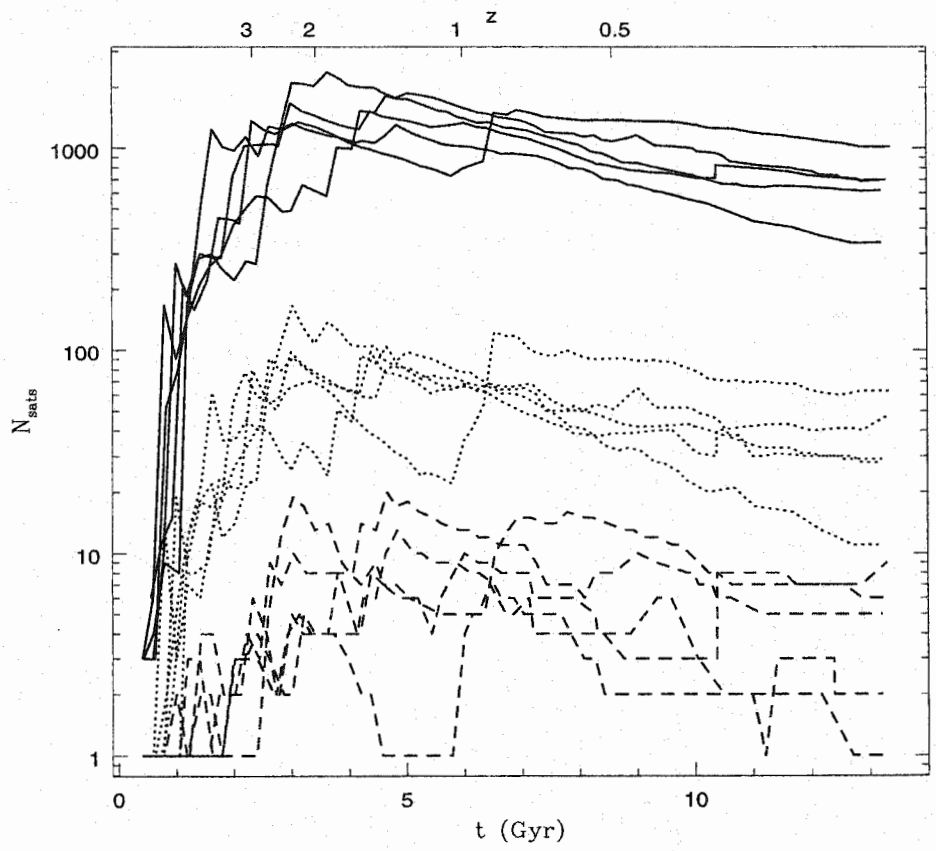

Figure 6.13 The total number of satellites in several LCDM halos as a function of time, in three mass ranges: all satellites (solid curves), those with $M > 5 \times 10^8\,\mathrm{M}_\odot$ (dotted curves), and those with $M > 5 \times 10^9\,\mathrm{M}_\odot$ (dashed curves).

Chapter 7

The Stellar Contents of Galactic Halos

In this chapter, the semi-analytic model of galaxy formation developed in chapters 5 and 6, which was based on the dynamics of dark matter halos and subhalos, is combined with a simple model of star formation, to predict the stellar contents of galaxy halos. I describe the observed stellar substructure in the halos of the Milky Way and Andromeda, and show that in standard CDM cosmologies, there cannot be a one-to-one correspondence between dark matter subhalos and satellite dwarf galaxies. The large discrepancy between the two may be a result of the reduced efficiency of star formation in less massive halos; a simple model of the truncation of star formation below a threshold circular velocity after the epoch of reionisation, for instance, produces a satellite luminosity function comparable to that seen in the Local Group. The structural properties and spatial distribution of the satellites produced by this model are also consistent with the observed properties of Local Group dwarfs. The more massive satellites in the Local Group appear to be systematically brighter than predicted, however. I present one possible scenario for the star-formation history of dwarf galaxies, and show that it produces a better match to the observed luminosity function, and explains the morphologies and star-formation histories of individual satellites in the Local Group. Many satellites are disrupted by tidal effects; their stellar debris presumably contributes to the bulge and stellar halo of their parent galaxy. The masses, density profiles and ages of the stellar halos formed by this process are shown to be consistent with local observations. On average, a number of satellites will have been disrupted recently around any given stellar galaxy, so the existence of coherent tidal streams in the stellar halo also seems plausible in this model.

Halos of cold dark matter are relatively simple and almost scale-invariant. Extensive numerical simulations have established their formation rates (e.g. Tormen 1998; Gross *et al.* 1998; Governato *et al.* 1999), their density profiles (e.g. Navarro *et al.*

1996, 1997; Moore *et al.* 1998), their net angular momentum (e.g. Barnes & Efstathiou 1987; Ryden 1988; Warren *et al.* 1992; Steinmetz & Bartlemann 1995; Cole & Lacey 1996) and the masses, circular velocities and spatial distribution of their substructure (e.g. Ghigna *et al.* 1998; Klypin *et al.* 1999b; Moore *et al.* 1999; Springel *et al.* 2000), fairly consistently, although there is still disagreement on certain points. In the previous chapter, I showed that a semi-analytic model can generate dark matter halos which have these basic properties, and resemble numerical halos in detail as well. In real-world galaxy halos, luminous substructure is much easier to observe than dark substructure, however, and thus the current challenge in galaxy formation models, whether numerical or semi-analytic, is to develop plausible models for star formation.

Unfortunately, star formation is an inherently complicated process, sensitive to the geometry, size and chemical composition of the star forming region, and as a result is much harder to model than the dynamical evolution of dark matter. This is particularly true on the scale of dwarf galaxies, which have only small reservoirs of cold gas from which stars can form, and whose potentials are weak enough that the energy balance in their gas is very sensitive to internal processes and environmental effects (Babul & Rees 1992). Any model of star formation, whether numerical or semi-analytic, is necessarily approximate; the basic energy exchanges within a star-forming region occur on scales thousands of times smaller than the finest numerical resolution limit, and their geometry and interaction is sufficiently complex to defy an analytic treatment. Numerical models of these processes benefit from a full and robust treatment of the large-scale context for star formation in galaxies, and are starting to produce convincing results (Gerritsen & de Blok 1999; Navarro & Steinmetz 2000; Springel 2000). Semi-analytic methods, while less robust, are much faster, which is an important advantage given the model-dependence and the number of free parameters in star formation recipes. In this chapter, I will add a very simple model of star formation to the semi-analytic model of halo evolution developed in the last chapter,

and test its predictions for the stellar substructure of galactic halos, to determine whether local observations are broadly consistent with hierarchical galaxy formation.

7.1 Local Observations of Halo Substructure

The distribution of stars within galactic halos is best observed locally, both because they are too faint to see in more distant systems, and because it is possible to get a clearer sense of their 3-D distribution in nearby space. The Milky Way is part of the Local Group of galaxies, an overdense region about 1 Mpc in size. In structural terms, the Local Group probably consists of two or possibly more overlapping halos, those of the Milky Way and of the Andromeda galaxy, giant spiral galaxies of similar size and morphology. (A number of other nearby galaxies exist in subgroups, some of which may be bound to the Local Group, but most of which lie outside it.) The contents of these halos offer the best observational comparison with models of halo substructure. Beyond their central stellar components, each contains roughly a dozen satellite dwarf galaxies. The basic properties of these satellites are listed in tables 7.1 and 7.2 below.

The dwarfs vary in luminosity, morphology, (cold) gas content and star-formation history, with some evidence for systematic trends in the latter properties with luminosity. In particular, some morphologies are exclusive to low-luminosity or high-luminosity systems. I will discuss this further below. It is also worth noting that the lists in tables 7.1 and 7.2 are probably incomplete below $M_V = -10$ (van den Bergh 2000), and that there may be other substructure in the Local Group with cold gas but little or no associated starlight, the so-called 'high-velocity clouds' of neutral hydrogen (Blitz & Robishaw 2000), although this remains controversial due to the lack of well-established distances to these objects (Zwaan 2001). The halos of luminous galaxies also contain tens or hundreds of globular clusters, aggregates of

Table 7.1 Dwarf Galaxy Satellites of the Milky Way

Name	Type	M_V	L_V ($L_\odot$)	V_c or ($\sqrt{2}\sigma$)	size (′)	[Fe/H]	star formation	cold gas
LMC	Irr	−18.05	1310	79.0	645	−0.8	active + burst	yes
SMC	Irr	−16.7	335	42.0	320	−1.1	active + bursts	yes
Sagittarius	dSph-N	−13.4	18.1	21.6	600	−1.0	old ?	no
Fornax	dSph	−13.2	15.5	14.8	71	−1.3	old + 1 burst	?
Leo I	dSph	−11.9	4.79	12.4	12.6	−1.5	old + 2 bursts	no
Sculptor	dSph	−11.1	2.15	9.33	76.5	−1.8	old + 1 burst?	yes
Leo II	dSph	−9.6	0.58	9.5	8.7	−1.9	old	no
Sextans	dSph	−9.5	0.50	9.33	160	−1.7	old + 1 burst	yes
Carina	dSph	−9.3	0.43	9.6	28.8	−2.0	old + 3 bursts	no
Ursa Minor	dSph	−8.9	0.29	13.2	50.6	−2.2	old	no
Draco	dSph	−8.8	0.26	13.4	28.3	−2.0	old	no

stars less massive but much denser than most dwarf galaxies. Given their numbers and densities, the origin of most globular clusters is probably very different from that of the dwarf satellites, but some of the most massive clusters may be the nuclei of disrupted dwarf galaxies (e.g. ω Centauri in the Milky Way (Hilker & Richtler 2000; Lee *et al.* 1999) and Mayall II in Andromeda (Meylan *et al.* 2001)). Finally, I note that while the properties of the recently discovered dwarfs spheroidals And IV–VI are poorly determined, they are probably similar to those of And I–III (Mateo 1998).

How do these observations compare with the substructure predicted by dissipationless models of halo formation? The answer, unfortunately, is not very well. The total mass of the Local Group dwarfs is highly model-dependent, but the circular velocities or velocity dispersions of their stellar components should give some indication

Table 7.2 Dwarf Galaxy Satellites of Andromeda

Name	Type	M_V	L_V ($L_\odot$)	V_c or ($\sqrt{2}\sigma$)	size ($'$)	[Fe/H]	star formation	cold gas
M 33	Sc	−18.9	2870	130	70.8	−1.1	active	yes
M 32	E	−16.7	383	65.1	9	−1.1	old ?	no
NGC 205	E	−16.6	366	10.0	6.2	−0.8	old + 2 bursts	no
NGC 185	dE	−15.5	125	35.4	16	−1.22	old + 1 burst	?
NGC 147	dE	−15.5	131	31.1	20	−1.1	old	no
IC 10	dIrr	−15.7	160	30.0	5	−1.37	old + 1 burst	yes
IC 1613	Irr	−14.7	63.6	20.9	11	−1.3	active + bursts?	yes
EGB 0427+63	dIrr	−12.6	9.12	33.0	2	−1.5	active	yes
And I	dSph	−11.9	4.71	9.9	13.4	−1.5	old	no
And II	dSph	−11.1	2.35	9.9	17.2	−1.6	old	no
LGS 3	dIrr	−10.5	1.33	12.7	14.5	−1.8	old ?	yes
And III	dSph	−10.3	1.13	9.9	4.5	−2.0	old + 2 bursts	yes
And IV	dSph	*	*	*	*	*	old	no
And V	dSph	*	*	*	*	*	old	yes
And VI	dSph	*	*	*	*	*	old + bursts?	yes

* The properties of And IV–VI are not well determined, but can be assumed to be similar to those of And I–III (Mateo 1998).

of the total mass of luminous and dark material associated with each one, independent of the stellar luminosity. Figure 7.1 shows the cumulative number of subhalos predicted by the semi-analytic model, as a function of peak circular velocity[1] , compared with the average number of satellites within the virial radii of the Milky Way or Andromeda, as a function of observed circular velocity (or a comparable quantity, $\sqrt{2}$ times the velocity dispersion, for the non-rotating systems). On the face of it, the theoretical prediction is terrible; at the low circular velocities it overestimates the number of dwarf satellites by more than an order of magnitude.

This result was highlighted recently in the numerical work of Klypin *et al.* (1999b) and Moore *et al.* (1999), but it is not new; Kauffmann & White (1993) found the same excess of small objects in their semi-analytic modelling of the Local Group, and similar results appeared in other galaxy formation models based on Press-Schechter predictions (White & Rees 1978; Dekel & Silk 1986; Cole 1991; White & Frenk 1991; Lacey & Silk 1991). The recent numerical results, however, conclusively refuted one hypothesis which had been advanced to explain away this excess, namely that small objects were disrupted or combined together by merging (e.g. White & Rees 1978; Kauffmann & White 1993). In fact, all CDM halos have central cusps which may be able to resist disruption, and small CDM halos are denser than large ones on average, so mergers with larger halos are unlikely to destroy them completely. Given this trend in density with mass, the large number of low-mass subhalos found in semi-analytic models is a straightforward consequence of the slope of the Press-Schechter mass spectrum, which in turn is determined by the slope of the CDM power spectrum. As a result, a basic feature of any hierarchical model based on pure CDM is that dark matter halos are very lumpy, sufficiently so that each lump cannot plausibly be related to a surviving, luminous companion of a massive galaxy.

[1]The circular velocity of dark matter halos reaches a maximum value at a characteristic radius, equal to 1.25 scale radii for the Moore profile used here (see figure 2.1 and equation 6.3). I will refer to this velocity and this radius as the peak circular velocity $V_{\rm p}$ and peak radius $r_{\rm p}$ respectively.

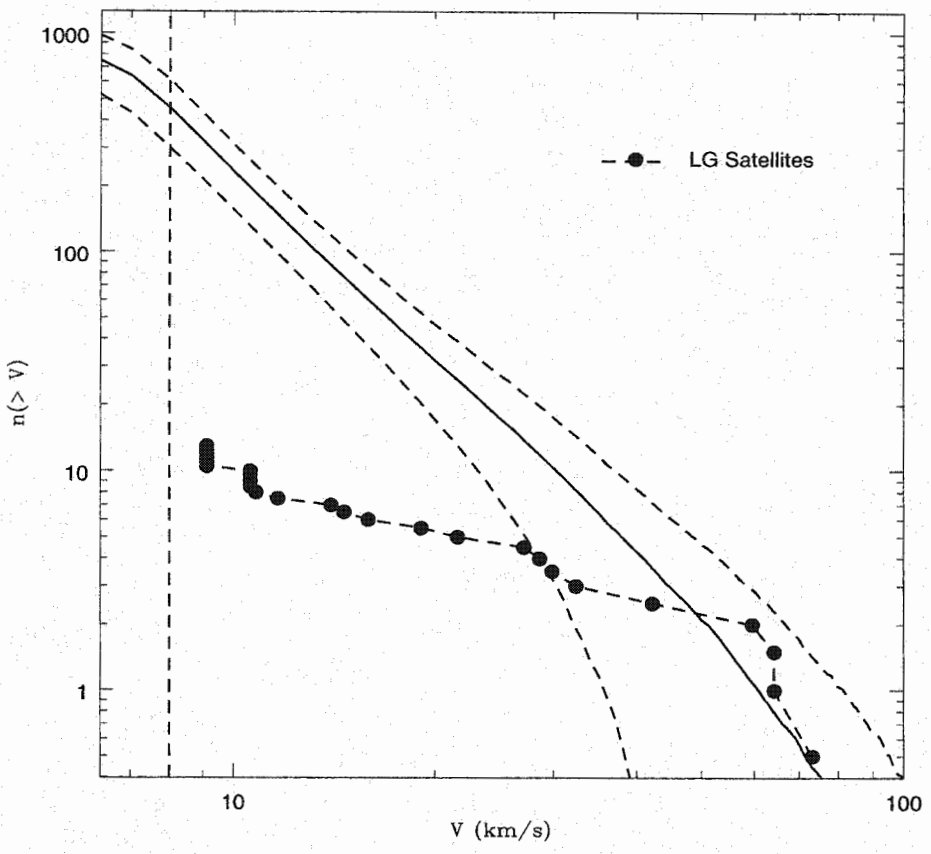

Figure 7.1 A comparison between the semi-analytic cumulative velocity function (solid line, with 1-σ deviation indicated), and the average number of satellites in the two main halos of the Local Group.

This discrepancy has been used to argue that pure CDM models may be wrong (e.g. Moore *et al.* 1999; Spergel & Steinhardt 2000; Cen 2001). A less radical possibility is that halos do contain large numbers of small subhalos, but that galaxy formation is less efficient on these small scales, so that few of these low-mass subhalos form stars (Klypin *et al.* 1999b; Moore *et al.* 1999; Bullock *et al.* 2000). In particular, the high-energy background radiation which has filled the universe since the reionisation epoch may have heated the gas in small halos to the point where it could not cool, condense and form stars (Babul & Rees 1992). In the rest of this chapter, I will develop a simple model for the net efficiency of star-formation in halos of different

masses, based on this premise, and determine the properties of dwarf satellites in this model.

7.2 The Luminosity Function of Galactic Satellites

7.2.1 Suppressing Star Formation in Dwarf Galaxies

While star formation in giant galaxies is observed to follow certain regular patterns, star formation in small galaxies is complicated by several effects. The potentials of small galaxies are shallow, such that the characteristic velocities associated with star formation are close to the escape velocity of the system. In particular, star formation leads to supernovae explosions within 10^7 years, and these can deliver enough mechanical and thermal energy to surrounding gas to eject it from the star forming region. (Even in massive galaxies, the energy released by supernovae is thought to regulate star formation by ionising, disrupting and heating the clouds of cold gas from which stars form. The self-regulation of star formation is normally called 'feedback'.) Overall, internal energy sources in dwarf galaxies are comparable to the gravitational energy of the system, leading to strongly regulated star formation.

The collapse and virialisation of dark matter halos gives their primordial gas a specific amount of thermal energy, heating it to a specific temperature. This 'virial' temperature is very low for small halos, around 10^4 K, as a result of their shallow potentials. External heat sources can produce enough energy to heat the gas within a halo to this temperature, so environmental effects will also play a role in regulating star formation in low-mass halos (Babul & Rees 1992; Efstathiou 1992; Gnedin 1996; Forcada-Miró 1997).

An important example of this type of regulation is heating due to reionisation. Baryons in the early universe first combined into electrically neutral atoms at around $z \simeq 1150$, the epoch at which the CMB photons were last scattered. At some later

time, however, the flux of high-energy background radiation from stars, or possibly from active galactic nuclei, was sufficient to re-ionise the neutral gas in the intergalactic medium. The redshift of this epoch of reionisation is estimated to lie between z $\simeq$ 6 (the redshift of the most distant observations of absorption-line systems, all of which show the universe to be ionised) and z $\simeq$ 10–20 (Haiman & Loeb 1997), based on models of radiative transfer in the early universe (Gnedin 1996, 2000) and studies of the CMB (Schmaltzing, Sommer-Larsen & Goetz 2001). Reionisation has important consequences for dwarf galaxy formation. Many studies have shown that the gas in the shallow potentials of low-mass halos is sufficiently heated by the background flux after reionisation that it cannot cool and condense to form stars (e.g. Babul & Rees 1992; Efstathiou 1992; Gnedin 1996, 1999; Forcada-Miró 1997; Barkana & Loeb 1999). Thus, after the epoch of reionisation $z_{\rm ri}$, star formation is suppressed in halos with circular velocities of less than $V_{\rm crit} \simeq 35\,{\rm km\,s^{-1}}$ (below $V \simeq 18\,{\rm km\,s^{-1}}$, the gas may actually be photo-evaporated from halos by background radiation).

It is straightforward to test the consequences of this effect with merger trees. I will consider the following simple model:

- prior to reionisation, cooling and star formation occur with a fixed efficiency $\epsilon_{\rm sf}$ in all halos, so that by $z_{\rm ri}$, a fraction $\epsilon_{\rm sf}\Omega_{\rm b}/\Omega$ of the mass of any halo is in cold gas or stars
- in large halos, star formation continues with the same efficiency after $z_{\rm ri}$, instantaneously transforming a fixed fraction of any newly accreted mass into cold gas or stars
- in halos with velocities less than $V_{\rm crit}$, no star formation occurs after $z_{\rm ri}$, although the stellar mass of the halo may grow through mergers

Given that the circular velocity of a halo can increase through mergers or accretion, this model also implies that intermediate-mass objects may stop forming stars at $z_{\rm ri}$, but start to form stars again at later times, as their velocity rises above $V_{\rm crit}$. Figure 7.2 illustrates schematically the change in the total and stellar masses (parameterised

by the velocities, V and $f_* \times V$ respectively) for massive, intermediate and low-mass halos. Note that the original gas content of small halos is assumed to remain available for star formation, once they grow or merge into systems over the velocity threshold. While reionisation can photo-evaporate the gas in very small halos, removing it from the halo completely, this gas is assumed to be reincorporated into the system as it grows larger. This explains the jump in stellar content for the intermediate-mass case shown in figure 7.2, once it reaches $V_{\rm crit}$. The model also neglects the timescale for gas cooling and star formation, and does not distinguish between cold gas and stars, counting them as a single component. The actual rate of star formation in galaxies will be discussed further in chapter 8.

Overall, this model of truncated star formation drastically reduces the number of galaxies (that is halos containing stars) with circular velocities less than $V_{\rm crit}$, and implies that low-mass halos that did form stars did so at very early times. Above this limit, a fixed fraction of the mass of a halo is in the form of stars and cool gas. The resulting truncation of the luminosity function produces rough agreement with the numbers of dwarf galaxies observed in the Local Group, as has recently been demonstrated by Bullock *et al.* (2000) (see also earlier analytic or semi-analytic work by Efstathiou (1992); Chiba & Nath (1994), Kepner, Babul & Spergel (1997), Nagashima, Gouda, & Sugiura (1999), and Barkana & Loeb (1999)). I will proceed to investigate this agreement in more detail below. Reionisation is also an appealing model for the suppression of star formation, in that of the three free parameters it introduces ($\epsilon_{\rm sf}$, $V_{\rm crit}$, and $z_{\rm ri}$), two ($\epsilon_{\rm sf}$ and $V_{\rm crit}$) are fairly well determined from other considerations. It is worth noting, however, that other processes such as feedback may also play an important role in limiting star formation in dwarf galaxies. In particular, the lowest-mass objects, which often formed before reionisation and were therefore unaffected by it, may have lost their gas through these other mechanisms. The predicted effect of feedback is more model-dependent, however, so I will restrict myself to a simple model where reionisation is the only process limiting early star

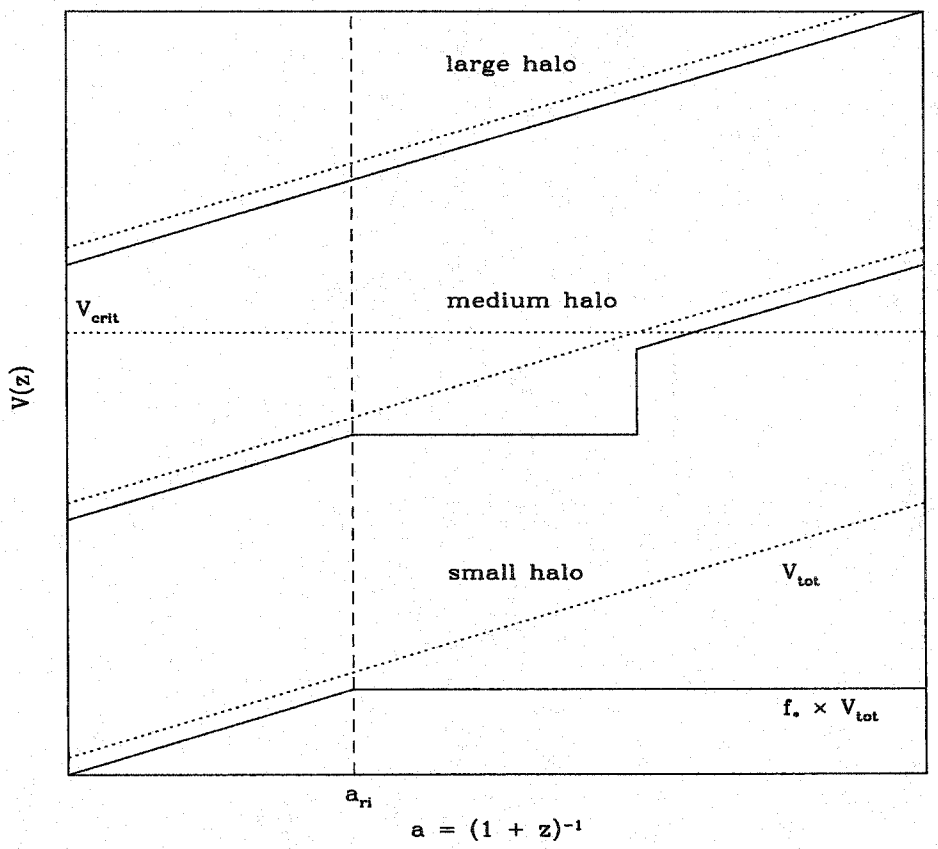

Figure 7.2 A model for the truncation of star formation in small halos, due to reionisation. The diagonal dotted lines indicate the circular velocities of three different halos as they change with scale factor, while the solid lines indicate schematically the evolution of the stellar fraction in each system.

formation.

7.2.2 Implementation and Results

In the model of regulated star formation described above, massive halos wind up with a fixed fraction ϵ_{sf} of their baryonic mass in cold gas and stars, while low-mass halos have some fraction between ϵ_{sf} and 0 in this form, depending on their mass at reionisation. The overall efficiency of cooling and star formation, ϵ_{sf}, was chosen so as to produce a stellar mass of approximately $8 \times 10^{10}\,M_{\odot}$ in the main halo, in keeping

with estimates of the stellar and gaseous mass of the Milky Way. Requiring that:

$$1.6 \times 10^{12} \epsilon_{\rm sf} \frac{\Omega_{\rm b}}{\Omega} = 8 \times 10^{10}, \tag{7.1}$$

gave $\epsilon_{\rm sf} = 0.9$ for the SCDM models and $\epsilon_{\rm sf} = 0.5$ for the LCDM models. These efficiencies are in the range suggested by numerical studies of star formation (Katz 1992; Navarro & Steinmetz 2000). The critical velocity used was $V_{\rm crit} = 35\,{\rm km\,s^{-1}}$, the value expected from theoretical studies (Ikeuchi 1986; Rees 1986; Babul & Rees 1992). The reionisation epoch was left as a free parameter, subject to the limits mentioned above ($6 \leq z \leq 20$). Figure 7.3 shows the ratio of the stellar to the total mass for the subhalos merging with the main system in a typical SCDM run, with $z_{\rm ri} = 15$. Most systems containing stars have a fixed fraction of their mass, $\epsilon_{\rm sf}\Omega_{\rm b}/\Omega$, in this form, and lie along a diagonal line in figure 7.3. Only the minority of systems which accreted some mass after $z_{\rm ri}$, but had velocities below $V_{\rm crit}$ when they merged with the main halo, deviate from this locus.

To relate this stellar mass to an observable luminosity, one requires a (stellar) mass-to-light ratio Γ. To get a rough idea of the resulting luminosity functions, I will initially use $\Gamma = 3$, a value typical of old stellar systems (Tantalo *et al.* 1996; Maraston 1999). Figure 7.4 shows the predicted cumulative luminosity function of satellites around a giant galaxy like the Milky Way at the present day. The solid curves show the results for reionisation epochs $z_{\rm ri} = 6$, 9, 12, and 15 (with the 1-σ variance indicated for the latter case). The points are the average for the satellites of the Milky Way and Andromeda, with the satellites of Andromeda shifted down by a factor of 0.75 in luminosity, to reflect the larger mass of this system. For both the predicted and observed values, I have only counted satellites within the virial radius of the main system, which makes the observed luminosity functions uncertain by a few counts. The counts are also likely to be incomplete below $M_V = -10$ (van den Bergh 2000); this is indicated schematically by the error bars, which show the effect of doubling the counts below this magnitude. Systematic trends in morphology with magnitude may

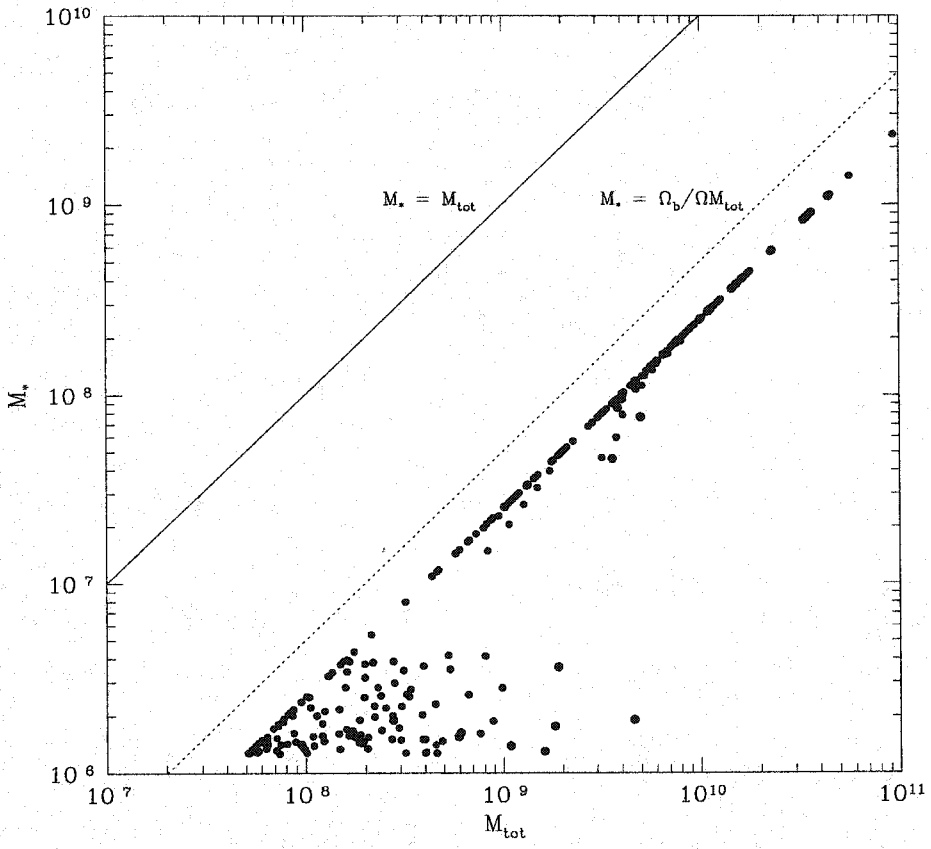

Figure 7.3 The mass in stars and cool gas (M_*), plotted vs the total mass $M_{\rm tot}$, for some of the subhalos in a typical run. The diagonal lines indicate the loci for $M_* = M_{\rm tot}$ (solid line) and $M_* = M_{\rm baryonic} = \Omega_{\rm b}/\Omega\, M_{\rm tot}$ (dotted line).

produce differences in the stellar mass-to-light ratio. I have used different point styles to distinguish dwarf irregulars (open squares), dwarf ellipticals (thick open circles) at the bright end from the more common and fainter dwarf spheroidals (filled circles).

Clearly, for early reionisation epochs, the predicted luminosity functions are now in much better agreement with the observations. This is a direct consequence of the truncation of star formation at low circular velocities, which produces a flat cumulative distribution below the threshold velocity $V_{\rm crit}$, matching to the relatively flat cumulative luminosity function observed in the Local Group. This general result

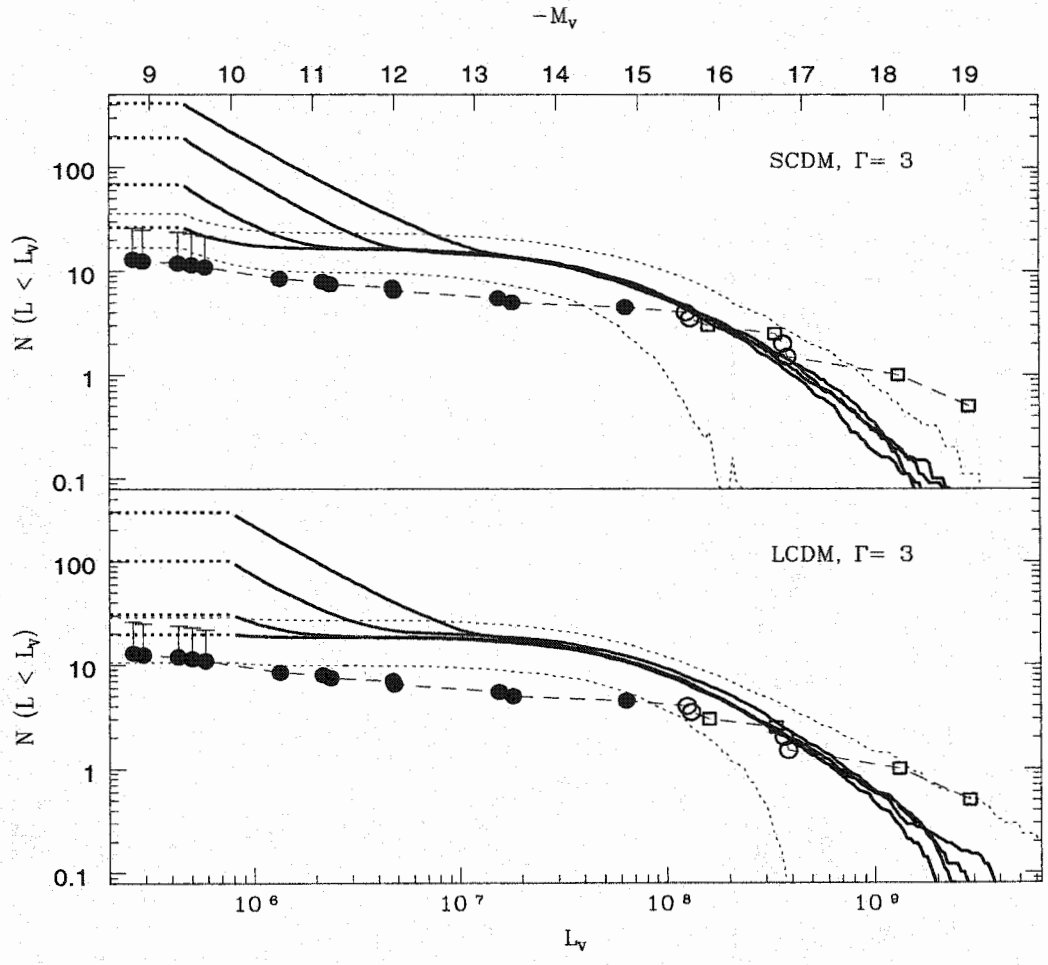

Figure 7.4 The average cumulative luminosity functions for dwarf satellites in the SA model, assuming a constant mass-to-light ratio of $\Gamma = 3$. The four curves are the results for $z_{\rm ri} = 6$, 9, 12, and 15 from top to bottom at the faint end. The points are the average luminosity functions within the virial radii of the Milky Way and Andromeda. The symbol types indicate morphology: open squares – dwarf irregulars and spirals, open circles – dwarf ellipticals, filled circles – dwarf spheroidals. The horizontal line and error bars indicate possible incompleteness below $M_V = -10$. The thin dotted lines indicate the 1-σ variance in the SA results for $z_{\rm ri} = 15$.

was established previously by Bullock *et al.* (2000), using a slightly different implementation of the same physical model. It is important to note that the horizontal normalisation of the luminosity function should be considered uncertain by a factor of as much as 2, due to uncertainties in the total mass of the Milky Way's halo (which determines the mass of the merger tree), the total mass of the stellar components of the Galaxy (which I used to determine $\epsilon_{\rm sf}$), and by the general cosmological uncertainty in $\Omega_{\rm b}$. Independent of this shift, there are two discrepancies in the shape of the luminosity function that are worth commenting on.

First, the semi-analytic luminosity functions have a steep slope at the low-luminosity end, which is inconsistent with observations for $z_{\rm ri} \leq 9$. This discrepancy stems from the fact that very small systems assembled most of their mass before $z_{\rm ri}$, and are therefore unaffected by reionisation in this simple picture. In the model of Bullock *et al.* (2000), this effect is not apparent because the substructure in side-branches of the merger tree is not considered. This produces a flatter spectrum of subhalo masses, as shown in figure 6.6. Simulations indicate that the actual spectrum of subhalos is steeper than this, however, so that this extra substructure should be taken into account.

The steep faint-end slope of the luminosity function in figure 7.4 may be real; for $z_{\rm ri} \geq 12$ it could correspond to faint, undetected dwarf galaxies, high-velocity clouds with few stars, or possibly globular clusters. The latter objects are small, dense, numerous and have lower luminosities than dwarf galaxies, but are also much more compact, suggesting that they may have formed in very different conditions. As discussed previously, feedback is also expected to reduce the efficiency of star formation in small objects, compounding the effects of reionisation. This could reduce the numbers of objects at the faint end of the luminosity function to the level required to match observations, so speculation as to the nature of these objects may be premature. Taken at face value, however, the model suggests that the reionisation epoch must be ≥ 12, as opposed to the value of 8 found by Bullock *et al.* (2000). This is

in agreement with a recent estimates of z $\simeq$ 10–20, by Haiman & Loeb (1997) and Schmaltzing *et al.* (2001);

There is also a smaller discrepancy at the bright end of the luminosity function; the brightest Local Group satellites are a few times brighter than expected from the simple model with a constant Γ. Here the explanation is straightforward; these objects have star formation histories and morphologies which differ systematically from those of fainter dwarfs, and should be brighter due to differences in their stellar mass-to-light ratio. The predicted mass-to-light ratios for stellar systems can be calculated using population synthesis models, but depend on several intrinsic physical parameters, including the age, metallicity and star formation history of the population, as well as several model parameters that should in principle be fixed by observations, including the shape of the initial mass function, the fraction of non-luminous remnants (brown dwarfs of black holes) in a population, and various corrections to the computed model atmospheres (Carraro *et al.* 2001; Bruzual 2000). For old, passively evolving systems, models predict a mass-to-light ratio of $\simeq$ 3 in the V band (e.g. Tantalo *et al.* 1996), although this value is probably uncertain by a factor of 50–100%. For active star-formation regions, on the other hand, Γ should be 0.5–0.7 or less (the value 0.5 was computed using the Starburst99 models (Leitherer *et al.* 1999), for instance, for a 1 Gyr old population with a metallicity of [Fe/H] = -1.5). The local stellar mass-to-light ratio of the galactic disk lies between these extremes, at a value of $\Gamma \simeq 2$ (Binney & Tremaine 1987).

Figure 7.5 shows the luminosity function for a reionisation epoch $z_{\rm ri} = 15$, plotted assuming $\Gamma = 3$ and $\Gamma = 0.5$. We see that these two distributions bracket the observed luminosity functions, with the fainter objects matching the higher value of Γ while the brighter objects match the lower value. Given that the brighter objects are irregulars actively forming stars in many cases (open squares), while most of the faint objects are dwarf spheroidals with no evidence of recent star formation, it seems entirely possible that the two groups are drawn from the original mass function predicted by

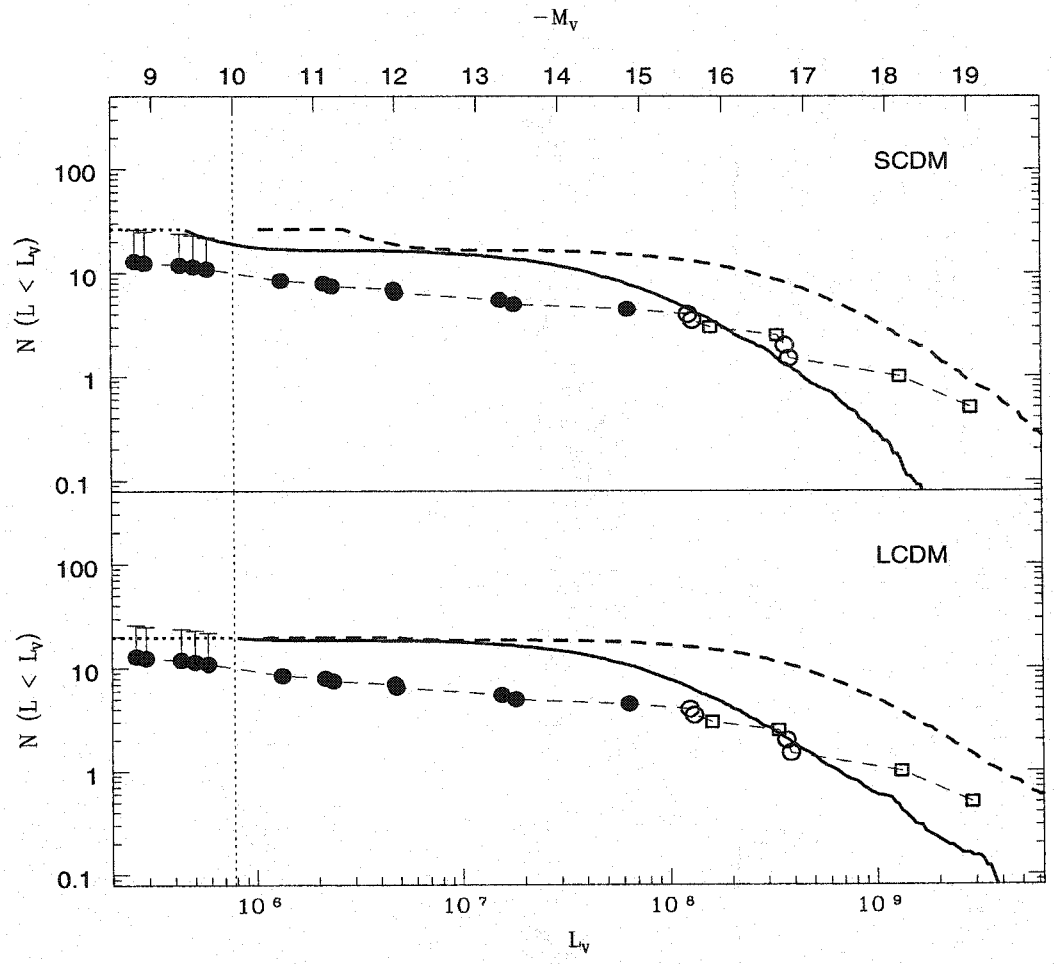

Figure 7.5 As figure 7.4, but for $z_{\rm ri} = 15$, and two different mass-to-light ratios, $\Gamma = 3$ (solid line) and $\Gamma = 0.5$ (thick dashed line).

the semi-analytic model. I shall re-examine this possibility in more detail below, and discuss the evolutionary status of the bright early-type galaxies, which also appear to deviate from the high–Γ curve.

7.2.3 Other Properties

Before considering the evolutionary history of the SA dwarf galaxies in detail, it is worth seeing whether they resemble the local dwarf galaxies in their structure and spatial distribution. The stellar masses determined in the previous section are much

smaller than the total masses of individual subhalos, so we expect the structure and internal dynamics of the stellar distributions to be dominated by dark matter, as is observed in the less massive Local Group dwarfs (Mateo 1998). In particular, the characteristic velocity of each stellar system should be determined by its surrounding subhalo, while its spatial extent should lie within the tidal limits of the subhalo. Figure 7.6 shows the peak velocities (bottom panels) and peak radii (top panels) for dark matter halos in the SA model, plotted versus their corresponding stellar luminosity, assuming a fixed mass-to-light ratio of 3. The large points indicate the limiting radii and peak circular velocities $V_{\rm p}$ (or $\sqrt{2}\sigma$ for the non-rotating systems) observed for Local Group dwarfs with well-determined distances (symbol types are as in figure 7.4). If we assume that the stars within each subhalo extend out to roughly the peak radius of the halo, and have a circular velocity roughly equal to $V_{\rm p}$, we see that the model results agree almost perfectly with the observations. (This correspondence also implies that the total mass-to-light ratios estimated for the dwarfs in the SA model would match observed values.) One interesting consequence of this picture is that the dark halos around many dwarf galaxies should extend out much further than the limits of their stellar component. Some dwarfs may have lost a large fraction of this dark mass through tidal stripping, explaining the observed (stellar) tidal streams around dwarf galaxies. I will discuss this point later, in section 7.4.3.

One can also compare the spatial distribution of the SA dwarfs to the observed distribution of Milky Way satellites (I will not consider the satellites of Andromeda as their distances from the centre of that system are much less well determined). Figure 7.7 shows this comparison, as a function of luminosity, for two typical runs. Overall, it is hard to compare the observed and predicted distributions, due to the small numbers of objects involved, and the uncertainties in the observed distances, but it is certainly possible to find systems which closely resemble the satellites of the Milky Way.

In summary, by accounting for the effects of reionisation on star formation in

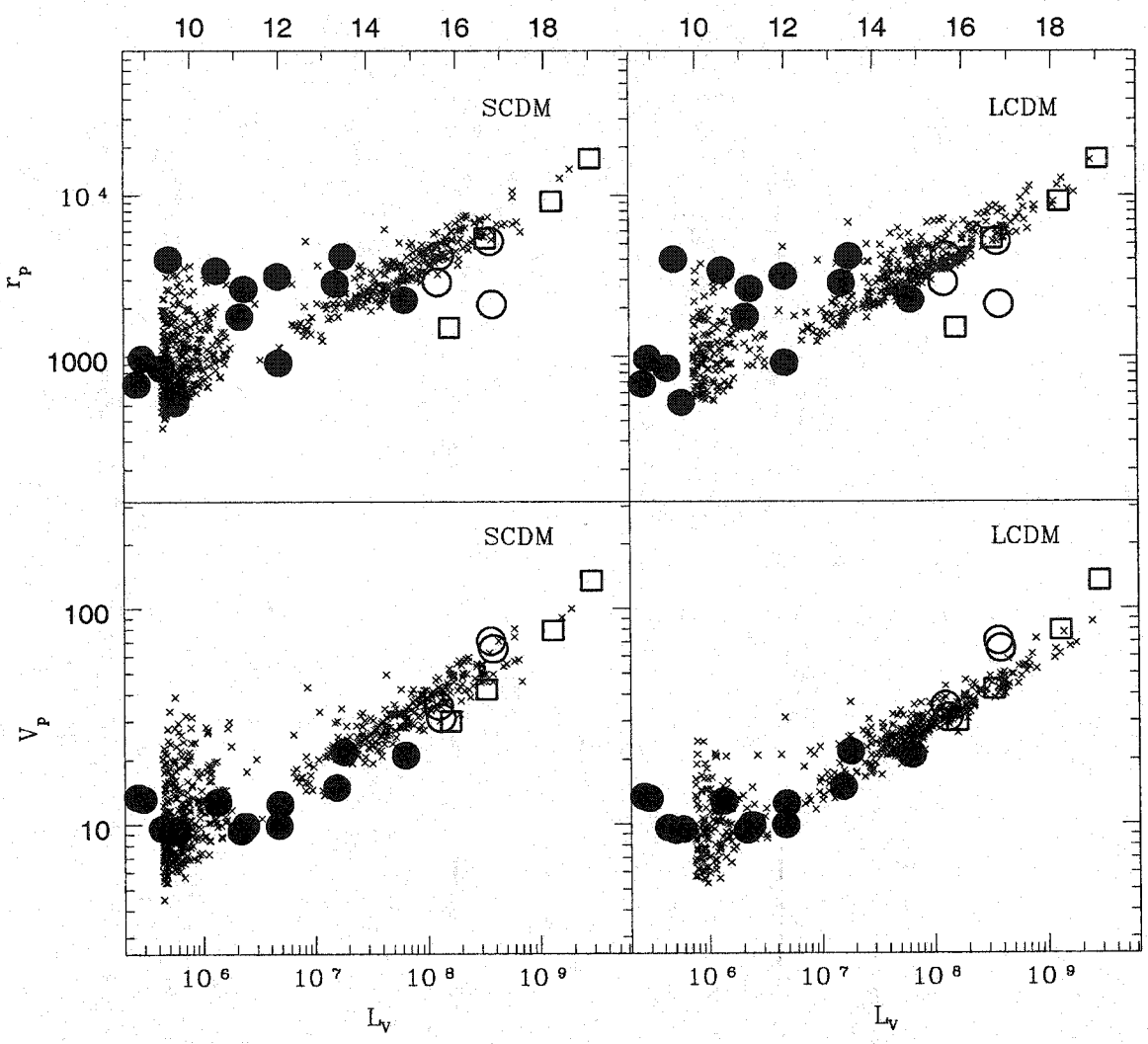

Figure 7.6 The peak-velocity radii and peak velocities for model dwarf galaxies (small points), compared with Local Group dwarfs with well-determined distances (large points – symbol types are as in figure 7.4).

small halos, it is possible to generate reasonable numbers of dwarf galaxies, that is groups of stars within satellite subhalos, in the semi-analytic model. These objects appear to have luminosities, peak velocities or velocity dispersions, sizes and a spatial distribution consistent with the observed properties of local dwarfs. The model can

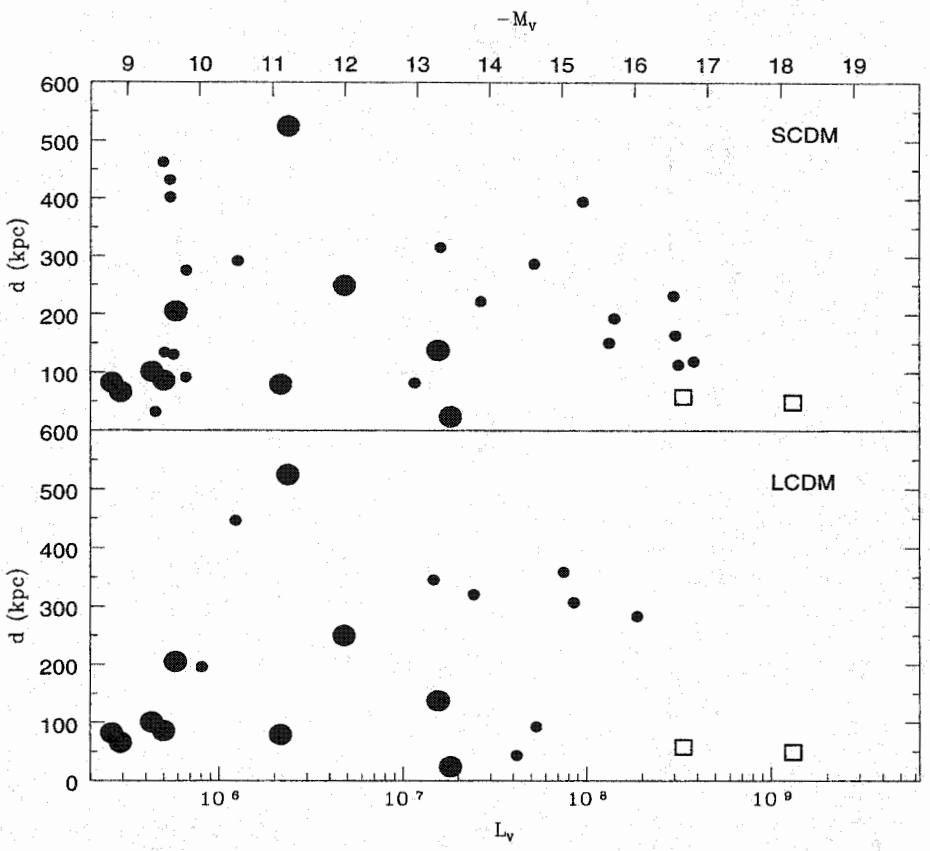

Figure 7.7 The radial distribution of satellites in two typical SA runs (small points), compared with the distribution of the Milky Way's satellites (large points). Symbol types are as in figure 7.4.

also provide more detailed information about their evolutionary histories, which I will consider in the next section.

7.3 The Star-formation History of Local Satellites

While the semi-analytic model, combined with heating after reionisation, gives total stellar masses for galactic satellites which are approximately correct, a wealth of other information is also available observationally, about the morphologies and star-formation histories of local dwarfs. In this section I will consider some of the

other processes that may affect star-formation in dwarf galaxies, in order to permit a closer comparison with observed systems.

Until they fall into the main halo, the star formation histories of the side-branches of the merger tree are dominated by the effects of reionisation. As explained in the previous section, in massive halos, star formation is assumed to be continuous, while in low-mass systems it will cease, possibly indefinitely, at $z_{\rm ri}$. Since reionisation occurs very early in the history of the universe, most low-mass systems will only contain old stellar populations, whereas massive systems will contain a mixture of stellar ages. Subsequent evolution, once systems fall into a larger halo, will generally emphasise these trends. Low-mass systems may lose their hot gas through ram-pressure stripping (Lin & Faber 1983; Kormendy 1985; Mayer *et al.* 2001), while massive systems will retain their gas and continue to form stars. This picture is broadly consistent with the observed properties of galactic satellites, as seen in tables 7.1–7.2; low-mass systems tend to be old, passively evolving dwarf spheroidals, while more massive systems have more gas, more active star formation and a greater range of morphologies.

There are some exceptions to this general trend. Some massive satellites resemble giant ellipticals in their morphology and concentration (e.g. M 32 and NGC 205), suggesting they may be smaller versions of field ellipticals which experienced a disruptive encounter either before or after falling into the main halo (see chapter 8 for further discussion of the formation of field ellipticals). Some low-mass satellites are also undergoing bursts of star formation at the present day (e.g. LGS 3), which change their colour and brightness, and both massive and low mass satellites show evidence for bursts of star formation in the past (cf. tables 7.1–7.2).

These fairly ubiquitous bursts of star formation must be due to some stochastic process occurring within the main halo, since many are more recent than the infall times for typical satellites. One possibility is that collisions with other satellites, or with the disk of the main galaxy, trigger these episodes. This is plausible, both because even small galaxies still have gas associated with them in some cases, e.g.

Sculptor and Sextans and half the dSph companions of Andromeda (see tables 7.1–7.2), and also because collisions between subhalos are relatively common.

While there is no clearcut way to distinguish collisions from less disruptive encounters, a satellite of mass m is sure to be strongly perturbed if its peak radius $r_{\rm p}$ overlaps with the peak radius of another system of similar mass $m' = fm$, or if it passes within $r_{\rm p}$ of the disk. Figure 7.8 shows the distribution of the number of such collisions, and the time of the last collision, for $f = 0.2$. We see that $\simeq$ 20–25% of all systems have experienced a collision or an encounter with the disk, and a number of these have occurred within the past 5–6 Gyr (since t = 7–8 Gyr) (although very recent encounters are rarer in the SCDM halos). Thus it is possible that these objects would have experienced starbursts and have younger, brighter stellar populations. Some objects have also experienced many encounters; examination of the individual runs show that these are generally systems in sub-groups, which fall into the halo at a single merger epoch, and remain on similar orbits thereafter.

Overall, it therefore seems quite easy to explain the presence of multiple episodes of star formation in even the smallest dwarfs, provided they retain at least some of their original gas (or accrete new gas through Bondi accretion – see Silk, Wyse & Shields 1987). Violent encounters may also play a role in determining the morphology of massive dwarfs. These objects retain large quantities of gas, and encounters between them, or with the disk, may result in strong starbursts, disruption of any disk present, and the formation of a relaxed remnant that is more elliptical in appearance. In this way, encounters may transform dwarf irregulars like the LMC into dwarf ellipticals like M 32. The same process (termed 'galaxy harassment'), operating on larger scales, has been proposed as a way of converting cluster spirals in to S0s or ellipticals, for instance (Moore, Katz & Lake 1996).

This leads to the following, synthetic picture of dwarf galaxy evolution:

- halos originally form stars independently, with an overall efficiency determined by reionisation and possibly also feedback; they may or may not retain their

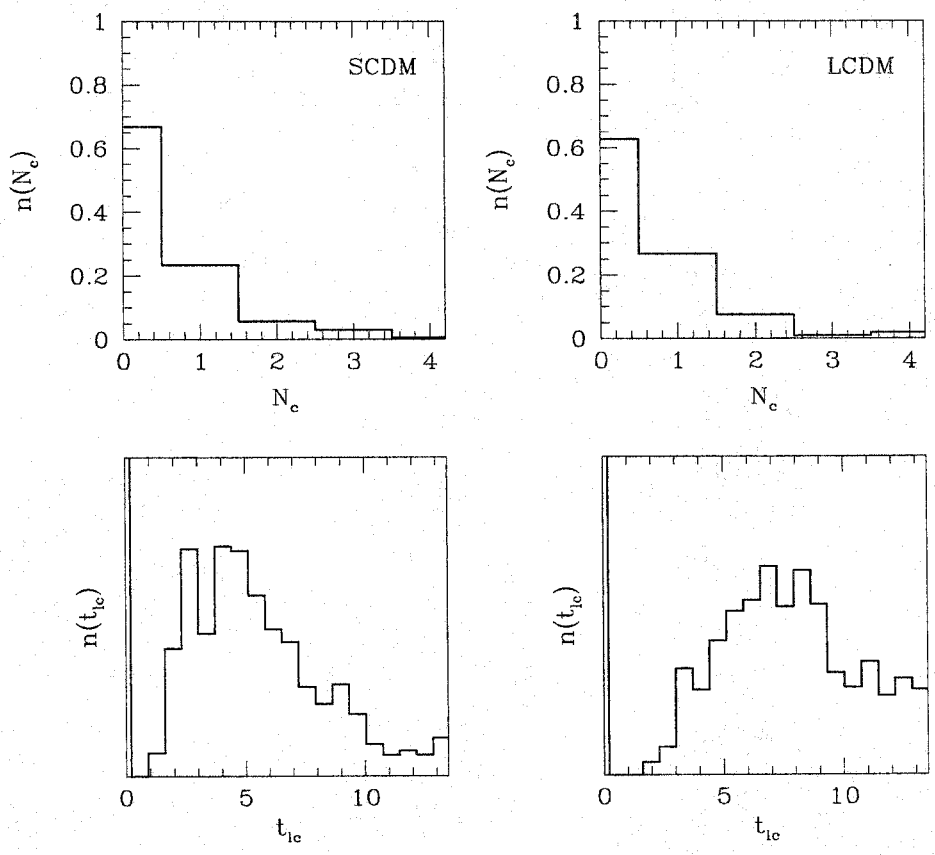

Figure 7.8 The number of close encounters, as defined in the text (top panels), and the time of the last close encounter (bottom panels), for dwarf satellites in the SCDM and LCDM halos.

primordial gas, depending on their mass

- most small subhalos form their stars prior to reionisation; they may also lose some or all of their gas then or after they fall into the main halo, through ram-pressure stripping; most of their stars are old (e.g. Ursa Minor)
- high mass subhalos continue to form stars after reionisation, and after they fall into the main halo, as dwarf irregulars (e.g. Magellanic clouds) or small spirals (e.g. M 33)
- encounters trigger occasional starbursts in low-mass systems (e.g. LGS 3)
- encounters may transform high-mass irregulars into dwarf ellipticals (e.g. M 32)

While this model does explain some of the general trends observed in the Local Group, it is probably not complete, and requires further testing, as well as detailed observations of other groups, to confirm it. Assigning morphologies and luminosities on the basis of this model does produce very convincing luminosity functions, however; figure 7.9 shows several composite luminosity functions, where the SA dwarfs have been assigned a morphology on the basis of having had a collision or encounter within the past 3 Gyr. The mass-to-light ratios used are 3 for passive, low-mass systems (dwarf spheroidals), 2 for disturbed high-mass systems (dwarf ellipticals), and 0.5 for star forming systems (dwarf irregulars). The top panel shows the average luminosity function for the satellites within the virial radii of Andromeda and the Milky way, for comparison.

There are several remaining uncertainties in this picture of dwarf galaxy evolution. Local Group is a very small sample of objects, which individually show evidence for many complex phenomena, leaving a number of open questions. Is the gas loss imputed for low-mass dwarfs due to environmental effects within the halo, such as ram-pressure stripping, or does it predate their infall into the halo? Can low-mass satellites retain primordial gas within the halo? What is an appropriate criterion for determining encounter rates between halos? Overall, these questions need to be addressed by more detailed modelling of the gas dynamics of dwarf satellites, which might also shed light on the related question of the nature of high-velocity clouds. Another aspect of dwarf galaxy evolution that is more amenable to semi-analytic modelling is the question of tidal disruption rates and the formation of the stellar halo of the Milky Way. I will turn to these questions in the final section of this chapter.

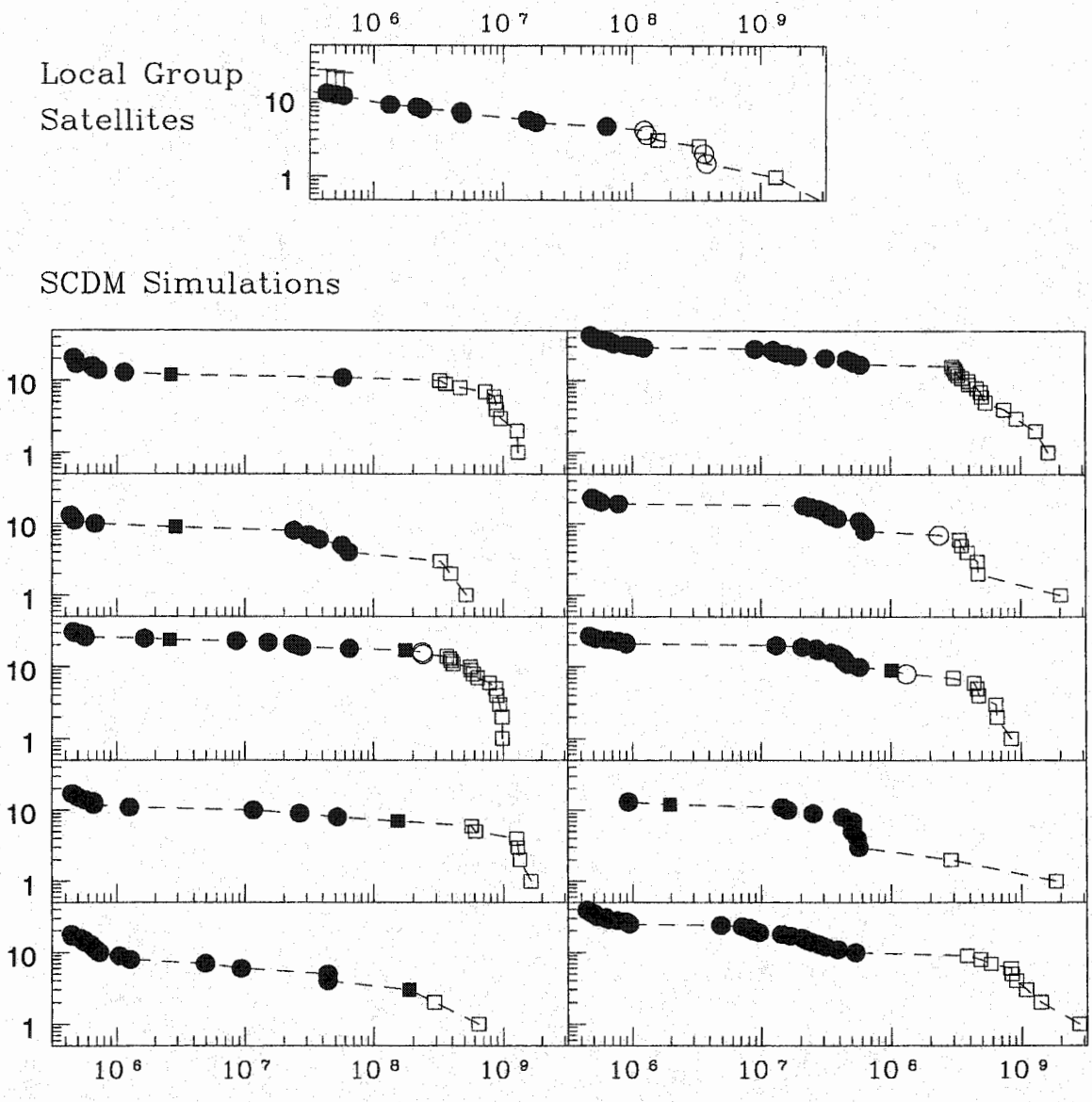

Figure 7.9 Sample luminosity functions for dwarf satellites, compared with the average luminosity functions of the satellites of the Milky Way and Andromeda (top panel). Mass-to-light ratios and morphologies have been assigned as described in the text, and are indicated by the different symbols (symbol types as in figure 7.4). Filled squares indicate disturbed, low-mass systems.

7.4 Tidal Disruption and the Formation of the Galactic Stellar Halo

7.4.1 The Density Profile of the Stellar Halo

Tidal stripping and satellite disruption were discussed extensively in the first part of the thesis. In chapters 3–4, I considered the evolution of a satellite with a

simple, single-component density profile. In this section, I will be more concerned with the disruption of a dwarf galaxy, that is a group of stars contained within a much larger halo of dark matter. Since the dark matter will dominate the dynamics of all but the most massive satellite galaxies, however, this should not alter the problem substantially. I also showed in chapter 4 that the analytic model of satellite evolution is only valid while the satellite retains at least 10–20% of its initial mass; beyond this point, disruption occurs fairly quickly, but the precise details of the process become too complicated to model. In general, I will consider a satellite to be disrupted when it is truncated to within the peak radius of its rotation curve. This disruption criterion was shown to produce the same number of surviving subhalos as seen in numerical simulations of halo formation, so it is probably reasonable, but I will also discuss the effect of using more or less restrictive criteria below.

In the SA model, once a satellite is disrupted, its stellar mass is added to the spheroidal component of the main galaxy, which includes the bulge and the stellar halo. If an extremely massive satellite collides with the main galaxy, it will disrupt the disk. This process may add to the mass of the stellar halo (e.g. Bekki 1998d), but I will assume that it contributes primarily to the bulge, while the stellar halo builds up mainly from satellite debris. I will examine disk disruption and bulge formation in more detail in chapter 8, and restrict my attention to the satellite disruption in this section.

There are several basic observational tests for a model of the formation of the stellar halo; its mass is a few times $10^9\,\mathrm{M}_\odot$ (Gilmore, King, & van der Kruit 1989), its density profile goes as $r^{-3.5}$ (Gilmore *et al.* 1989), and it is predominantly old (Unavane, Wyse & Gilmore 1996). The semi-analytic model produces larger masses of disrupted material – $M_{\rm sh} \simeq 1.5\times10^{10}\,\mathrm{M}_\odot$ for the SCDM trees, or $M_{\rm sh} \simeq 2.5\times10^{10}\,\mathrm{M}_\odot$ for the LCDM trees – but some of this material will wind up in the bulge. These masses are comparable to the total mass of stellar material in surviving satellites, so using a different disruption criterion will change them substantially. The criterion

used here is required to match the numbers of surviving subhalos seen in numerical simulations however, as shown in chapter 6. Changing the efficiency of star formation in satellite halos would also affect these mass estimates, at the expense of requiring different mass-to-light ratios to match the luminosity function of Local Group satellites.

The mean density profile produced by satellite debris will depend on its subsequent orbital evolution. While the SA model gives detailed information about the position, velocity and mass of stripped material as it leaves the satellite, subsequent orbital evolution is complicated by the fact that the size and shape of the background potential change with time. Furthermore, the need to resolve individual debris streams from hundreds of satellites with large numbers of particles puts the problem beyond the means of the analytic model described in part I. Pending a numerical treatment of this problem, one can get a rough idea of the radial distribution of debris by plotting the density of stripped or disrupted material as a function of the last point in the satellite's orbit before it was disrupted (although mass loss is triggered by a passage through pericentre, it lasts for a large fraction of the orbital period, so the orbital phase at which satellites are disrupted should be fairly random – cf. chapter 4.)

Figure 7.10 shows the average density profile of satellite debris, versus the radius at which the satellite was disrupted, for SCDM and LCDM (points). Overall the debris is centrally concentrated, and generally comparable with the observed density profile of the stellar halo in the Milky Way; the dotted line, for instance, shows a power-law of slope -3.5, corresponding to the distribution of halo material observed locally (Gilmore *et al.* 1989). Thus, while the formation of the stellar halo requires further numerical study, the SA results are broadly consistent with observations of the halo of the Milky Way.

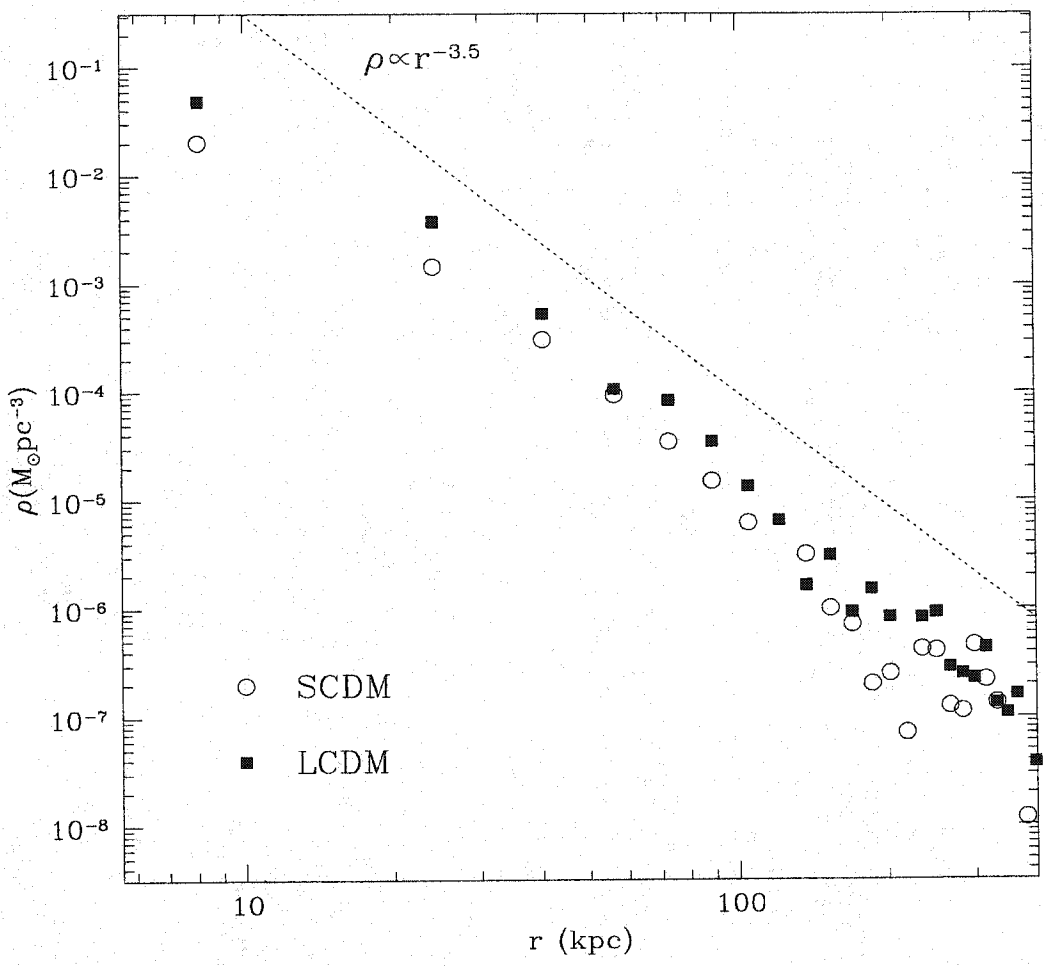

Figure 7.10 The approximate density profile of satellite debris, for several SCDM and LCDM trees (open circles and solid squares respectively). The dotted line shows a slope of -3.5, similar to that observed in the Milky Way (Gilmore *et al.* 1989).

7.4.2 The Age of the Stellar Halo

Another interesting test of the model is in the overall age of the stars left in the stellar halo. Unavane *et al.* (1996) pointed out that there are strong observational constraints on the amount of intermediate-age material present in the stellar halo, which in turn limits how much material could have been accreted from the Local Group satellites, as they appear today. They divided the halo into three metallicity

ranges, $-1 >$ [Fe/H] > -1.55, constituting 40% of the halo, $-1.55 >$ [Fe/H] > -1.95, constituting 30% of the halo, and [Fe/H] < -1.95, constituting the remaining 30%. Based on the observed numbers and metallicities of blue stars in the halo, they then estimated that no more than a fraction of the stars in each of these three metallicity ranges could be substantially ($\simeq 2$ Gyr) younger than the dominant, old population. The upper limits were 28%, 6%, and 3%, for the three ranges respectively.

Since I have not included a full population synthesis code in the SA model, it is impossible to determine the metallicities of the different stellar populations in the dwarf satellites directly (such a calculation would also be very sensitive to the efficiency and dynamics of star formation in these small objects, and thus highly uncertain). Observed Local Group satellites show a strong correlation between metallicity and luminosity or circular velocity, however, as shown in figure 7.11. The solid symbols show the dwarf spheroidals and transitional objects, for which the correlation is strongest. It is weaker for the irregulars and dwarf ellipticals (open symbols), but these objects may have lower and variable mass-to-light ratios, which move them off the relationship.

The strong correlation for objects with similar mass-to-light ratios, implies a corresponding correlation between stellar mass and metallicity. This may be because the same process limiting star formation in small halos also removes metals from these halos more effectively, or because star formation only occurs in a single early burst in the smallest systems. More recent bursts of star formation should make systems brighter by one to two magnitudes or more, but they do not appear to affect their metallicity substantially, as seen in figure 7.11. If one uses the stellar mass of the system as an indicator of its metallicity, it is thus possible to determine the approximate metallicity distribution of material in the stellar halo. The limits mentioned above then become limits on the amount of young material contributed to the stellar halo by massive, intermediate-mass and low-mass systems respectively.

Star formation in low-mass objects (those with velocities less than $V_{\rm crit}$) is turned

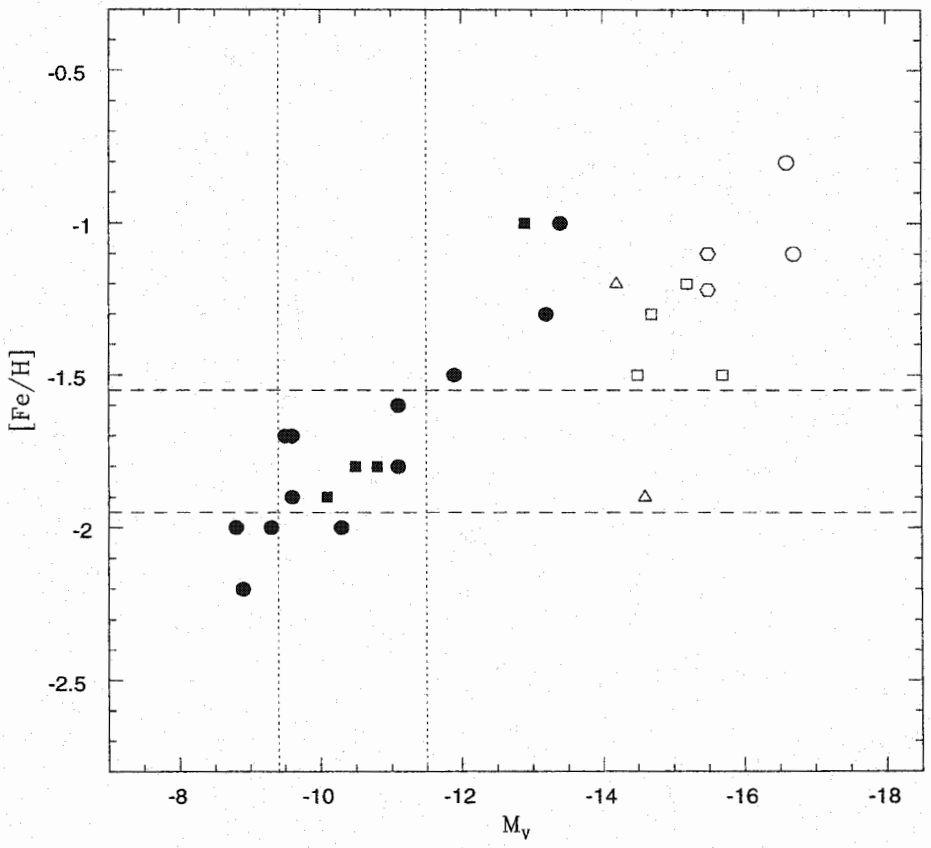

Figure 7.11 The metallicity of Local Group dwarf galaxies, versus their luminosity. A strong correlation exists for dwarf spheroidals or dwarf spheroidal/irregulars (solid circles and squares). Dwarf ellipticals, ellipticals, and irregulars lie off this relationship, (open symbols) possibly because they are systematically brighter for a given stellar mass. The horizontal lines divide the objects into three metallicity classes; the vertical lines show the corresponding approximate division into three magnitude classes, for the solid symbols.

off at the onset of reionisation, which corresponds to a very early time, so all their stars are old. To estimate the fraction of younger stars in more massive satellites, I assume that massive satellites have formed stars at a constant rate up until their disruption time. I also define the dominant old population as consisting of stars which formed before $t_{\rm old}$. The fraction of younger stars contributed to the halo by a satellite

disrupted at time $t_{\rm dis}$ is then simply:

$$f_{\rm y} = \frac{t_{\rm dis} - t_{\rm old}}{t_{\rm dis}} \ . \tag{7.2}$$

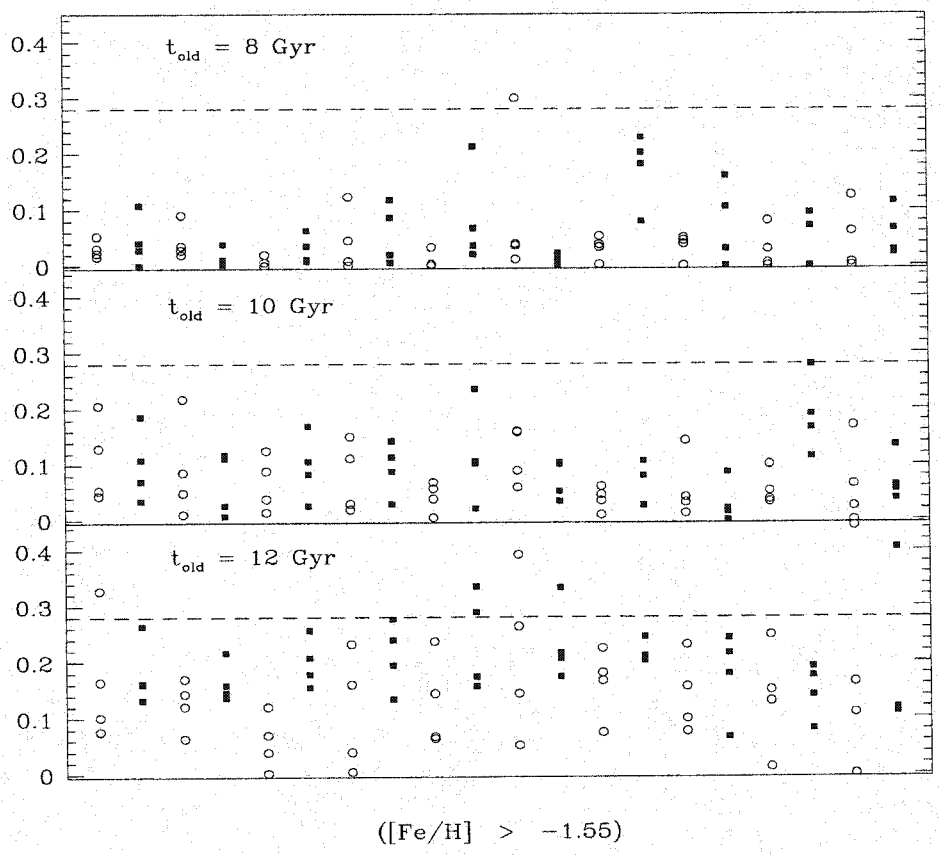

Figure 7.12 The fraction of intermediate-age, metal-rich material in the stellar halo, for a set of SCDM halos (open circles) and LCDM halos (solid circles). The three panels show the results assuming a different minimum age for the old population. The horizontal lines indicate upper limits for the stellar halo of the Milky Way, from Unavane *et al.* (1996).

Figure 7.12 shows the fraction of intermediate-age material in the richest metallicity bin, for a set of SCDM runs (open circles) and a set of LCDM runs (solid circles). Since the age of the universe is only 13.5 Gyr in the cosmologies considered here, it is not clear exactly what value of $t_{\rm old}$ to use; the three panels show results for $t_{\rm old}$

= 8, 10 and 12 Gyr The horizontal lines indicate the limit determined by Unavane *et al.* (1996). We see that the SA halos are compatible with these limits, provided $t_{\rm old} \leq 10$ Gyr. Overall, the stellar halos formed in the SA model seem similar to the observed stellar halo of the Milky Way. (In the two other metallicity bins, not shown here, almost all material is old.) Aside from the problem of choosing an appropriate value for $t_{\rm old}$ in the younger cosmologies considered here, there are several complications that may affect this result, however. First, I have ignored the contribution from material formed in bursts; this could increase the young fraction slightly, particularly in the last two bins. Furthermore, I have assumed that the disk, when disrupted, contributes to the bulge rather than the stellar halo. If young disk material were mixed in with the stellar halo material by disruption or violent heating, this could also raise the fraction of intermediate-age material in the stellar halo to an unacceptable level. I will show in the next chapter that most disk disruptions occur at early times, however, so that any material added to the stellar halo by the process would also appear to be old. Overall, the massive contributions to the stellar halo and bulge, both from disrupted disks and from disrupted satellites, tend to occur at early times in the SCDM and LCDM trees, so the age of the stellar halo is easily explained.

7.4.3 Tidal Streams in the Halo

When a satellite loses mass through tidal stripping or disruption, the stripped particles or stars remain on similar orbits with slightly different energies. A corresponding difference in orbital velocity stretches the stripped material out in long, coherent tidal streams, which remain thin and strongly associated for several orbital periods. Given that an orbital period in the outer stellar halo is several Gyr, it is plausible from theoretical arguments alone that some of these streams would be visible at the present day. There has been much recent interest in the existence of tidal

streams, prompted by the discovery of Sagittarius, a dwarf galaxy currently being disrupted by our Galaxy (Ibata, Gilmore and Irwin 1994), and several possible fainter streams (Helmi *et al.* 1999; Majewski *et al.* 2000). Thus it is is worth considering whether tidal streams are common in the SA results.

Unfortunately, if the stars in dwarf galaxies occupy the centres of much larger dark matter potential, then they will only escape to form visible streams when the dwarfs are close to being disrupted, and as I argued earlier, the exact timing of this process is hard to determine in the analytic model of satellite dynamics. One can get a sense of the frequency of tidal streams, however, by examining how many systems exist close to the disruption limit, or how many have been disrupted recently. Satellites are considered to be disrupted when they have been truncated to the radius of their peak circular velocity, when they disrupt the disk, or when they pass within the halo core radius (3.5 kpc at the present day) of centre of the potential. The latter criterion is put in to avoid numerical errors in the orbital calculations; in practice, it only accounts for 10% of the disrupted satellites, and is only important at early times when the main halo is small. Most satellites are disrupted by progressive mass loss. Figure 7.13 shows the distribution of the ratio of the truncation radius to the peak velocity radius, $x = r_{\rm t}/r_{\rm p}$, for all the satellites in a set of SA runs, including those that have been disrupted. For 'live' systems, I have used the values of $r_{\rm t}$ and $r_{\rm p}$ at $z = 0$, while for disrupted systems, I have used the values at the time the satellite was disrupted. Since the model does not follow the dynamics of disrupted systems, the details of the distribution around the disruption limit $r_{\rm t}/r_{\rm p} = 1$ are unreliable, but we see in general, disrupted systems have been stripped to 1–2 times their peak radius. Some of these systems may show tidal tails while they are in the early stages of disruption, while in the latter stages they may form tidal streams.

Finally, it is also possible to examine when satellites were disrupted, since the tidal debris of recently disrupted systems may remain visible for several orbits. Figure 7.14 shows the distribution of times since satellites were disrupted, in the SCDM runs. We

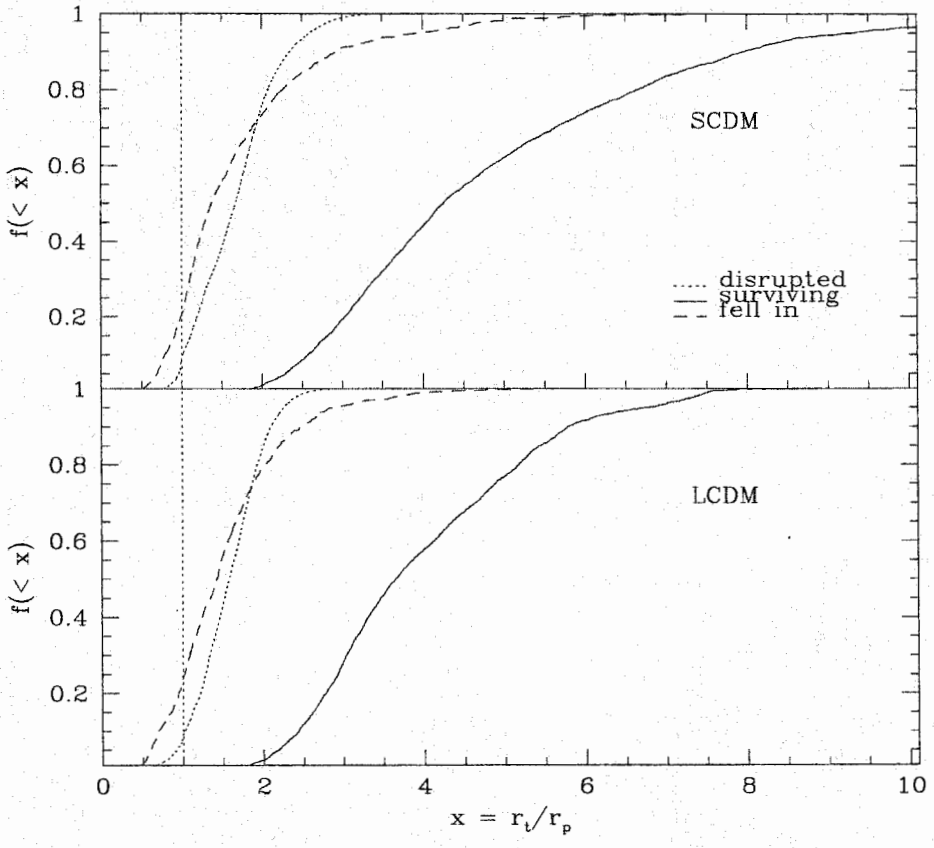

Figure 7.13 Cumulative distributions of the tidal radius of satellites, relative to their peak velocity radius, in the SCDM and LCDM runs. The numbers include systems which have been disrupted (e.g. all those with $r_t/r_p < 1$).

see that 10–20% of systems have been disrupted within the past 4–6 billion years (2–3 dynamical times), so coherent tidal streams could be quite common in this model. This plot also shows that the rare satellites which fall in or hit the disk (dashed lines) do so at earlier times. Halo disruption and the subsequent orbital evolution of stellar material is sufficiently complex that this process should be re-investigated numerically; these preliminary results provide a motivation for numerical studies with realistic, two-component model dwarfs, falling into a halo at a cosmological rate.

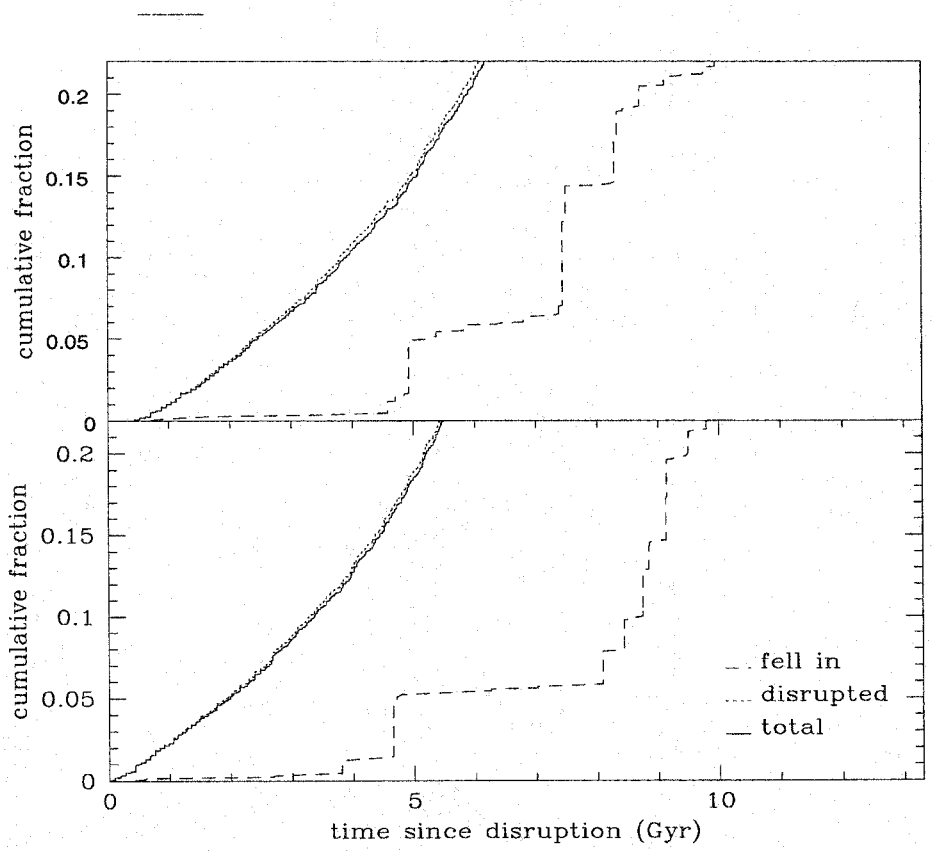

Figure 7.14 The cumulative distribution of times since disruption, for satellites which have been disrupted of have fallen in by $z = 0$ (solid lines). The dotted line shows the distribution for those disrupted, while the dashed line shows the distribution for those which have fallen in or hit the disk.

7.5 Summary

In this chapter, I have applied the basic framework of the semi-analytic model to the study of the stellar contents of subhalos, within the halo of a giant galaxy like the Milky Way. I have argued that observations of local satellite galaxies require either a radical revision of CDM, or a less radical restriction of star formation in small halos, due to feedback, external heating at the epoch of reionisation, or similar processes.

The effects of reionisation can be represented using a simple model with three main parameters, the epoch of reionisation $z_{\rm ri}$, the velocity threshold $V_{\rm crit}$ below which it

truncates star formation, and the efficiency of star formation $\epsilon_{\rm sf}$, above this threshold. Setting these parameters to reasonable values produces a distribution of stellar masses similar to that of the Local Group satellites. Assuming that the dark halo of each object dominates its dynamics, it also possible to assign sizes and peak velocities to these model dwarf galaxies.

I show that this approach reproduces not only the luminosity function of local satellites, but also their sizes, circular velocities and spatial distribution, with one or two notable discrepancies. There is an excess of faint objects, which may reflect the need for additional feedback to further limit star formation in the smallest halos, but could also indicate the existence of fainter objects in the Local Group which have escaped detection. The most luminous local satellites are also somewhat brighter than predicted in the simplest model, but this is expected from differences in their stellar populations and stellar mass-to-light ratio. In particular, the SA model predicts that the stellar mass function should resemble a Schechter function, whereas the observed luminosity function is flatter than this. Assigning lower mass-to-light ratios to massive, star-forming satellites corrects this discrepancy.

Considering the evolutionary histories of satellite in more detail, I show that many have experienced encounters with the disk or with other satellites. This may account for the starbursts known to have occurred in the past in some small Local Group dwarfs. In more massive systems, it may lead to a transformation from dwarf irregulars into dwarf ellipticals. Assigning morphologies and mass-to-light ratios to dwarf galaxies based on their recent evolutionary history, it is possible to construct morphologically mixed luminosity functions which resemble those observed in the Local Group.

Finally, satellite disruption is fairly common in this model, despite the presence of dark matter halos surrounding dwarf galaxies. Although the detailed structure of this debris requires further investigation, the mass, age and spatial distribution of material stripped from satellites are roughly consistent with the properties of the spheroidal

components of the Milky Way. The existence of many systems close to the disruption limit, and of many systems which have recently been disrupted, also suggests that tidal streams should be common in this model, consistent with observations. Overall, the SA model succeeds in producing a detailed and plausible description of the evolution of dwarf galaxies and stellar halos around giant galaxies, consistent with observations in the Local Group. It remains to determine whether the evolution of the central components of giant galaxies, the disk and the bulge, is also consistent with observations. This will be the subject of the final chapter of this thesis.

Chapter 8

The Survival of Galactic Disks

In this chapter I apply the semi-analytic model to the problem of disk survival in CDM cosmologies. After a brief discussion of previous work on this problem, I consider two processes which may affect disks, *collisions*, that is direct encounters with subhalos, and *indirect heating*, which arises from distant interactions with halo substructure. While the collision rate depends on the precise definition used to classify encounters, I show that for reasonable definitions, disks are likely to have experienced their last major collision long ago, so that the existence of thin, old disks does not pose a problem for CDM models. If the stellar contents of the disk are converted into a spheroidal component when it is disrupted, however, then the bulge-to-disk mass ratios of bright galaxies may be too high in LCDM. This result places a constraint on the earliest time at which disks can have formed their stars, whereas the observed age of the disk of the Milky Way constrains the latest time at which it formed its contents. Small collisions can heat the disk substantially without disrupting it, and disks are also subject to indirect heating from the fluctuations in the galactic potential produced by orbiting subhalos. I present analytic estimates of the magnitude of direct and indirect heating, and show that they both predict considerable disk heating at early times, but less disk heating recently. These results are consistent with the observed velocity dispersion of the thin disk at the present day, and may also explain some of the increased velocity dispersion in older disk stars. In particular, the sharp jump in velocity dispersion that marks the transition to the thick disk corresponds to an age when minor mergers are common, suggesting that heating by halo substructure may be responsible for the formation of this component in disk galaxies.

While hierarchical galaxy formation, mediated by cold dark matter, is a well-established model for which there is much indirect support, it does rely on several key elements that have not yet been tested directly, including the form of initial

spectrum of density fluctuations on small scales, the precise physical properties of dark matter, and the global parameters describing the universe. It is therefore of particular concern when the predictions of this model appear to be in conflict with the observed properties of galaxies. An example of one such potential conflict was mentioned in the introduction; hierarchical galaxy formation requires frequent mergers between halos, which in turn might be expected to leave most galaxies with disturbed morphologies, give them a wide range of ages and dynamical states. Observationally, however, most isolated galaxies have disks, which constitute a large fraction of the stellar mass of the galaxy, and are relatively thin. These observations all suggest that the majority of field galaxies have not been disturbed by mergers in the recent past. Local observations of the Milky Way also support this conclusion: the disk of the Milky Way is about three times as massive as its bulge and corona, is dynamically cold, and has been that way for at least 8 Gyr, judging from observations of the velocity dispersion of old disk stars.

The effect of mergers on the long-term evolution of disk galaxies was first quantified in theoretical studies by Tremaine, Ostriker and others (Ostriker & Tremaine 1975; Tremaine 1976, 1981; see also Carlberg & Hartwick 1989; Ostriker 1990), and numerical studies by Quinn, Goodman and Hernquist (Quinn & Goodman 1986; Hernquist & Quinn 1988, 1989). This work faced two difficulties; that of determining accurate merger rates for a given cosmology, and that of determining the effect of individual mergers on the luminosity, thickness and dynamics of the disk. The strongest evidence for a conflict between cosmology and observed galactic structure was found by Tóth and Ostriker (1992, TO hereafter), who estimated the efficiency of disk heating by infalling satellites, and used the age, scale height and velocity dispersion of the thin disk of the Milky Way to constrain how much mass it could have accreted over in the recent past. Their estimate was that any merger with an object with more than 10% of the mass of the disk within the past 5 Gyr would have heated it substantially more than is observed. They concluded that SCDM might be ruled

out on this basis.

Uncertainties in the overall efficiency of heating derived in this analytic work, and the limitations of the earliest numerical simulation of a minor merger (Quinn & Goodman 1986) inspired a number of subsequent numerical studies. Quinn *et al.* (1993) and Walker, Mihos & Hernquist (1996, WMH hereafter) built directly on the results of the first study, extending it to a fully self-consistent, high-resolution simulation. In these papers, they established several important corrections to the estimate of TO, including the dependence of heating on orbital inclination, and the dissipation of the energy of the satellite in several sinks not accounted for in TO's calculation, including the orbital energy of stripped mass, and global, coherent modes in the disk. Huang & Carlberg (1997, HC hereafter) took a different approach, simulating quite different encounters between stiffer disks and initially distant, low-density satellites. They found both that their satellites were heavily stripped before they reached the disk, and that the disk absorbed some of the effect of the merger by tilting coherently towards the satellite, thereby reducing heating. Most recently, Velázquez and White (1999) carried out a large number of high-resolution simulations, which confirmed the importance of orbital inclination and mass loss in minor encounters.

The conclusion of these studies appears to be that although minor mergers do heat disks, the process may be 2–3 times less efficient than estimated by TO. It is hard to make more precise estimates, however, as they depend on the specific properties of the encounter. Some of the basic differences seen in these simulations can be understood using the analytic model developed in part I. Figure 8.1, for instance, shows the orbital evolution and mass-loss histories for three different satellites, VW's satellite model S1 (solid line; see chapter 2), a more concentrated satellite similar to that used by HC (dotted curve), and a dense satellite similar to the more concentrated of the two models considered by TO (dashed line). Although I have given these satellite models the same mass, and started them off on identical orbits (the orbits considered by VW – cf. chapter 2), their evolution is radically different, largely because of differences in

mean density. In particular, the least dense satellite (that of HC) loses most of its mass quickly, and experiences little orbital decay thereafter, while the much denser satellite of TO falls into the disk quickly, taking most of its mass with it. This different dynamical evolution may explain some of the disagreement between the amount of disk heating reported in these studies. Figure 8.2, for instance, shows the fraction of its original mass still bound to the satellite as it crosses the disk, as a function of the radius of the disk crossing, for many different orbits. The squares correspond to the TO satellite, the triangle to the VW satellite, and the circles to the satellite of HC; solid symbols indicate the average value in the region of the outer disk indicated by the dashed lines. If disk heating depends on the square of the bound mass of the satellite, as suggested by the analytic expressions mentioned below, the difference in mass loss in the three cases may explain the much stronger heating found by TO, relative to HC and VW (Taylor & Babul 2001).

These numerical simulations, while a substantial improvement on the earliest work, are still limited by the overall difficulty of the problem. Generating numerical models of galactic disks which remain stable over cosmological times is extremely challenging, and the detailed dynamics of disks may require millions of particles to resolve fully (Quinn & Goodman 1996; Velázquez & White 1999). Given that the disk must then be embedded in a much larger dark matter halo, this means it will be some time before minor mergers are routinely simulated in realistic halo-galaxy systems.

More generally though, it is worth pointing out that no set of merger simulations based on idealised initial conditions, no matter how accurate, can determine the overall likelihood of disk survival in a hierarchical universe. The question of disk survival is ultimately a statistical one; given estimates of the damage done by encounters, either those of TO, or revised versions based on subsequent numerical studies, how often does the average disk experience an encounter sufficiently intense to heat or disrupt it? Since mergers occur stochastically in hierarchical models, at

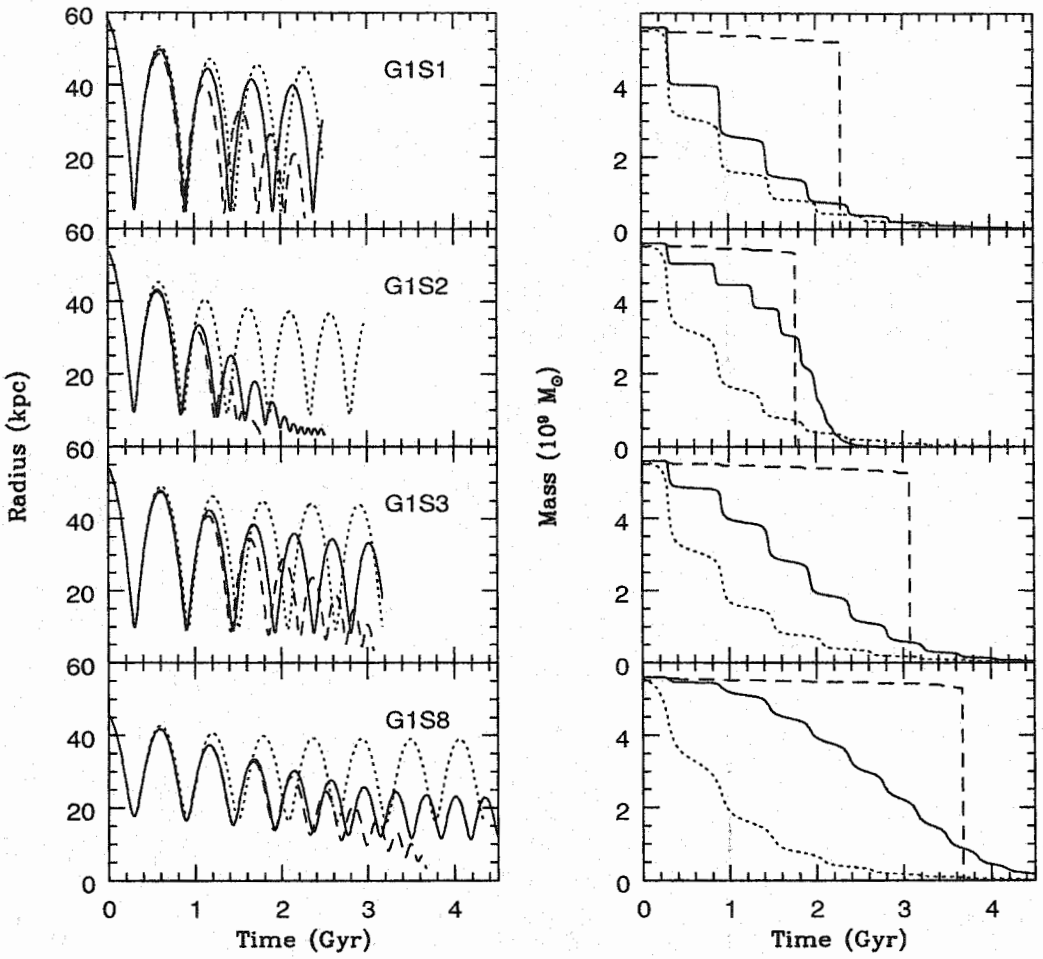

Figure 8.1 A comparison of the orbital evolution and mass-loss histories for satellites similar to those used by VW (solid lines), HC (dotted lines), and TO (dashed lines). Orbits and masses were calculated using the analytic model.

a mean rate that evolves over time, and since the previous history of merging systems may determine the outcome of an encounter between them, predicting the rate of disruptive encounters accurately is far from trivial. Any attempt to answer this question must necessarily consider an ensemble of many different systems, with representative merger histories. This was attempted in a cursory way by Kauffmann & White (1993), using a semi-analytic model, and by Navarro, Frenk & White (1994), using large-scale simulations. Both studies suggested that disruptive mergers are less

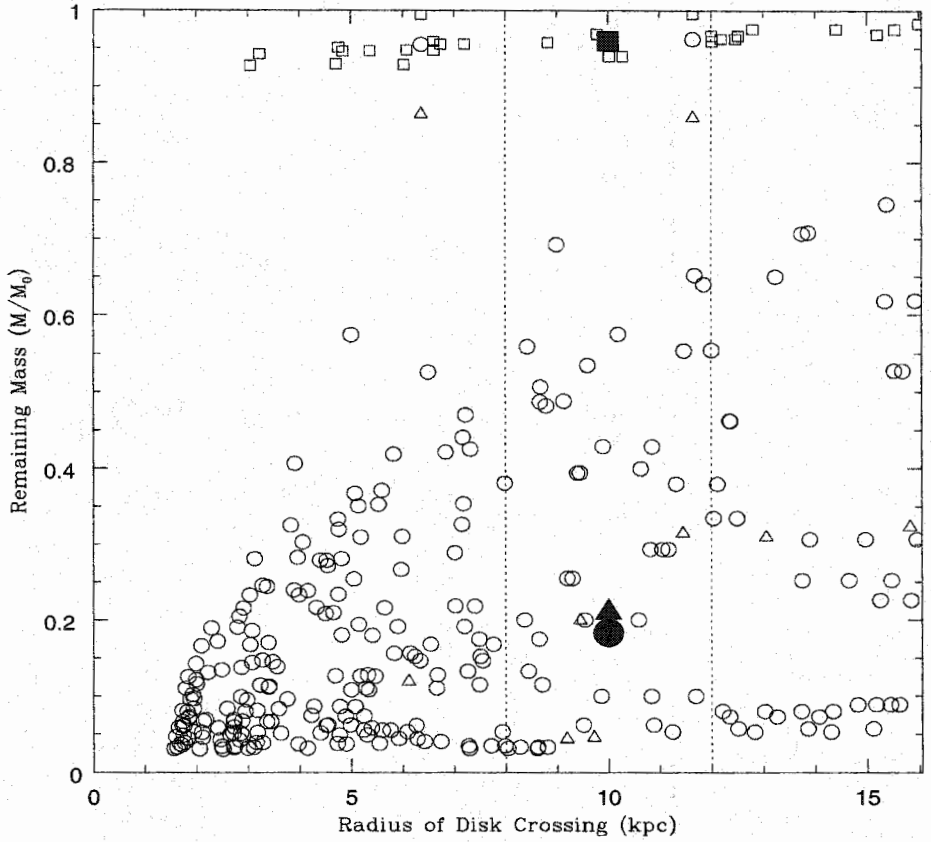

Figure 8.2 Disk-crossing events for the different satellite models used in figure 8.1, plotted in terms of radius and satellite mass fraction at the time of the event (see text for details). Symbol types are: square – TO, triangle – VW, circle – HC. The solid symbols indicate the average value for all disk crossings in the region of the outer disk indicated by the dashed lines.

common than expected, but neither studied the dynamics of the encounters in much detail. Numerical simulations are only just starting to form realistic disks *ab initio* (e.g. Navarro & Steinmetz 2000; Springel 2000), and simulating the merger histories of a large population of disk galaxies in their cosmological context, with sufficient resolution to improve substantially on previous work, is well beyond current computational capabilities. The methods developed in this thesis do allow a reinvestigation of the problem by semi-analytic means, however. In the rest of this chapter, I will use

my model to address two main issues, first the frequency of *collisions*, direct encounters between the disk and a satellite, and second the combined effect of many small or distant encounters, which produce a steady *heating* of the disk.

8.1 Estimating Collision Rates

8.1.1 Defining Major Mergers

I will start by considering the frequency of collisions, and more specifically the rate of major mergers, that is encounters which completely disrupt a galactic disk. If two massive systems encounter each other on a bound, parabolic or even mildly hyperbolic orbit, they will merge due to dynamical friction. It seems reasonable to suppose that if the systems are stellar disks of equal mass, embedded in larger halos of equal mass, and that the disks are inclined to one-another, the resulting merger remnant at late times will be much less flattened, due to conservation of angular momentum. This basic mechanism for turning two flattened disk galaxies into a single, more spherical elliptical, the 'merger hypothesis', was first suggested by Toomre & Toomre (1972), and has been demonstrated to work effectively in many subsequent numerical simulations (e.g. Negroponte & White 1983; Barnes 1988, 1992; Hernquist 1992; Mihos & Hernquist 1994; Steinmetz & Buchner 1995; Barnes & Hernquist 1996; Heyl, Hernquist & Spergel 1996; Bekki & Shioya 1997).

It is much less clear whether mergers between a primary galaxy and a less massive secondary will disrupt the disk of the primary, or simply heat it. While this problem has only been addressed recently in simulations, there are now several simulations of unequal-mass mergers (Bekki 1998a, 1998b, 1998c; Naab, Burkert & Hernquist 1999; Bendo & Barnes 2000; Cretton *et al.* 2001), which result in one outcome or the other, depending mainly on the mass ratio of the two systems. Thus it seems reasonable to distinguish between heating and disruption events based on a critical mass ratio,

as is commonly done in semi-analytic models (Cole 1991; Kauffmann & White 1993; Cole *et al.* 1994; Somerville & Kolatt 1999).

Before I estimate the critical mass ratios which distinguish between disruptive events, heating events, and less important collisions, it is worth noting a major source of confusion in determining appropriate values for these ratios from numerical simulations. The 'major merger' simulations mentioned in the preceding paragraph consider encounters between disk or disk/bulge galaxies, within dark matter halos 4–5 times as massive, starting at distances of $\simeq$ 100 kpc. Their general conclusion seems to be that ratios of 3:1 or less are required to disrupt the disk, while 10:1 encounters heat it. The 'minor merger' simulations of VW, HC, WMH, and others considered single-component satellites, introduced at some smaller distance from the centre of halos varying in greatly in size, from 35 kpc (WMH) or less to 84 kpc (VW). Their conclusion seems to be that a 5:1 encounter with the disk may be sufficient to thicken it substantially. The analytic calculation of TO, on the other hand, was expressed in terms of the mass the (single-component) satellite had as it crossed a given radius, on a decaying, quasi-circular orbit starting at 20 kpc. It concluded that a 10:1 encounter could thicken the disk. In each case, the mass ratio quoted is the ratio of the mass of the main system (the whole halo and galaxy in the case of major mergers, or the disk only in the case of minor mergers) to the mass of the satellite, *at its initial orbital position.*

In a simple model where the primary and secondary have exactly the same density profile, and the secondary falls in slowly enough to be tidally stripped to its Jacobi limit, a 3:1 merger at the virial radius will produce a 3:1 merger at 100 kpc, 60 kpc, 20 kpc or any other radius (albeit after a delay while the secondary falls in), provided that this mass ratio is taken to be the mass within a given radius in the primary, relative to the mass which remains bound to the satellite as it crosses that radius. Realistic satellite orbits are highly radial, however (Tormen 1997), and mass loss is correspondingly less efficient. Figure 8.3, for instance, shows the fraction of

its mass that each satellite in a SA halo has retained as a function of its merger epoch. It is apparent that most satellites lose a well-defined fraction, about 45%, of their mass on the first pericentric passage, and about 73% in the second passage (beyond a certain redshift the orbital phase no longer correlates with merger epoch, so there is more scatter in the results). Furthermore, small satellites will pass close to the disk several times before their orbits decay completely. Thus, the mass ratios quoted in different types of studies are not, in fact, directly comparable – a 3:1 halo merger between self-similar galactic systems may be quite different from an encounter between a single-component satellite and a disk three times more massive, sitting at the centre of a much larger dark halo.

A final complication to this extremely confusing situation is the fact that disks may re-form, even after a disruptive collision. If a disk is primarily gaseous, as it would have been at early times, then a violent encounter may simply heat this gas and eject it into the surrounding halo. Over time, this gas can cool and form a new thin disk, with a net angular momentum vector intermediate between those of its progenitors. Thus merger remnants may have thin disks at late times, despite the disruption of their original disks. There are some limits on the efficiency of this process, however; several studies suggest that about 75% of the gas involved in a major merger is immediately converted into stars, and therefore no longer available for subsequent disk formation (Barnes & Hernquist 1996; Mihos & Hernquist 1996).

In summary, there is no simple way to relate the results of major and minor merger simulations to cosmological rates of disk disruption. Disruption may require a 3:1 halo merger, an encounter between a galactic disk and a satellite with 50% of its mass or more, or something in between. Furthermore, it is not obvious how the rates of these different events would compare in a cosmological setting, given the complex orbital evolution of satellites, and its dependence on the properties of the initial orbit and the mass and density of the smaller system. Not every 3:1 halo merger will lead to a major merger with the disk, as the infalling halo may have substructure or may be

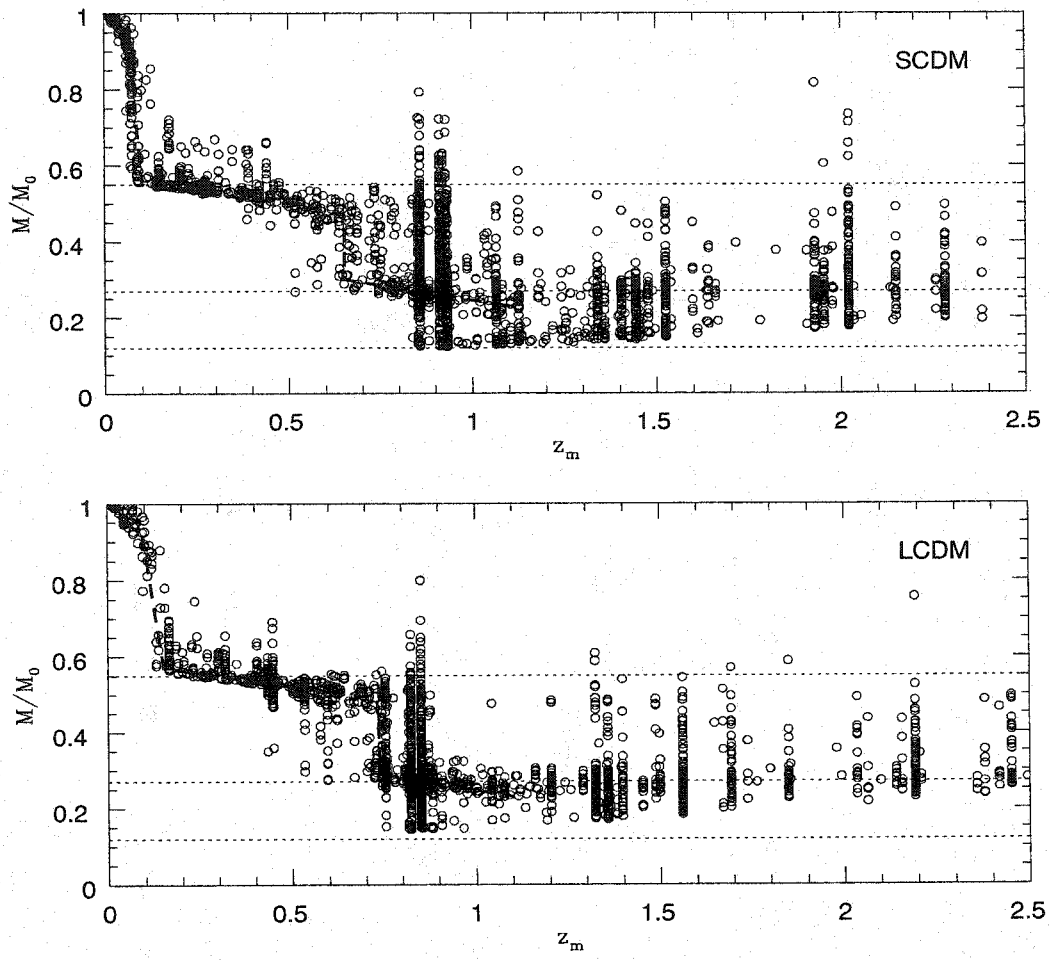

Figure 8.3 The fractional mass loss for subhalos in the SCDM and LCDM halos described in chapter 6, as a function of merger epoch. The thick dashed lines indicate the general pattern of mass loss as systems fall into the halo; most systems lose 45% of their mass on the first pericentric passage (highest dotted line), and a further 30% on the second passage (middle line). Systems below the lowest dotted line have been disrupted.

tidally stripped before it falls in, or the disk may grow substantially during the infall time, nor does every major merger with the disk result from a recent major merger at the virial radius. To make some headway in this problem, I will first examine the disruption rates predicted by the SA model, given different possible disruption

criteria.

8.1.2 Comparing Disruption Criteria

Figure 8.4 compares the disk ages determined using various different disruption criteria suggested by the results of merger simulations. Each panel shows the cumulative distribution of ages, that is the distribution of lookback times since the disk was last disrupted. The dotted horizontal line indicates a fraction of 50% of the trees (which were not selected in any way, and as a result do not necessarily have a large disk at the present epoch). A disk was considered disrupted when:

- a satellite came within 4 $R_{\rm d}$ of the centre of the potential ($\simeq$ 15 kpc in present-day systems)
- the mass ratio characterising encounter, r, exceeded the threshold $f_{\rm c}$

The figure shows results for:

a) $r \equiv M_{\rm t}/M_{\rm s,0}$ and $f_{\rm c} = 6$ (top panel, solid line)

b) as a), but for $f_{\rm c} = 12$ (top panel, dotted line)

c) $r \equiv M_{\rm g}/M_{\rm s}$ and $f_{\rm c} = 1$ (middle panel, solid line)

d) as c), for $f_{\rm c} = 2$ (middle panel, dotted line)

e) $r \equiv M_{\rm d}/M_{\rm s}$ and $f_{\rm c} = 1$ (bottom panel, solid line)

f) as e), for $f_{\rm c} = 2$ (bottom panel, dotted line)

where $M_{\rm t}$ is the total mass of the main system and $M_{\rm s,0}$ is the total mass of the satellite, each measured *when the satellite first crossed the virial radius of the main system*, whereas $M_{\rm g}$, the mass of the galaxy (disk+spheroid), $M_{\rm d}$, the mass of the disk, and $M_{\rm s}$, the mass of the satellite, are measured *at the time of the collision.* The first two criteria are closest to those used in simulations of major mergers, which

normally compare the initial masses of self-similar systems at large distances from each other, although the halos in these simulations are fairly small, relative to the central galaxies. The last four criteria are more suitable for comparison with minor merger simulations, which normally compare the satellite mass to the mass of the main galaxy.

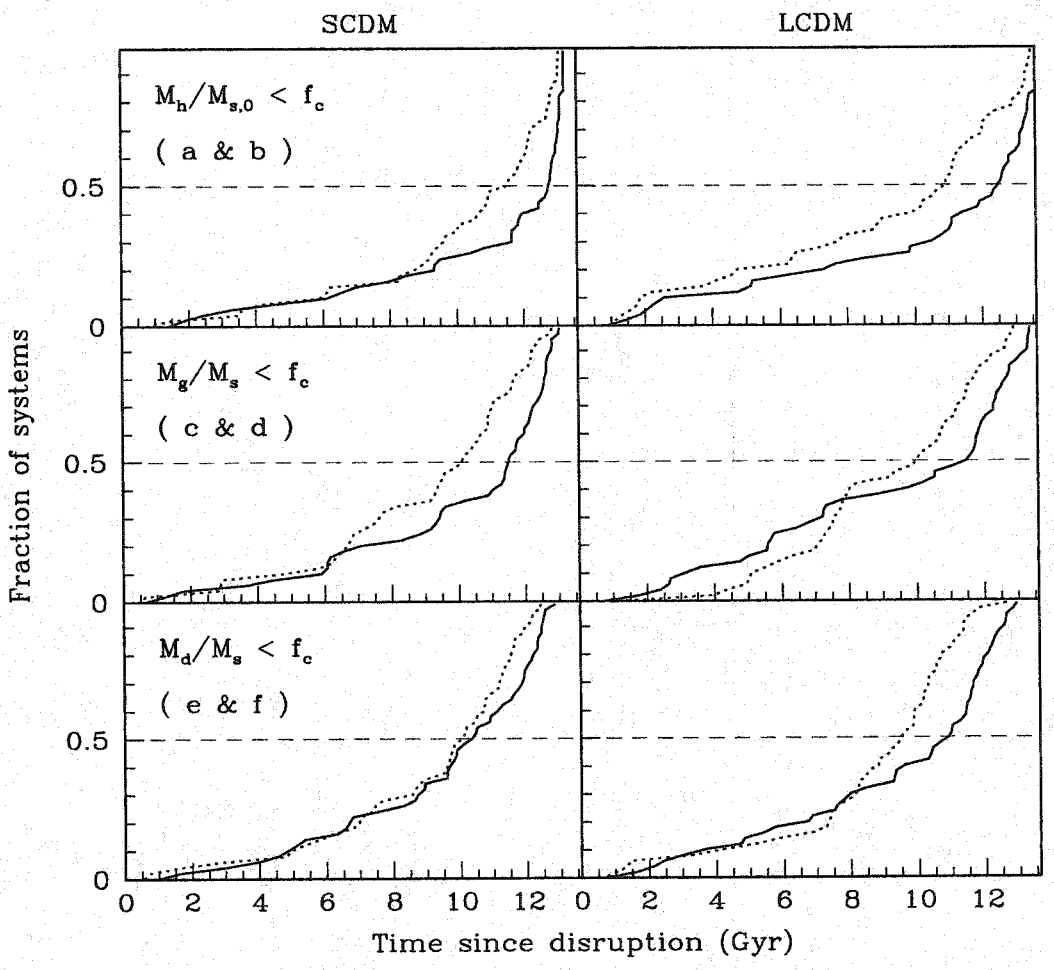

Figure 8.4 The cumulative distribution of present-day disk ages, that is the time since the disk was last disrupted, predicted by the SA model for SCDM and LCDM. The three sets of panels and two line styles indicate the ages determined using 6 different criteria defined in the text. The age of the universe is approximately 13.5 Gyr in either cosmology.

We can draw a number of conclusions from a comparison of these different curves. The first is that the disruption criteria used are decreasingly restrictive, going from top to bottom. Even for the least restrictive criterion, however, the average disk formed more than 9 Gyr ago (at $z \simeq 1.5$–2.5, depending on the cosmology). Thus the existence of an old disk in the Milky Way is not surprising. While it may be possible that disks are disrupted by even weaker encounters than assumed in case f), that is by encounters with an object with less than half the mass of the disk, this seems unlikely based on the simulation results. Conversely, even if disks are only disrupted by huge mergers, such as the 6:1 case in the top panel (which, for our fiducial galaxy model, would correspond to a subhalo which had five times the mass of the disk, when it crossed the virial radius, or about two times the mass of the disk, after its first pericentric passage, falling into the centre of the potential), they still have median ages of around 12 Gyr. This is simply because they could not have formed much earlier in the cosmologies considered here. At these early times, structure is forming rapidly, and halos are acquiring a large fraction of their mass with a very short period of time, which explains why disk ages are relatively insensitive to the exact disruption criterion used.[1]

In what follows, I will take e) or f) as determining disk disruption, given that these criteria depend on the mass of the disk and of the satellite at the time of the encounter, and are therefore easier to interpret, and closely related to the minor merger criterion considered later in this chapter. There is one subtle difference worth noting between criteria c) and d) and criteria e) and f). Once a disk has been disrupted once, its parent galaxy will start forming a new one as before, but the new disk will be much less massive than the surrounding bulge initially. Because of its small mass, this proto-disk is frequently disrupted by infalling satellites, in models using criteria e) or f). Thus, an intial disruptive merger is often followed by many subsequent disruptions,

[1] It also justifies an approximate treatment of satellite dynamics, as mentioned in chapter 4, since it implies that major mergers are rare at late times.

until the disk has a chance to grow larger or the merger rate becomes low enough for it to survive to the present undisturbed. Since a massive bulge may act to stabilise the disk slightly, the actual rate of subsequent disruptions following an initial disk disruption may be overestimated here, but it is clear from figure 8.4 that this does reduce the average age of the disk substantially.

8.1.3 The Estimated Rate of Disk Disruption

Given a definition of disk disruption, specifically criterion e) or f) above, let us now re-examine the rate of disk disruption in more detail. Figures 8.5 and 8.6 show the distribution of disk ages and formation epochs (that is the times when the disk was last disrupted, using this criteria), for SCDM and LCDM respectively. (These are the same data shown in the bottom panel of figure 8.4, plotted differentially.) The solid histogram is for criterion e), while the dashed histogram is is for criterion f). From figures 8.4–8.6, we see that in the SCDM model, two-thirds of disks are 8–12 Gyr old, a third have been disrupted more recently, and up to 20% have been disrupted in the last 5 Gyr. The latter objects presumably correspond to field ellipticals or S0s, which are bulge-dominated and constitute roughly this fraction of giant galaxies (Loveday 1996). The systems with the oldest disks were last disrupted at $z \simeq 2$–3, while the youngest disks formed at $z \simeq 1$. For LCDM, the relative numbers of old and young disks are similar, although the disruption ages are slightly younger. In either cosmology, a typical spiral galaxy (that is a giant galaxy with a large disk) has been in place since $z \simeq 1$, about 9–10 Gyr ago, even assuming disks are easily disrupted (criterion f), dashed histograms); if they are harder to disrupt (criterion e), solid histograms), the mean ages are even older. These ages and redshifts refer to the time when the disk was last disrupted; how much of the disk's mass was in place by this epoch is a separate question I will address next.

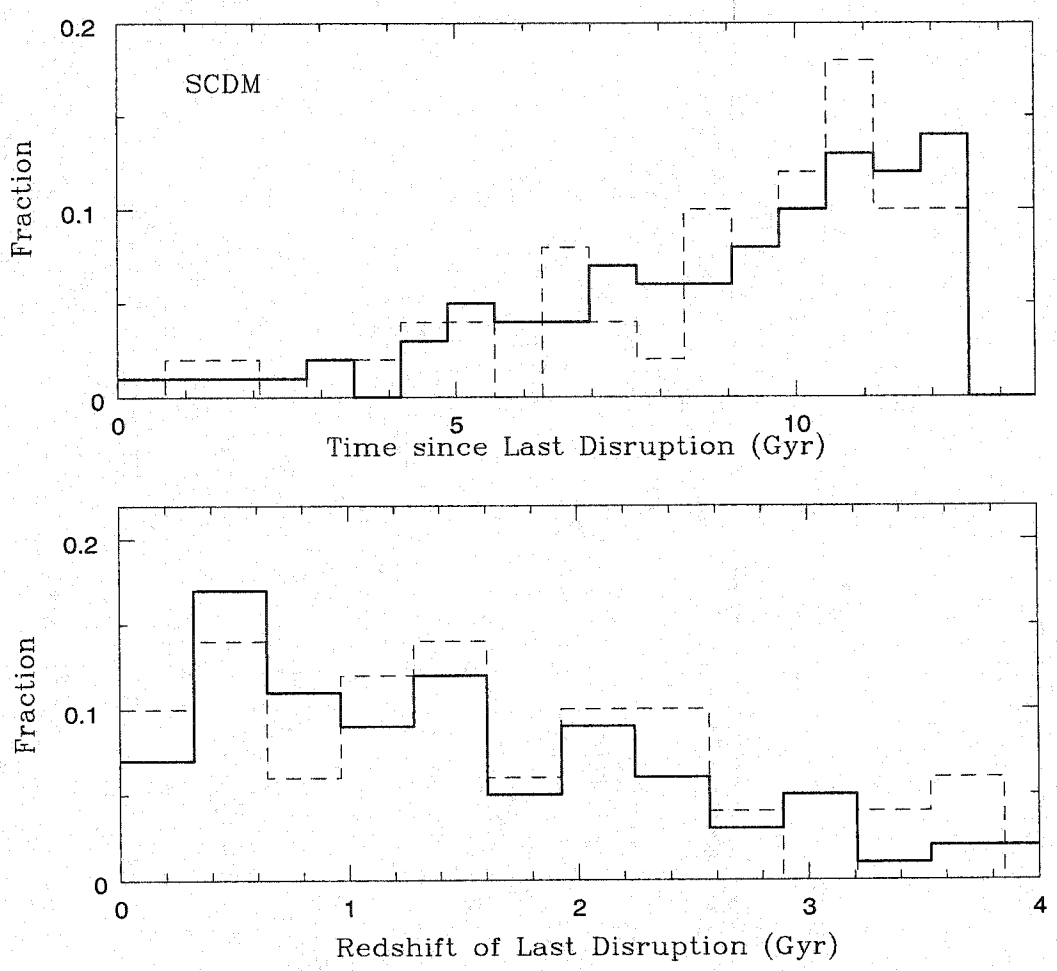

Figure 8.5 The distribution of disk ages and formation epochs, determined using criterion e) (solid histogram) of f) (dashed histogram), for the SCDM galaxies.

8.1.4 The Mass of the Bulge

Finally, there is one last test a model of disk disruption has to pass. Simulations of major mergers show that disks which are completely disrupted form less flattened distributions of stars, similar to elliptical galaxies. This may provide an origin both for isolated, field ellipticals (Toomre & Toomre 1972) and also the central bulges of spiral galaxies (Kauffman, Guiderdoni, & White 1994; Baugh *et al.* 1998). The central bulge of the Milky Way is old, metal rich, only slightly flattened (Gilmore *et al.* 1989), and shows signs of having formed in a rapid starburst (e.g. Prochaska

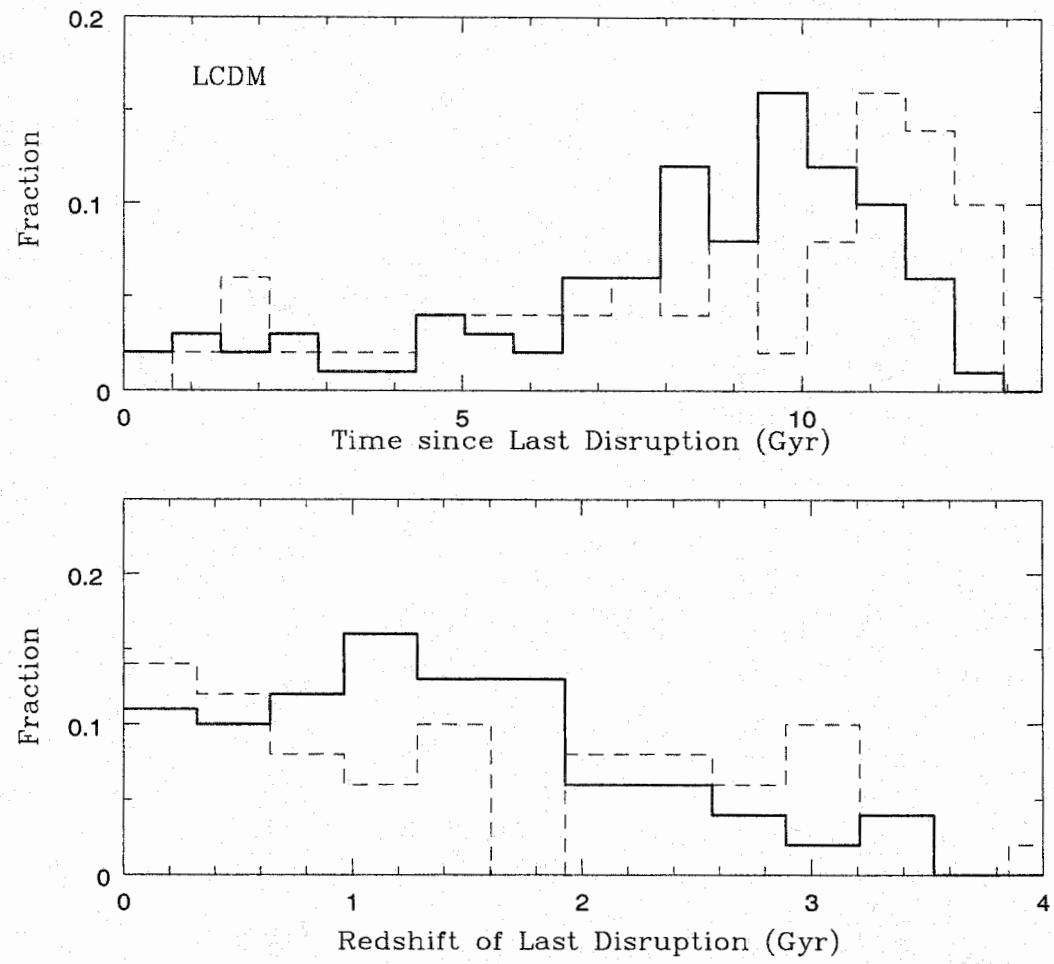

Figure 8.6 As figure 8.5, for the LCDM galaxies.

et al. 2000). Thus it seems plausible that it represents the remains of an early disk, disrupted by a major merger. The total mass of the bulge is less certain; estimates put it at 1/3–1/6 the mass of the disk (Dehnen & Binney 1998). Major mergers in the SA model lead to old disks, but do they also produce reasonable bulges?

In the standard picture established by numerical simulations, a major merger with a gaseous disk causes a starburst, which consumes some fraction of the available gas (Barnes 1996; Mihos & Hernquist 1996). Both the original disk stars and those formed in the burst are distributed spherically, but any gas remaining after the event can cool and form a new disk. Thus, only some fraction of the disk mass is added

to a spherical component in each major merger. Unfortunately, the fraction is not well known. Mihos & Hernquist (1996) found that approximately 95% of the disk mass was converted into a bulge in a major merger, but the disks considered in these simulations were only 20% gas. Early disks and early satellites would have had a much higher gas fraction, so the overall efficiency of the starburst following disk disruption may have been lower. Recently, Barnes (2001) has also found that the conversion efficiency can vary between about 50% and 90% depending on the mass ratio and orbital parameters of the encounter.

Figure 8.7 shows the distribution of disk-to-bulge mass ratios predicted by the SA model at the present day, in SCDM and LCDM cosmologies, using criterion e) (dashed histograms) or f) (solid histograms). In the top panels, the conversion efficiency is taken to be 50%, while in the bottom panels it is taken to be 95%. Even if disk disruption is rare (criterion e), the median bulge-to-disk mass ratio is 0.45 in SCDM and 1.1 in LCDM, and if it is more common (criterion f), the median ratio is 0.9 in SCDM and 1.7 in LCDM. While it is expected that 10–15% of the systems will correspond to field ellipticals, and a further 15% to S0s, the majority of isolated, giant galaxies are spirals (Loveday 1996) with $M_{\rm B}/M_{\rm D} < 1$. Thus, these mass ratios seem far too high.

Making a more quantitative statement is problematic, due to uncertainties in the observed mass ratios of isolated galaxies. Bulge-to-disk luminosity ratios are typically 0.6 or more for field ellipticals, 0.4–0.6 for S0s, and less than 0.4 for spirals (Simien & de Vaucouleurs 1986), but these luminosities are measured in the B-band, where the mass-to-light ratios of the two components may differ by a factor of 2 or more. Studies of galaxy rotation curves (Vega-Beltran *et al.* 1999; Pignatelli & Galletta 1997; Corsini *et al.* 1999), for instance, find much larger mass ratios (e.g. 1.4 for the Sa NGC 2775, 2.8 for the Sa NGC 2179, and 5–9 for some S0s). The rotation-curve decompositions used in these studies do not always include a dark matter component, however, so it is not clear how much of the spherically distributed mass inferred

corresponds to the stellar bulge and halo.

If I ignore these complications and use the luminosity ratios 0.6 and 0.4 to distinguish between morphological types, for the disruption criterion e), then the SCDM galaxies are 60% spirals, 18% S0s, and 22% ellipticals, and the LCDM galaxies are 10% spirals, 22% S0s, and 68% ellipticals. The latter numbers are clearly inconsistent with the percentages 67/18/15, determined in surveys (Loveday 1996), and using a broader disruption criterion or increasing the conversion efficiency makes the discrepancy worse.

There are several possible solutions to this problem, aside from rejecting LCDM as a viable cosmology. The disruption rate calculated here could be overestimated, but it seems unlikely that a disk could encounter a satellite of equal mass and escape intact. The conversion efficiency is less well constrained; in particular, collisions between systems with large gas fractions may be much less efficient at forming stars in sudden bursts. If major mergers of gaseous objects is indeed less efficient, this may in turn have implications for the age, colour and chemical composition of the structural components of giant galaxies. Clearly, simulations of gaseous mergers are required to confirm this hypothesis, but it seems a plausible means of reducing the bulge-to-disk ratios to observed values, given the merger rates predicted above. Finally, it may be that the masses of bulges are genuinely high, as suggested by some of the observations mentioned previously. Accurate mass estimates for a larger sample of objects would be required to confirm this, however. Having considered the rate disruptive encounters in hierarchical models, I will now examine the dynamical evolution of disks, in halos filled with substructure.

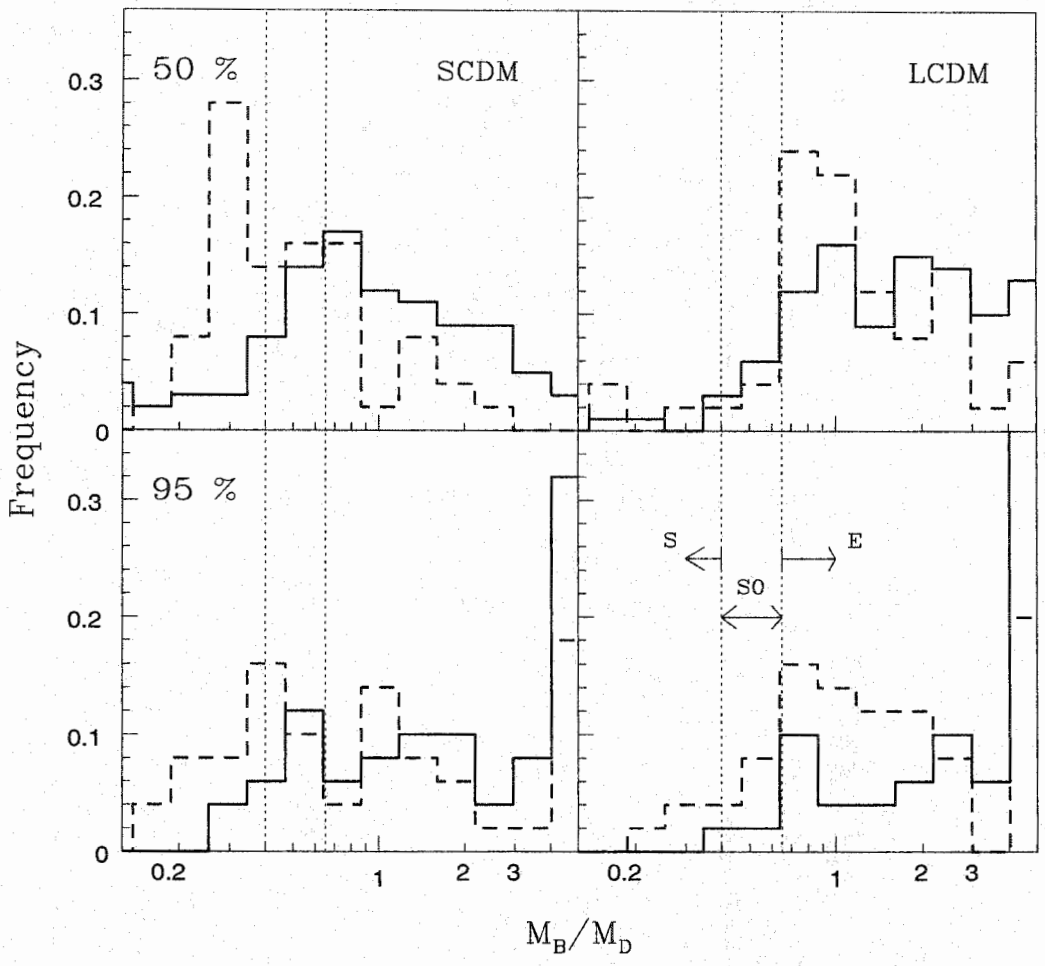

Figure 8.7 The distribution of bulge-to-disk mass ratios calculated for the SCDM and LCDM galaxies, using criteria e) (dashed histograms) and f) (solid histograms), assuming 50% or 95% of the disk mass is converted into a spheroid during each disruption (top and bottom panels respectively). The vertical dotted lines indicate the expected ranges for different morphological types.

8.2 The Consequences of Minor Mergers

Much of the theoretical and numerical work mentioned in the introduction to this chapter sought to determine not so much whether disks would be completely disrupted in hierarchical cosmologies, but rather whether they would be heated substantially by minor mergers below the threshold for disruption. In this section I will estimate

the frequency of such encounters, using the same approach as for major mergers, and also consider some other consequences of minor mergers.

8.2.1 The Origin of the Thick Disk

Local observations in the solar neighbourhood yield a clear relationship between the age of local stars and their velocity dispersion. Stars more than about 10 Gyr old show a substantially increased velocity dispersion, particularly in the vertical direction (Gilmore, Wyse & Jones 1995; Quillen & Garnett 2001; Fuchs 2000). This increased vertical dispersion produces a large scale height for the old population. There is a corresponding relationship between age and metallicity, with a similar jump at 10 Gyr (Buser 2000), and old, high-velocity disk stars also show enhanced abundances of α-elements, indicating that they may have formed in a rapid burst (Prochaska *et al.* 2000). These separate observations suggest the existence of a distinct component, referred to as the 'thick disk'. The thick disk probably corresponds to 20% of the mass of the disk or less, based on stellar age distributions, although these numbers are poorly determined (Rocha-Pinto *et al.* 2000; for cumulative plots see Fuchs 2000, Ferguson & Clarke 2001). Given that the kinematics of the thick disk are exactly those expected for a disk heated by a minor merger, it is interesting to see how often minor mergers produce disk thickening, and whether the age and mass of the thickened component in the SA galaxies are comparable to the properties of the thick disk of the Milky Way.

To determine when the disk was last thickened appreciably by a minor merger requires some merger criterion, just as disruption did. The same choices and complications arise as for major mergers, although it is clearer in this case that a criterion like to e) or f) should be used. For the same reasons mentioned above, I will choose a criterion similar to e) or f), but with a mass ratio of 5:1, to reflect the rough consensus from simulations that a merger with a satellite with 20–30% of the disk mass will

heat the disk substantially (WMH, HC, VW). Thus, I consider the disk to be tidally thickened every time it collides with a satellite which has more than 20% of its mass.

Figures 8.8 and 8.9 show the resulting distribution of ages and epochs of the last disk thickening event, for SCDM and LCDM respectively (with line styles as before). We see most of these events occur around the epoch of halo assembly, when major merger are frequent as well (cf. figures 8.5 and 8.6). As a result, a number of systems may have experienced a thickening event shortly after they formed.

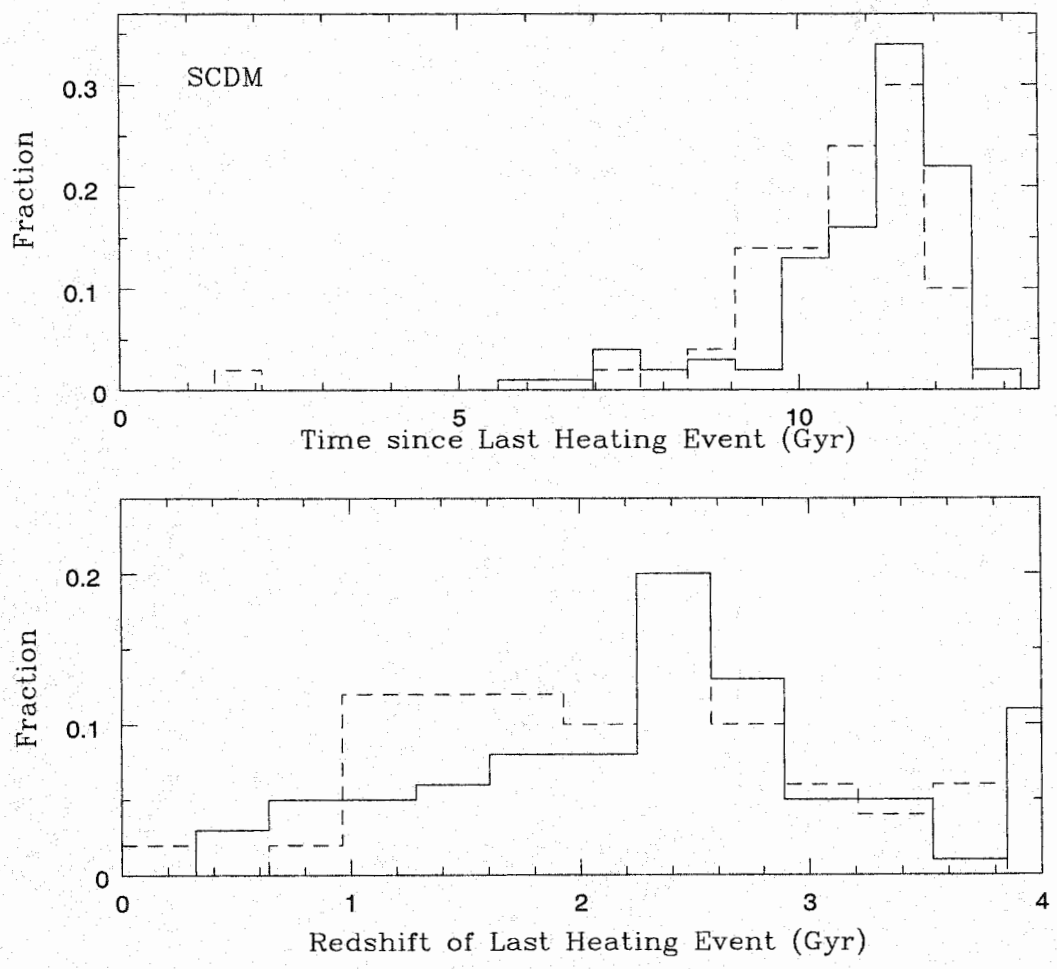

Figure 8.8 The distribution of times and redshifts since the disk was last heated by a minor merger, in SCDM.

Figure 8.10 shows the cumulative distribution of thick disk masses predicted by

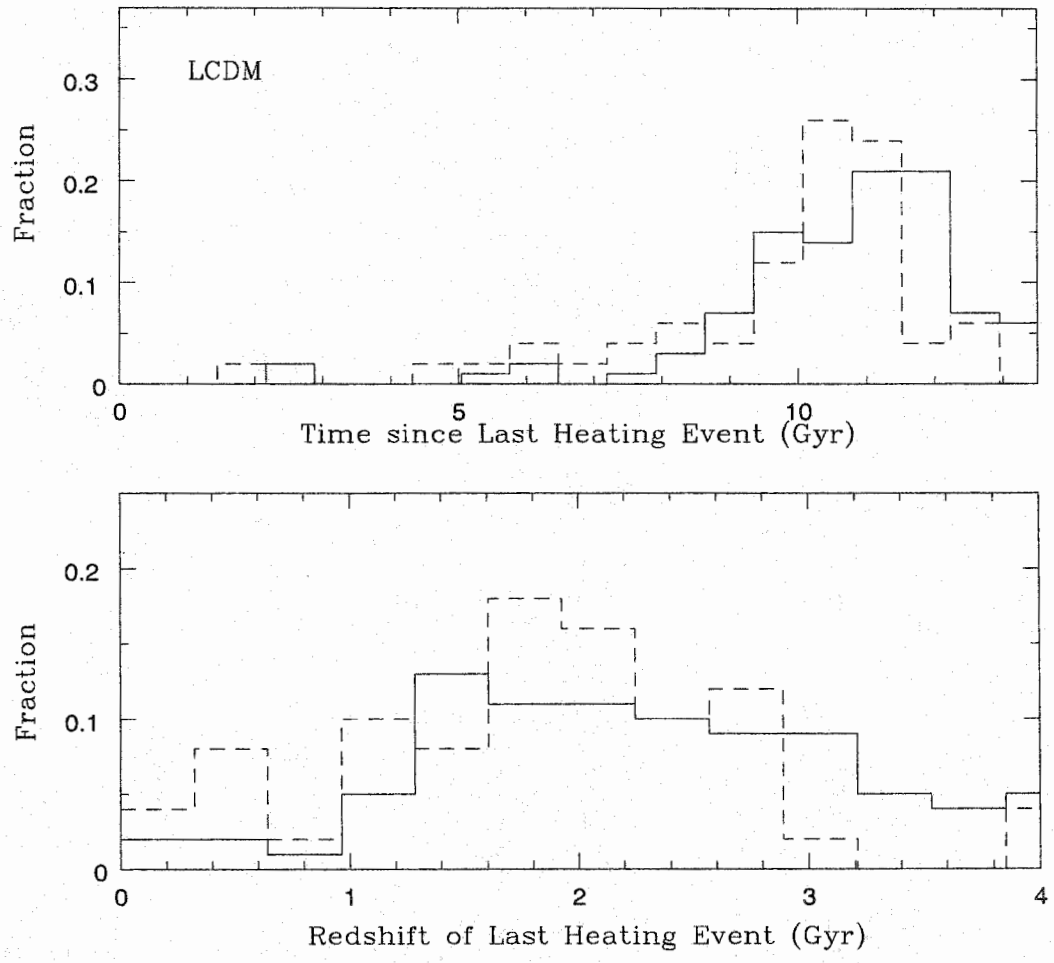

Figure 8.9 As figure 8.8, for LCDM.

the SA model, in SCDM and in LCDM, relative to the final mass of the disk in each system at $z = 0$. The solid lines are the results for criterion 'e' (the disk is disrupted by an encounter with a satellite of equal mass), while the dashed lines are for criterion 'f' (the disk is disrupted by an encounter with a satellite of half its mass). In the top panels, the mass of the thick disk is taken to be the mass of the disk at the time it was last heated. We see that approximately 30% of all systems have a thick disk, and that the thick disk can constitute a large fraction of the total mass of the disk at the present day. This model ignores an important complication, however; whereas stars constitute roughly 80% of the mass in the disk of the Milky Way at present (Dehnen

& Binney 1998), at earlier times, much more of the disk would have been gaseous. This gas could have been heated by a minor merger, but then have dissipated and cooled again, leaving no trace of this event. To give a rough idea of the resulting effect on the mass of the thick disk, in the bottom panel I show the mass of the thick disk, assuming that the thick disk material assembled quickly, and formed stars at a constant rate over time until it was heated, (that is assuming that at time t, a fraction $t/t_{\rm H}$ of the disk was stellar, where $t_{\rm H}$ is the Hubble time). We see that while the same fraction of systems have thick disks, these components now have 10–20% of the total disk mass, similar to the mass of the thick disk component in the Milky Way.

In summary, a model for tidal heating, based on numerical estimates of the threshold for this process, suggests that both the age and mass of the thin disk of the Milky Way, as well as the age and mass of the thick disk, are entirely plausible in hierarchical cosmologies. Given that these observational constraints were originally considered an argument against CDM (or at least SCDM), this result may seem a bit surprising. To clarify the physics behind this result, I will re-examine disk heating rates under some simplifying assumptions.

8.2.2 Explaining the Age of the Thin Disk

If early estimates of the rate of major and minor mergers were so high, why are they so much rarer in the SA model? Does the SA model overlook important physical processes, or does it account for processes ignored in previous work? First, it is worth noting that while the result in TO received much attention, it relied on early and uncertain estimates of the halo merger rate, as well as a simplified treatment of the efficiency of heating. While subsequent numerical work was mainly directed towards clarifying the latter issue, and produced estimates for the heating efficiency $\simeq 50\%$ smaller than those in TO, two papers have also examined the accuracy of the

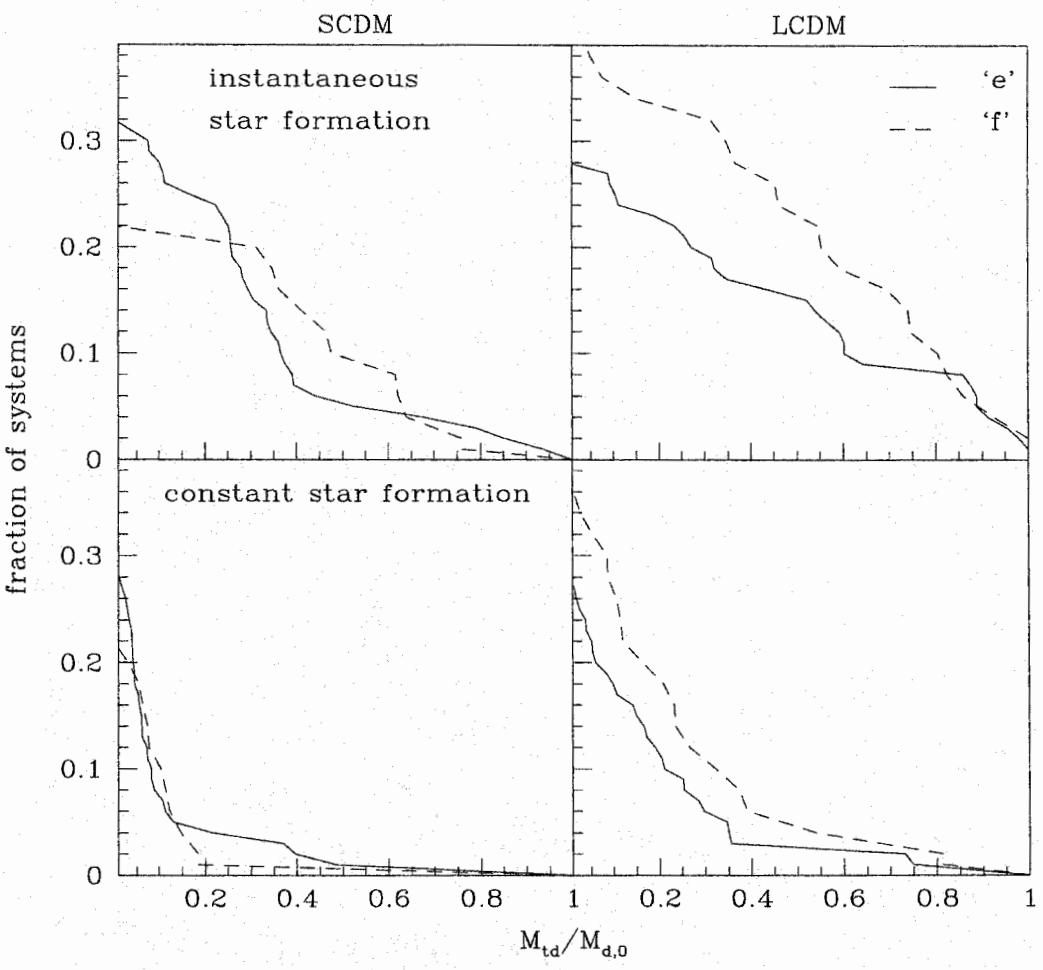

Figure 8.10 The distribution of thick disk masses in SCDM (solid histograms) and LCDM (dashed histograms), calculated assuming the disk is entirely stellar when disrupted (top panel), or forms stars at a constant rate (bottom panel).

assumed merger rate. Kauffmann & White (1993) derived a more rigorous estimate of the halo merger rate, using merger trees. Although their main conclusions were similar to those of TO, namely that the rates were unacceptably high in SCDM, but more reasonable in and open cosmology with $\Omega < 1$, they did note that the galaxy merger rate might be much smaller than the halo merger rate. Navarro *et al.* (1994) re-examined this problem with numerical simulations, finding that the infall time for satellites, which is comparable to the age of the universe at any epoch, introduced a

delay in merging which further reduced the rate of disk heating. Thus the average disk in their SCDM simulations had remained unheated for the last 5 Gyr. While these simulations suffered from a number of technical problems, leading to excessive cooling and the formation of very small galactic disks, they did suggest that the case for frequent disk disruption had been overstated. Combining their low merger rates with the stricter heating criteria determined by simulations of minor mergers, one obtains the estimate that most thin disks are 6–8 Gyr old, in keeping with the results presented above.

Since the physics of the SA model is modular, it is possible to quantify the distinct processes which reduce the disruption rate. Figure 8.11 shows the cumulative distributions of times since the disk was last disrupted, calculated (from top to bottom):

1. with the full model
2. with no dynamical friction
3. with no mass loss from the satellites
4. with neither dynamical friction nor mass loss compared with the rate at which halos experienced 20:1 of 6:1 mergers,
5. counting substructure within branches separately
6. ignoring substructure

Comparing 1) and 2), we see that the net effect of dynamical friction is to make the last major merger slightly more recent, presumably by causing the infall of satellites which would otherwise orbit indefinitely in the halo without hitting the disk. Turning off mass loss while retaining friction has a much larger, and opposite, effect; with mass loss, most disks are $\simeq$ 10 Gyr old, whereas without mass loss, they are $\simeq$ 4 Gyr old. Thus it is essential to account for mass loss when studying halo dynamics.

Comparing panels 3) and 4), we can also see that the high merger rate in the no-mass-loss case disappears if dynamical friction is turned off, resulting in a distribution of disruption times similar to the fiducial one. This indicates that the mergers in panel 3) are caused by the rapid orbital decay of massive satellites, in a model where they lose no mass.

Finally, we can compare these different rates to the corresponding halo merger rates. Panels 5) and 6) show the distribution of times when the halo last experienced a merger with a mass ratio of either 20:1 of more (dotted line), or 6:1 or more (solid line), at its virial radius. Comparing 5) with 6), we see that accounting for the substructure in branches of the merger tree, as described in chapter 6, reduces the rate of halo mergers of a given mass ratio slightly, because it decomposes branches into smaller subcomponents. The overall effect is more important for the more massive mergers, however (compare 6:1 and 20:1). Comparing 5) with the fiducial rate in 1), we see that subhalos with masses comparable to or greater than the central galaxy cross the virial radius at later times, but have not disrupted the disk by $z = 0$, either because they are still falling in, or because tidal stripping has reduced them in mass.

Overall, the results shown in figure 8.11 demonstrate the complexity of merger statistics, which depend on disk and halo growth rates, halo merger rates, infall times and mass-loss rates. It is because of this complexity that a semi-analytic model is required to estimate accurate disk disruption rates.

8.2.3 The Epoch of Major and Minor Mergers

Major and minor mergers may produce effects that are directly observable as they occur, including strong bursts of star formation, quasar and AGN activity, and nearby ULIRGS (Mihos & Hernquist 1996; Mihos 2001). It is not the goal of this thesis to predict realistic rates for this activity. To do so would require, at a minimum, generating merger trees for objects of many different masses, and convolving the

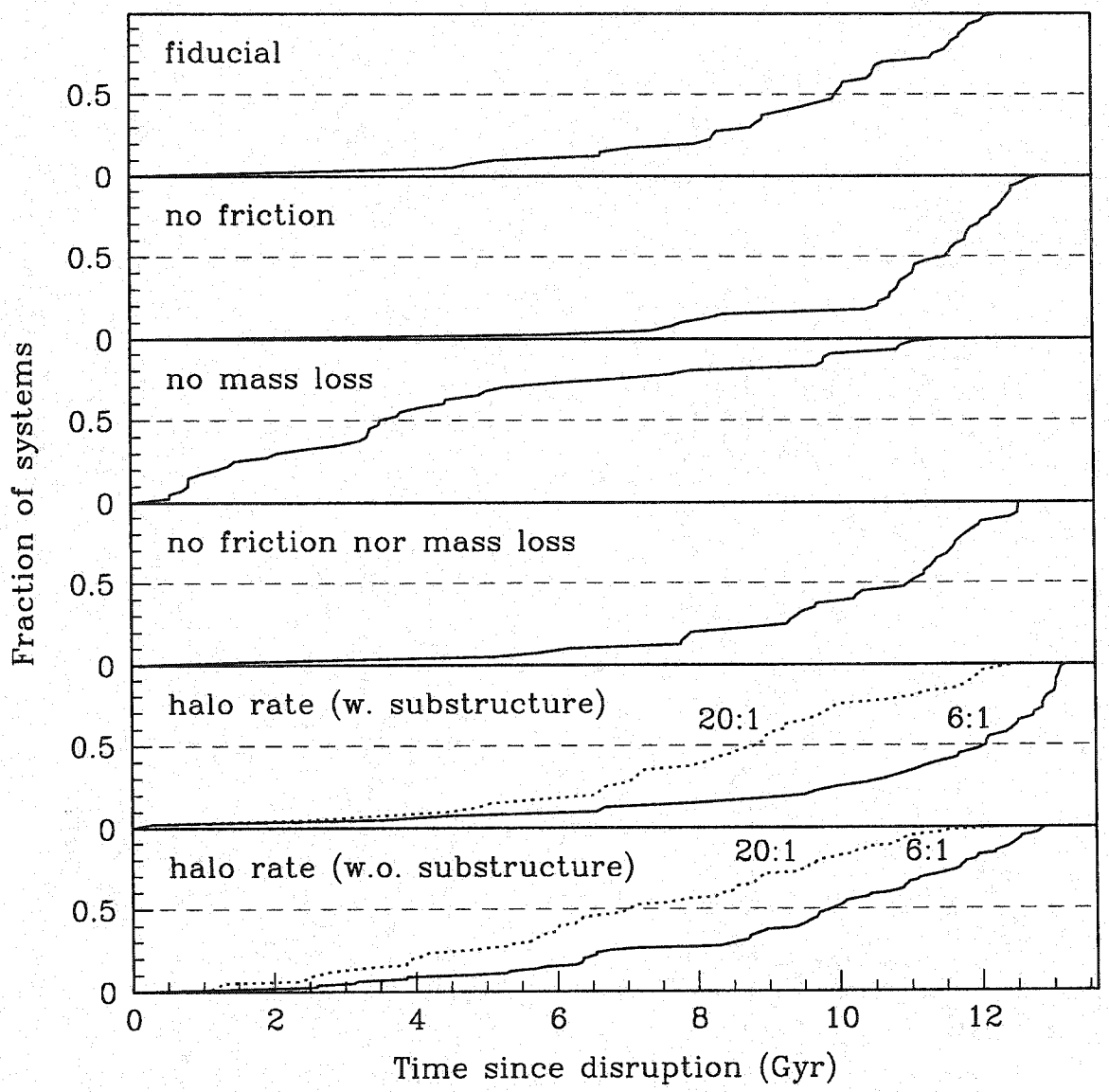

Figure 8.11 The cumulative distribution of disk ages, calculated with different physical processes turned off (middle panels), compared with the fiducial rate (top panel) and the corresponding halo rates (bottom panels).

results with the halo mass function. Furthermore, observations of merger-induced activity depend on many complicating selection effects. Nonetheless, the SA model can give a general sense of the rate and intensity of mergers with redshift.

Figure 8.12 shows the distribution of merger events in the SCDM and LCDM

runs, weighted by the total mass of material involved in each case. The large solid points indicate major mergers (mass ratios of 2:1 or less), while the small points indicate minor mergers (mass ratios of 10:1–2:1). Redshift is plotted logarithmically along the bottom axis. This figure shows that the distribution of major and minor merger events is similar, although minor events are more common, and that these events typically take place before $z \simeq 2$–3 in LCDM, or $z \simeq 1$ in SCDM. On the other hand, because galaxies are simultaneously growing in mass over time, it is the late mergers that involve the largest total masses of material. Thus, if mergers convert some fixed fraction of the available gaseous mass into stars, or funnel it into an active nucleus, we would expect these encounters to be the most luminous. This combination of increasing merger mass and decreasing merger rate may account for the peak in quasar activity around $z \simeq 2$–3 (Pei 1995), although the problem clearly requires more detailed treatment, as in Haiman & Menou (2000).

8.3 Disk Heating

So far in this chapter I have considered only the effects of direct collisions on the dynamical state of galactic disks. Thin disks are inherently unstable, however, and therefore sensitive to much weaker disturbances (indeed, it is extremely difficult to keep the disks in numerical simulations stable, even in isolation). Given the amount of substructure predicted to survive in galaxy halos, the combined effect of these weak disturbances may heat disks substantially. Direct collisions below the threshold mass ratio for minor mergers may also contribute to heating, thickening the disk gradually over time. In this last section, I will apply two previously developed analytic estimates of the heating rate produced by substructure, to determine the importance of this effect.

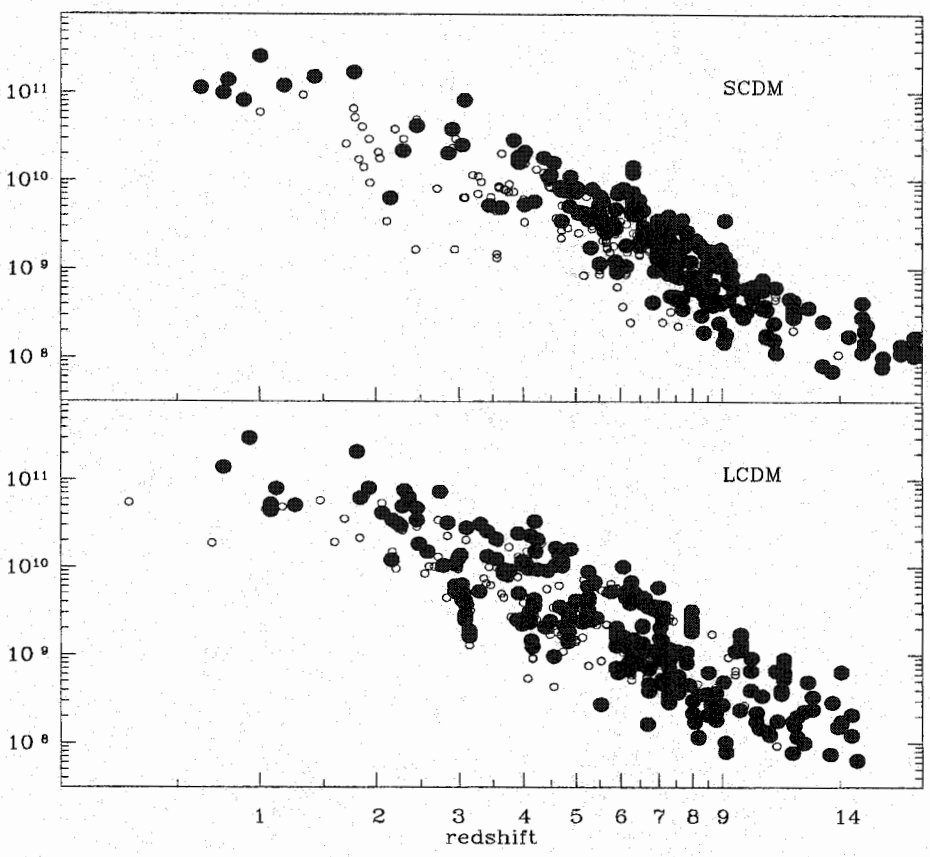

Figure 8.12 The redshift distribution of major and minor mergers (solid and open circles respectively), plotted versus the total mass of material involved.

8.3.1 Analytic Estimates

Heating is a collisional process, that is it depends on the transfer of kinetic energy from a perturbing satellite to individual disk stars or particles, via two-body interactions. To calculate the heating rate in a given encounter, one has to average over the effects of individual encounters, integrating either over the distribution of perturbers, or over the distribution of background particles. Heating thus represents the other part of the calculation that led to dynamical friction – in the simplest terms, one could imagine that the energy each satellite loses through dynamical friction winds up in the background as kinetic 'heat'. In practice, however, there are several sources of energy involved in the process, including the kinetic energy of the satellite but also

the potential energy of the disk, and the kinetic energy associated with its coherent rotation, while this energy is redistributed into several sinks, including the orbital energy of material stripped from the satellite and the kinetic energy of coherent disk modes, as well as the 'thermal' kinetic energy of the disk. Despite this complexity, heating is ultimately mediated by two-body interactions, and therefore by using something like a dynamical friction calculation (or more rigorously, fluctuation-dissipation theory), with a suitable efficiency parameter, it should be possible to determine an analytic expression for it.

In order to establish the relative efficiency of heating, one needs to derive a steady heating rate, based on the net effect of a single encounter. This requires averaging the effects of a single satellite on many nearby disk particles, or averaging the effects of a large number of evenly distributed satellites on a single disk particle. The first approach was developed by TO. It estimates a local heating rate proportional to the amount of energy lost to disk dynamical friction locally, summed over all satellites that pass through a given annulus. Given that disk dynamical friction is modulated by the local density of the disk, which drops off rapidly with height in the disk plane, this form of heating is appreciable only for close encounters with the disk, and therefore accounts for weaker versions of the disruptive encounters considered in the previous sections.

The other way to estimate a continuous heating rate is to average over the effects produced by a large number of more distant satellites. This approach was used by Lacey & Ostriker (1985; see also Lacey (1984) and Ipser & Semenzato (1985)) to determine the dynamical effect of a halo full of massive black holes. It requires an estimate of the local density of perturbers, which are assumed to be numerous and uniformly distributed, and in this sense will be approximate for the less common, more massive satellites, but does have the advantage of including the effects from the entire halo. I will proceed to compare these two heating estimates below.

8.3.2 Close Encounters

In a single hyperbolic encounter between two point particles, it is possible to calculate the net velocity changes for either particle, in the directions parallel to, and perpendicular to, the initial relative velocity (Binney & Tremaine 1987). Averaging over a series of encounters, for a given geometry and velocity distribution, gives the net changes in $< V_\perp >$, $< V_\parallel >$, $< V_\perp^2 >$ and $< V_\parallel^2 >$. By comparing these net changes, it is possible to determine the ratio of the energy lost by a satellite through dynamical friction, to the energy gained by the disk in plane, $\epsilon_\parallel$, and out of plane, $\epsilon_\perp$.

TO considered a satellite spiraling in on a quasi-circular orbit, and assumed that fractions $\epsilon_\parallel$ and $\epsilon_\perp$ of the orbital energy lost as it fell through a given annulus were contributed to the random motions of stars in that annulus. Thus, for instance, they assumed a change in the surface energy density:

$$\Delta\epsilon_z = \frac{\Delta E}{2\pi r dr} = \frac{\epsilon_\perp \Delta E_{\rm sat}}{2\pi r dr} \tag{8.1}$$

for an annulus extending from r to dr, where $\Delta\epsilon_z$ is the change surface energy density going into vertical motion, ΔE is the total energy input into the annulus, and $\Delta E_{\rm sat}$ is the energy lost by the satellite. They then considered the response of the disk and its return to virial equilibrium, showing that this energy produces a change in scale height:

$$\begin{aligned} \frac{\Delta H}{H} &= \frac{2}{3}\frac{\Delta\epsilon_z}{w_{\rm dd}}\left[1+\frac{8}{3}\delta(r)\right]^{-1} = \frac{2}{3}\frac{\Delta\epsilon_z}{\pi G\Sigma^2 H}\left[1+\frac{8}{3}\delta(r)\right]^{-1} \\ &= \frac{2}{3}\frac{\epsilon_\perp \Delta E_{\rm sat}}{2\pi^2 r dr G\Sigma^2 H}\left[1+\frac{8}{3}\delta(r)\right]^{-1} \end{aligned} \tag{8.2}$$

where $\Sigma = \Sigma(r)$ is the surface density of the disk, $w_{\rm dd}$ is the disk self-energy and $\delta(r)$ is the ratio of the disk and halo self-energies, $w_{\rm dd}$ and $w_{\rm dh}$. I will use this expression to calculate the change in disk height produced by each encounter with the disk, and the related expression:

$$\Delta\sigma_w^2 = \left(\frac{\Delta H}{H}\right)\left(\frac{w_{\rm dd}+4w_{\rm dh}}{\Sigma}\right) \tag{8.3}$$

for the change in the velocity dispersion.

For thin disks, $\delta(r)$ is small; for the disk model considered by TO, for instance, which has the same vertical profile as the model used in the SA code, $\delta(r) \simeq 0.1$. The exact value of $\epsilon_\perp$ also varies between 0.4 and 0.7, depending on the distribution of orbits in the disk. I will ignore the uncertainty in these two terms, and use an effective coefficient of 0.4 for the change in H. TO also assumed slowly decaying, quasi-circular orbits, such that satellites only passed through a given radius once. For general orbits, it is not clear how to distribute the energy lost during each encounter with the disk. I have chosen to distribute it over and area equal to the size of the satellite, assuming that this is roughly the size of the region that responds strongly to the encounter. Thus, a fraction $dr/(2r_{\rm sat})$ of the energy lost by the satellite goes into heating an annulus of width dr if it lies within $r_{\rm sat}$ of the satellite's position. The resulting expression for the change in scale height is then independent of the width of the annulus:

$$\Delta H = \frac{k_1 \Delta E_{\rm sat}/2r_{\rm sat}}{r\Sigma^2} \tag{8.4}$$

where $k_1 = \epsilon_{\perp,1}/3\pi^2 G$ and $\epsilon_{\perp,1} \simeq 0.4$. The corresponding change in the vertical velocity distribution is:

$$\Delta\sigma_w = \frac{k_2 \Delta E_{\rm sat}/2r_{\rm sat}}{r\Sigma} \tag{8.5}$$

where $k_2 = \epsilon_{\perp,2}/3\pi$ and $\epsilon_{\perp,2} \simeq 0.6$.

Figures 8.13 and 8.14, for instance, show the cumulative changes in ΔH and $\Delta\sigma_w$, produced by direct encounters since a given time, at $r = r_\odot = 8.5$ kpc in a typical disk, for four different SCDM and LCDM halos respectively. The dotted line shows the rate at which the disk builds up. We see that the only substantial direct heating is at early times, before most of the mass of the disk is in place. This heating can produce an old, thickened disk component in some cases (e.g. the second panel of figure 8.13). Because close encounters at late times are rare, however, and because the main mass of the disk builds up after these encounters cease to be common, the contribution of

direct heating to the mass-weighted velocity dispersion and mean scale-height of the disk will negligible. Since the average scale height and velocity dispersion of the disk are much larger than the initial values produced by star formation, other mechanisms must be responsible for this heating. This problem has been studied extensively (cf. Binney & Tremaine 1987 for an overview), and many possible heating agents, including spiral arms and scattering off giant molecular clouds, have been suggested. Another possibility, however, is that the potential fluctuations produced by satellites orbiting in the halo, which are weaker but more common than close encounters, can heat the disk substantially. I will consider this possibility next.

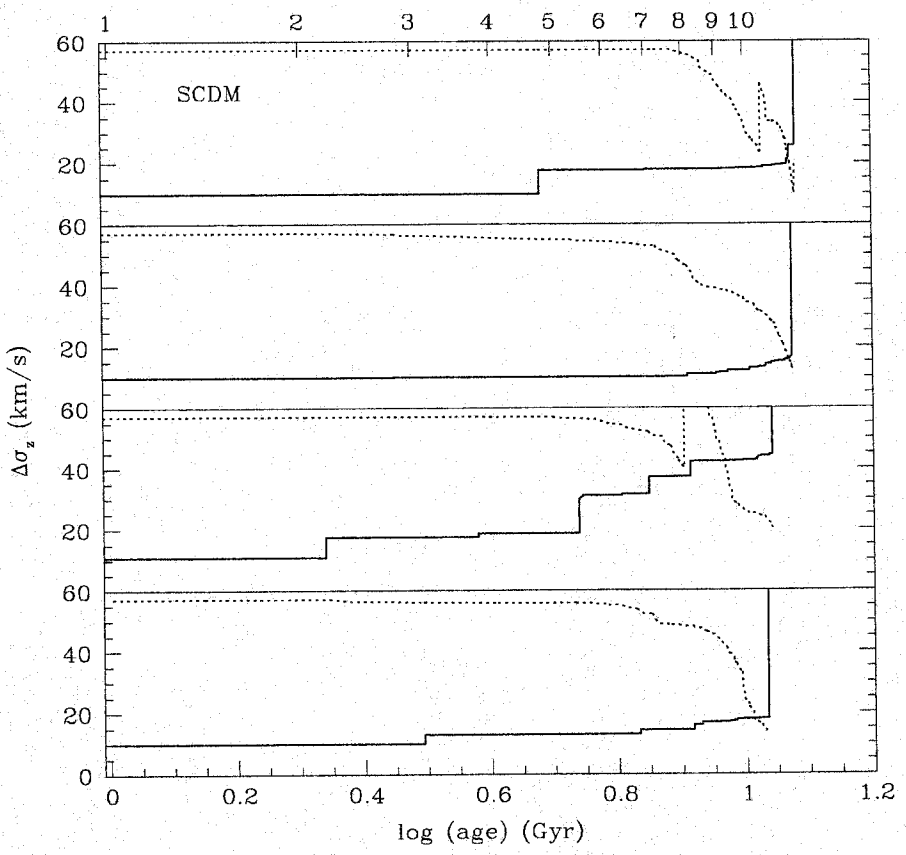

Figure 8.13 The change in the stellar vertical velocity dispersion, $\Delta\sigma_w$, produced over time by direct collisions with several typical SCDM disks, as a function of the age of the stars. The dotted line shows the increasing mass of the disk versus time.

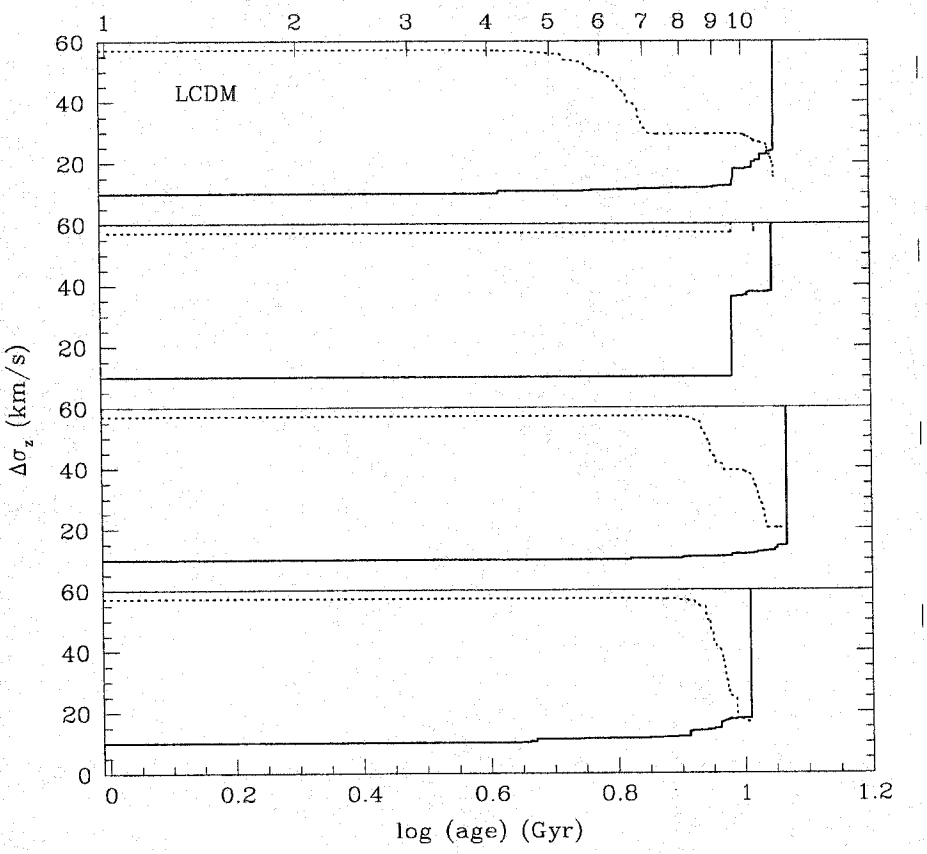

Figure 8.14 As figure 8.13, for LCDM.

8.3.3 Distant Encounters

Lacey (1984), Lacey & Ostriker (1985), and Ipser & Semenzato (1985) used the same coefficients for the change in $< V_\perp >$, $< V_\parallel >$, $< V_\perp^2 >$ and $< V_\parallel^2 >$ to the mean velocity change of an individual star, produced by a uniform distribution of small perturbers, orbiting in the halo. Specifically, they derived an expression for the change in the velocity dispersion of a population of disk stars:

$$\frac{d\sigma_w^2}{dt} = \frac{\sqrt{2}\pi G^2 n_{\rm p} M_{\rm p}^2 \ln\Lambda}{\sigma_{\rm p}} H(\gamma) \tag{8.6}$$

where $n_{\rm p}$ is the local number density of perturbers of mass $M_{\rm p}$ and velocity dispersion $\sigma_{\rm p}$, $\ln\Lambda$ is the Coulomb logarithm defined in chapter 2, and

$$H(\gamma) \simeq [(2\gamma^2 - 1)\mathrm{erf}(\gamma) + \gamma\,\mathrm{erf}'(\gamma)]/2\gamma^3, \tag{8.7}$$

with $\gamma \simeq V_c/\sqrt{2}\sigma_p$.

To apply this heating in a given timestep, one has to determine the local number density of perturbers, and account for differences in their masses. Since there are a finite number of satellites in th SA model, I do this by counting up the number of satellites in a given volume, and dividing by the volume to get a number density. The different masses are easily accounted for by weighting the contribution of each satellite by the square of its mass. The resulting heating rate depends only on a single parameter, the volume over which the mean density of satellites is calculated. This introduces some uncertainty in this heating calculation, but it still provides an interesting estimate of the disk heating rate, as I show below[2] .

Figures 8.15 and 8.16 show the vertical velocity dispersion for disk stars as function of their age, calculated using this approach, in three different SCDM runs and three LCDM runs (note that the stellar ages have been rescaled to match the age of the universe of 18.6 Gyr implied by the observations). Each solid line shows the result calculated for a different volume; the volumes used were spheres with radii of 20, 40, and 80 kpc. The scatter between these lines gives an estimate of the uncertainty in the heating rate, in each case. The points are the observed velocity dispersions of disk stars, compiled by Quillen & Garnett (2001) (there is some disagreement over the in-plane velocity dispersions determined in this work, cf. Fuchs (2000), but more general agreement on the vertical velocity dispersions, which are shown here). The total velocity dispersion was calculated by assuming that the value for the youngest of these points, $\simeq 9\,\mathrm{km\,s^{-1}}$, is the initial value produced by star formation.

As with the direct heating calculation, we see that most disk heating from potential fluctuations occurs at early times. In particular, this form of heating has been relatively unimportant for the past 3–5 Gyr. The increase in the observed velocity

[2]In fact, some of this volume-dependence may also cancel out, if the Coulomb logarithm, which is based on a weighted integral over the distribution of perturbers, is adjusted upwards for larger volumes.

dispersion during this time must therefore be due to the other known sources of heating, internal to the disk, such as spiral arms and giant molecular clouds. These are expected to produce heating which increases steadily as $t^{1/2}$. The dashed lines show the range of age-dispersion relations this form of heating would produce, based on the measured dispersions in the youngest bins. We see that although these mechanisms may account for the increase in heating back to $\simeq 5$ Gyr ago, prior to this potential fluctuations become an increasingly important source of heating. In particular, the observed velocity dispersions in the oldest bins (corresponding to the thick disk) are larger than expected from internal heating alone. Figures 8.15 and 8.16 show that this old component, which I suggested previously was the result of a minor merger, could also be the result of an indirect encounter.

In fact, the distinction between direct collisions and indirect encounters may be rather arbitrary, since the heating rate calculated using equation (8.6) is generally higher if the satellite density is averaged over a smaller volume, indicating that the dominant effect is from fairly close encounters. The contribution from the largest satellites also dominates over the more numerous small satellites. Unfortunately, the simplifying assumptions made in deriving equation (8.6) also break down in precisely these cases. It therefore seems worth examining the possibility that indirect encounters with the few dozen most massive satellites in a halo could heat the disk substantially over its lifetime, using numerical simulations. This will not be an easy task, since it requires forming a galactic disk in cosmologically realistic conditions, and keeping it stable for a Hubble time. It is a challenge for the next generation of numerical studies of disk evolution.

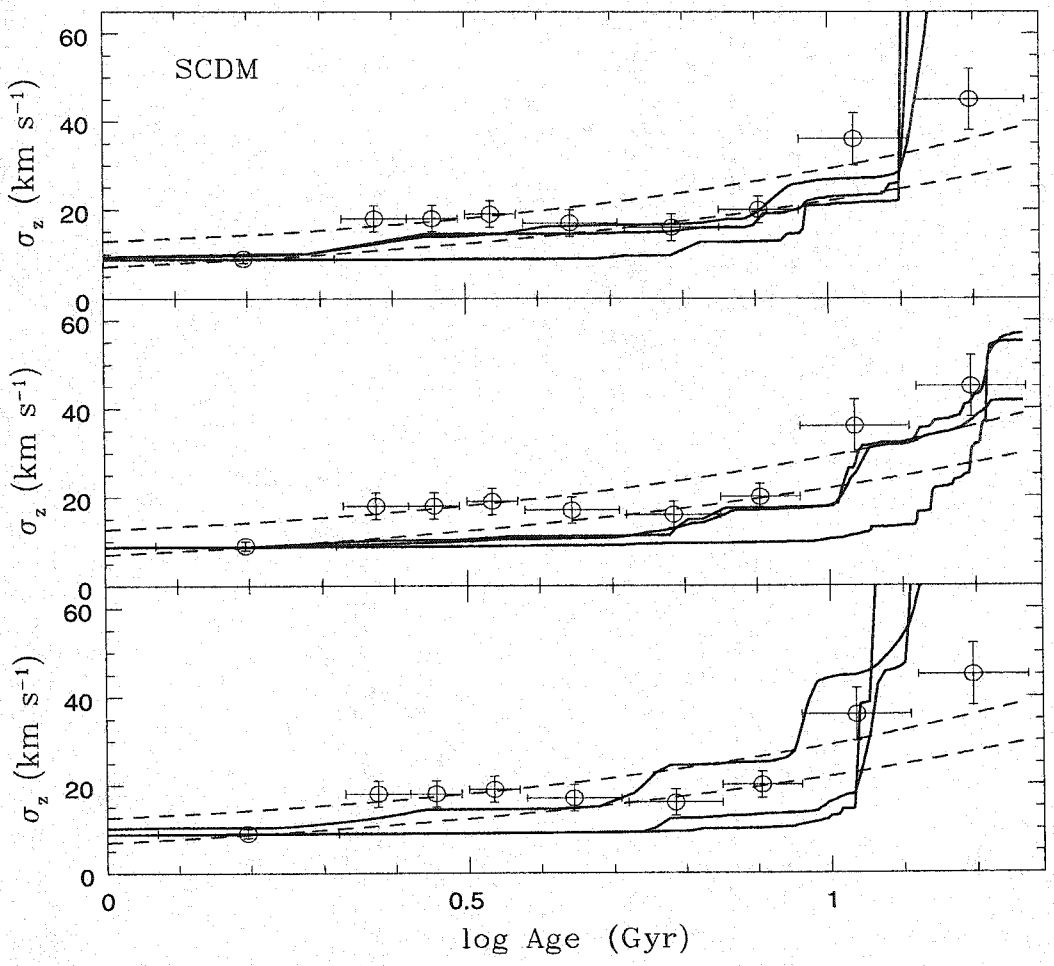

Figure 8.15 The velocity dispersion produced by potential fluctuations in three SCDM galaxies, as a function of stellar age. The solid lines show the results of averaging the density of satellites over different volumes in the heating calculation. The dashed lines show the effect of heating which goes as $t^{1/2}$. The data points and error bars are observations of disk stars, compiled by Quillen & Garnett (2001).

8.4 Summary

In this chapter, I have determined the frequency and effects of collisions and encounters, commonly thought to be a major threat to disk stability in hierarchical models of galaxy formation. The dynamics of individual encounters are sufficiently

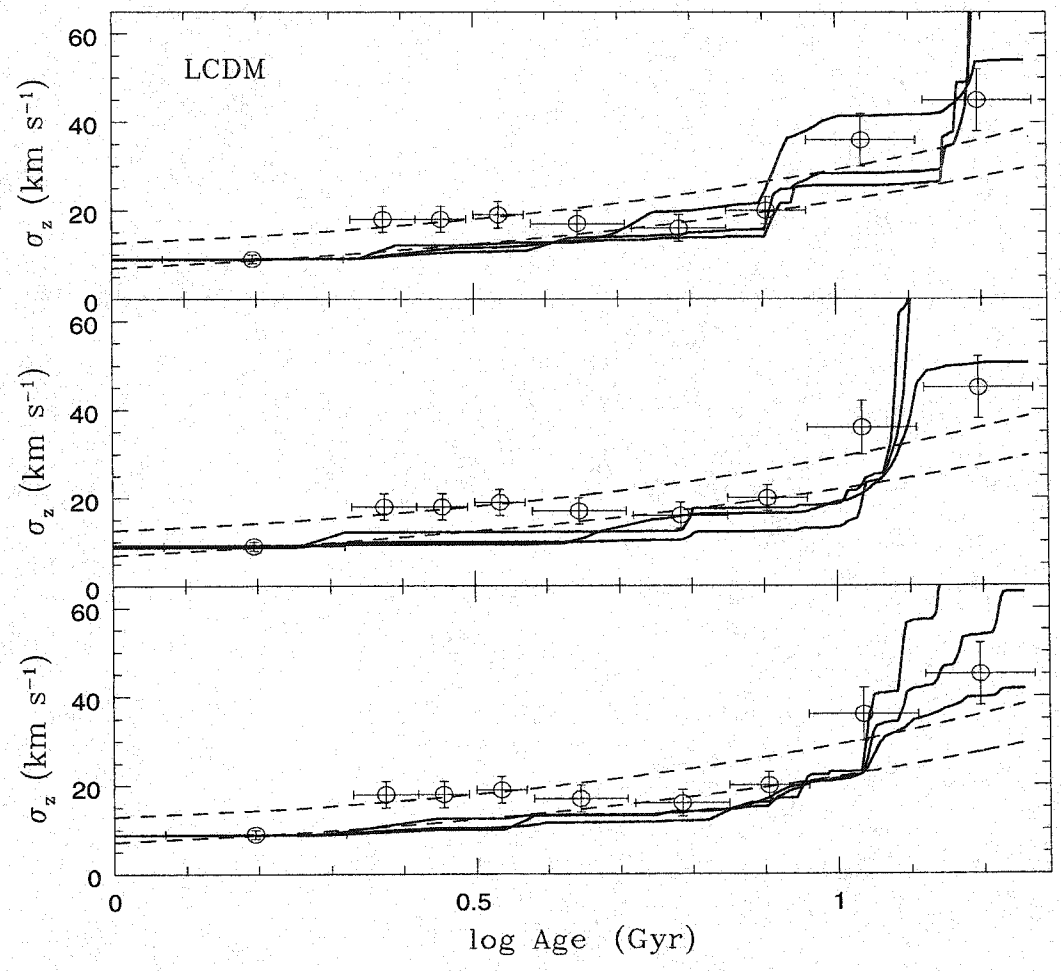

Figure 8.16 As figure 8.15, for three LCDM galaxies.

complex that they can only be determined accurately using numerical simulations. The rate of these encounters is also sufficiently difficult to estimate, however, that semi-analytic techniques are required to determine the implications of a given set of numerical results. Using the SA model developed in this thesis, I calculate the rate of disruptive mergers, which destroy disks completely, and of minor mergers which heat disks substantially, as well as the integrated effects of direct heating from smaller collisions and of indirect heating from potential fluctuations in the halo. I find that in general, old, thin disks with ages of 8 Gyr or more are common both in LCDM and in SCDM, and that small, thickened components 1–2 Gyr older than the main

part of the disk also arise naturally in approximately 30% of the systems in either cosmology. Thus, the observed structure of the Milky Way and other nearby galaxies does not seem to preclude hierarchical galaxy formation in a standard cosmology.

One or two unresolved problems remain, however. The mass ratios of galactic bulges and disks appear to be too large in LCDM – rough estimates for the simplest model suggest that more than half of the present-day galaxies correspond to S0s or ellipticals. While stricter disruption criteria would reduce this problem, they are probably not warranted; a more likely explanation is that early disks are primarily gaseous, and that this gas re-forms a new disk after a major merger, rather than being turned into a stellar bulge. Even allowing for this possibility, however, the average bulge-to-disk mass ratio is somewhat higher than expected. To elucidate this point, it would be useful to have a better sample of observed bulge-to-disk mass ratios, for galaxies of different morphological types.

The disk heating from the potential fluctuations produced by indirect encounters is also predicted to be quite large, the dominant effect coming from encounters with the nearest and most massive satellites. Unfortunately, in this regime, the analytic estimates of heating are least reliable. Thus, a detailed and careful numerical simulation of disk evolution in a halo filled with substructure will be required to verify the importance of this mechanism, in generating the velocity dispersion observed in old disk populations.

There are also some obvious extensions to the SA model which would provide a clearer picture of disk formation. The most important is the inclusion of realistic disk-formation and star-formation rates, based on analytic estimates or hydrodynamical results. This would be a particularly interesting first step towards a full semi-analytic model, combining the dissipative physics of previous models with the dynamical calculations developed in this thesis.

Chapter 9

Conclusion

9.1 The Motivation for this Work

Galaxy formation is undoubtedly a complex process. In hierarchical models, it proceeds through many distinct steps, including the gravitational collapse of dense regions in the early universe, baryonic dissipation, star formation, mergers between small star-forming regions, repeated heating and cooling of the gas around these objects, disk formation, and disk disruption; and these various steps are repeated on increasing mass scales and at decreasing rates as the universe expands. Clearly this sequence of events is too complex to be represented by any simple analytic model. The analytic models of galaxy formation that have been developed consider only limited aspects of the process, under greatly simplified assumptions.

For the two decades that galaxy formation has been a major topic of research, numerical models have therefore dominated the field. The earliest and most common models considered the dissipationless dynamics of matter as it assembled into the dense lumps from which galaxies formed; more recently, hydrodynamic codes have allowed the actual gas dynamics, and even star formation, to be modelled numerically. It seems likely that these hydrodynamical simulations will produce convincing galactic models in the near future. Despite their considerable predictive power in many areas, however, even these models represent only a crude approximation to the microscopic structure of star-forming regions, just as stellar models average over the detailed motion of gas on small scales, inside stellar interiors.

Numerical models have another, more significant disadvantage, however, namely the computational expense they incur. Current high-resolution simulations require

days, weeks or months of CPU time to follow the evolution of galaxy halos in detail. In heavily constrained problems with few free parameters, it is reasonable to run a few very large simulations, even if the expense is considerable. Thus, for instance, a set of four one-billion particle simulations of the evolution of the entire observable volume of the universe have recently been run on the world's most powerful supercomputers, each with one of the few sets of cosmological parameters of interest currently. Galaxy formation is more complex and less well understood than the formation of larger-scale structure, however; typical galaxy formation models involve dozens of free parameters and many uncertain assumptions. It is simply not possible to investigate every option computationally.

There are some broad patterns in galaxy formation, that make aspects of the problem amenable to a combination of numerical and analytic techniques, however. Galaxy halos, for instance, appear to have a regular structure with certain universal properties, judging from numerical simulations. Their substructure also seems to form in a predictable and regular way, through a fairly simple process of hierarchical merging on progressively large scales. Over the past decade semi-analytic methods, which combine numerical codes with analytic results, have capitalised on these regular features of galaxy formation, producing predictions for the general properties of large populations of galaxies, while leaving the detailed structure of individual galaxies to numerical techniques. In this thesis I have developed a new semi-analytic model which considers the evolution of individual galaxies, and in particular galaxy halos, in detail. This model attempts to bridge the gap between existing numerical and semi-analytic models, in order to develop a way of studying statistical problems in galaxy dynamics that is computationally feasible.

9.2 Galaxy Formation in Hierarchical Cold Dark Matter Models

According to the current cosmological paradigm, galaxies form within much larger and more massive halos of dark matter. The early dynamical evolution of galaxies is thus dominated by the dynamics of dark matter, which are much simpler to understand and to model. The initial evolution of dark matter can be understood in terms of a simple model of spherical collapse. An isolated fluctuation in the early universe, that is a region slightly denser than its surroundings, will amplify with time, becoming denser until it ceases to expand along with the background universe, but instead turns around and recollapses. In the latter stages of collapse, more complex processes will redistribute potential and kinetic energy, until the system reaches virial equilibrium, a relatively stable state for non-relativistic matter distributions interacting through gravity alone.

In the case of a symmetric, isolated density fluctuation, collapse and virialisation proceed at a well-defined pace, so that the size, mass and density of the resulting virialised structure can be predicted at all times. In the more realistic context of a universe filled with many density fluctuations on different scales, different rates of collapse will cause separate halos, or virialised regions, to form at different times, possibly incorporating material which has itself already collapsed on a smaller scale. In cosmologies dominated by CDM, in particular, the amplitude of fluctuations is greatest on small scales. As a result, by the time galaxy-scale halos form, they do so by merging with and incorporating many smaller halos. Thus structure forms hierarchically from small scales to large scales as the universe expands.

Since the density of halos is some roughly constant multiple of the background density of the universe at the time they formed, the smaller, older halos absorbed by galaxy halo will also be denser than the system into which they are incorporated. Analytic and numerical work shows that under these conditions, substructure will survive within the main halo, taking the form of small, dense, roughly spherical

cores, orbiting in a smooth, roughly spherical background. While the outer parts of galaxy halos are less relaxed and more irregular, within the virial radius a two-component model, consisting of a smooth spherical mass distribution following a universal density profile, and a set of much smaller, self-similar profiles corresponding to individual subhalos, or cores of accreted halos, provides a very good approximation to the matter distribution seen in simulations.

The dynamics of small, symmetric objects moving in a smooth potential can be described fairly accurately in analytic terms. Thus, given initial conditions within a galaxy halo, it is possible to construct a purely analytic model of its subsequent evolution (analytic in the sense that it relies on analytic formulae to describe the main dynamical processes, although these formulae are then applied through numerical integration in some cases), which capture the essential details seen in more expensive numerical simulations.

9.3 An Analytic Model of Satellite Dynamics

In the first part of the thesis (chapters 2–4), I develop an analytic model of subhalo dynamics, suitable for studying the dynamical evolution of galaxy halos. Beyond straightforward orbital calculations, which treat subhalos as unchanging point masses moving in a static potential, this model includes three other components, which provide the most important corrections to the point-mass approximation: dynamical friction, tidal stripping, and tidal heating.

Dynamical friction is discussed in chapter 2, after a brief description of the general outline of the dynamical code and the basic orbital calculations. It is a drag force, produced by two-body interactions between the subhalos and individual particles in the background halo, or more generally, in a smooth background, by the response of the background to the satellite's passage. While more rigorous descriptions of

dynamical friction exist, they are generally too complicated to apply in a simple model. Instead, I rely on Chandrasekhar's formula, an approximate result which is only formally valid in a very limited set of situations. I find, however, as many others have previously, that by adjusting the Coulomb logarithm, a free parameter which scales the magnitude of dynamical friction, I get good agreement with the deceleration measured in high-resolution simulations of close encounters between small satellites and disk galaxies. This part of the analytic model thus has a single free parameter, the (halo) Coulomb logarithm, $\ln \Lambda_{\rm h}$, which by comparison with the simulations of Velázquez & White (1999, VW hereafter), should have a value of approximately 2.4.

I include dynamical friction from the disk and from the spherical components of the background separately, both because their dynamics are quite different, and because this permits a heating calculation described in chapter 8. Disk friction is generally less well determined than halo friction, both analytically (because of the complexity of the calculation), and numerically (because it is typically much weaker than halo friction), except for orbits in the disk plane, where other effects may also influence satellite dynamics. Disk friction introduces a second free parameter in the analytic model (although it is of negligible importance for most satellites in a typical halo), namely the disk Coulomb logarithm, $\ln \Lambda_{\rm d}$. Once again, comparison with simulations suggests that $\ln \Lambda_{\rm d} \simeq 0.4$–$0.5$.

In chapter 3, I consider the dynamical effects of mass loss. Subhalos have some spatial extent, and therefore experience differential, or tidal, gravitational forces which may affect their structure. In the simplest case, tidal forces create an envelope around the core of the subhalo, analogous to the Roche lobes in stellar binaries, outside of which material is unbound and no longer follows the core in its orbit. This tidal truncation reduces the mass of a subhalo as it orbits, which in turn reduces the magnitude of the local response it produces in the background, and consequently the drag it experiences from dynamical friction. For massive satellites similar in density to their parent system, the interplay between mass loss and dynamical friction is

complex, and can change the overall evolution of the system radically, so it is crucial to have a good model for mass loss in order to predict the evolution of these systems accurately.

For static systems, that is satellites on unchanging circular orbits, the size of the tidal limit can be calculated approximately and is constant with time. Thus, over time these systems should lose all mass beyond this limit, and reach a steady state. This state is described most succinctly in terms of the ratio of the density of the satellite within its tidal limit to the density of the background system within the orbit of the satellite; in the simplest case, this density ratio has a value of 3. The orbits of subhalos typically seen in numerical simulations are very radial, however, and poorly described by this approximation. An alternate description of mass loss, called the impulse approximation, predicts mass loss in encounters much faster than the internal dynamical time of the satellite. I propose an ad-hoc description of mass loss which interpolates between these two limiting cases, and show that while lacking rigorous justification, it predicts mass-loss rates much more accurately than the tidal limit approximation alone. Formally, this description of mass loss has no free parameters, although the use of the the instantaneous orbital period as the characteristic timescale, while reasonable, is somewhat arbitrary.

Finally, in chapter 4 I discuss some of the complexities of tidal interactions not included in the simple model of mass loss. Generally, tidal forces will increase the internal energy to orbiting satellites, affecting their structure even within the tidal limit. The theory of tidal heating has been developed previously in great detail, to model the dynamics of globular clusters. I show how it can be added to the mass-loss model in a particularly convenient way. The close connection between the two is no accident, but reflects the fact that both effects stem from the same tidal forces, differing only in the characteristic timescale over which they operate. By adding the effects of tidal heating to mass loss and dynamical friction, I show that the analytic model predicts the evolution of numerical satellites to very good accuracy. Tidal

heating introduces a third and final free parameter, the heating efficiency ϵ_h, predicted to be ≥ 2. I find that the value $\epsilon_h = 3$, determined by matching the VW simulations, also produces a very good match to the simulations of Hayashi and Navarro (2001, HN hereafter), despite their use of a very different potential and satellite model.

The three main components of the analytic model are all expressed in terms of global properties of the subhalo, which account for their speed, relative to Fokker-Planck or restricted numerical methods. Unfortunately, it is not obvious how to account for changes in satellite structure within this simple framework. In the last part of chapter 4, I show some results on the internal evolution of subhalos, from simulations by HN. Although these results do not provide a clear explanation for the observed behaviour of the density profile, they do suggest that the simple approximation of self-similar evolution is valid as a first-order description of satellite evolution.

Combining this result with the previous components of the analytic model, I produce a code which can predict the changing position and velocity (or orbital radius, phase and period), as well as the mass, size and peak circular velocity of subhalos as they orbit within a larger system, to an accuracy of 10–20%. This analytic model becomes unreliable only when the satellite is very large or very massive compared to the main system, if it evolves for many orbits, or if it loses most of its mass (more than 80–90%). The model is based on scalar calculations, involving global properties of the satellite, so its computational order is $O(N_{\rm sats}N_{\rm steps})$, where $N_{\rm sats}$ is the number of satellites $N_{\rm steps}$ is the number of timesteps. As a result, it only takes a few minutes on a desktop system to follow the evolution of several thousand satellites over the age of the universe. In this sense, the analytic model is 5,000–10,000 times faster than comparable numerical models, and thus provides a radically different way of studying halo dynamics.

9.4 A Semi-analytic Model of Galaxy Formation

Given initial conditions, the analytic model provides a very fast way of studying the subsequent dynamical evolution of a galaxy halo. Producing detailed and accurate initial conditions is a separate challenge, however. Here too, I rely on a substantial body of previous analytic and numerical work on structure formation. The collapse of dark matter into bound halos has been studied by a generation of numerical modellers, producing agreement on some of the basic properties of halos, including their formation rates, spatial distribution, mean densities, density profiles, spin, flattening or triaxiality, and most recently their substructure, which appears to have its own regular distributions of mass, peak velocity, and position within the large halo. Thus, at least in classical CDM cosmologies, it is possible to produce a detailed and convincing model for the structure of individual halos at any given epoch.

While large-scale simulations now follow the formation of well-resolved halos in considerable numbers, this is a comparatively recent development. Earlier models of galaxy formation in the field therefore relied on analytic estimates of the formation and growth rates of halos. The most developed of these, extended Press-Schechter methods developed in the 1990s, actually predict, in a statistical sense, the full accretion and merger histories of individual halos. By incorporating these statistics into Monte-Carlo codes, a number of groups have developed methods for generating large numbers of representative halos with detailed evolutionary histories, at very little computational expense. While these methods are approximate, and have some intrinsic limitations, they provide a very fast way of generating initial conditions for subsequent studies of galaxy evolution.

In chapter 5, I describe my implementation of the most recent of these 'merger tree' codes. Press-Schechter statistics predict the number of collapsed halos on a given mass scale at any epoch, based on a simple model of the collapse of an isolated density fluctuation. The more recent extension of this method provides estimated growth and

merger rates between halos. Using these extended Press-Schechter statistics, Merger tree codes then determine a possible decomposition of a single present-day halo into its progenitors at any earlier time, as well as the sequence of mergers which produced the final object. Thus, in the context of the analytic model, these codes predict the evolving mass of the main halo since some early time, the original masses of its satellites, and the times at which they merged with the main system, as well as similar information for each satellite, down to some limiting resolutions. Each run produces a unique halo, and typically takes a few minutes of CPU time on a desktop machine.

There are several subtleties involved in using the output from merger trees as the input to the dynamical model developed in part I. I discuss these in chapter 6. First, the properties of the central galaxy in the main system, which can have a substantial effect on satellite evolution, need to be determined. For this I use simple, semi-empirical rules based on the observed appearance of giant galaxies, and inferences about galaxy formation.

A more complex problem is the treatment of substructure within subhalos merging with the main system. While some semi-analytic models ignore this complication, it is clearly unrealistic to assume that in an equal mass merger between two halos, for instance, one contains many small subhalos which survive the merger, whereas the other consists of a single monolithic object, which contributes no substructure to the final merger remnant. In fact, I show that this assumption leads to a systematic underestimate of halo substructure, compared to the results seen in simulations. The converse assumption, namely that subhalos consist entirely of undissolved substructure, is equally incorrect. To model the formation of a halo consistently would require separate dynamical calculations for each subhalo, making the process prohibitively expensive. Instead, I use an approximate treatment of halo dynamics in sub-branches, to determine which substructure is disrupted or absorbed within each subhalo falling into the main halo, and which substructure survives to continue its evolution in the main halo as a separate entity. I show that this 'pruning' technique produces results

in good agreement with numerical simulations.

Suitably pruned merger trees, the set of rules describing the growth of the central galaxy in the main halo, and the analytic model of subhalo dynamics, together constitute a full, semi-analytic model of galaxy formation, which makes the investigation of statistical problems in galaxy dynamics computationally feasible. The last two chapters of the thesis are devoted to two important examples, the origins and distribution of the stellar contents of galactic halos, and the survival of galactic disks.

9.5 The Contents of Galactic Halos

The different stellar populations distributed throughout galaxy halos, in particular the halo of the Milky Way, provide some of the most detailed clues about the process of hierarchical galaxy formation. The first, paradoxical feature of the halos of the Milky Way and Andromeda, the only giant galaxies we can observe in detail, is that they contain few distinct dwarf galaxies, whereas dark matter halos in simulations contain thousands of subhalos with comparable or greater circular velocities. Two possible explanations for the incredible discrepancy between the two (almost two orders of magnitude!) are that star formation in small subhalos is extremely inefficient, or that CDM is wrong in some fundamental way.

I investigate one plausible model for the truncation of star formation in small halos, suggested by Babul & Rees (1992) and others, and considered recently in this context by Bullock *et al.* (2000). In this model, star formation in small objects is suppressed, due to the heat deposited into the gas in these objects by intense background radiation, which arises at the epoch of reionisation. I find that this process produces plausible numbers of dwarf satellites, as established by Bullock *et al.* (2000), but that one or two features of the predicted luminosity function require further explanation. In particular, different dwarf galaxy morphologies, evolutionary histories

and mass-to-light ratios suggest a number of other processes have modified their stellar contents since they fell into the halo of their parent galaxy. Estimating the effects of ram-pressure stripping, collisions and encounters on these objects produces satellite luminosity functions very similar to those of the Milky Way and Andromeda. I also show that the properties of the dark matter halos associated with the dwarf galaxies, specifically their extent, circular velocity and spatial distribution, are consistent with Local Group observations.

Not all dwarf satellites survive the process of galaxy formation; many may be disrupted, spreading their debris throughout the stellar halo of the Milky Way. Observations of the halo have determined its total mass, density profile, age and metallicity. I show that the disruption of dwarf galaxies in the SA model produces a distribution of tidal debris roughly consistent with these observations, and that in particular, most of the material is fairly old. While the details of dwarf disruption are uncertain in the simple analysis of the analytic model, some systems have been disrupted fairly recently in a typical SA stellar halo, so coherent streams of stellar debris, similar to those observed in the Milky Way, may have survived to the present day.

9.6 The Survival of Galactic Disks

Finally, in the last chapter, I address one of the main questions motivating this work; in hierarchical cosmologies, how is it that galactic disks survive heating and disruption by frequent mergers? I consider the effects of two distinct processes which can disturb disks: direct collisions, in which satellites pass through disks, heating them violently or disrupting them, and distant encounters which heat disks slowly over time.

The rate of collisions is particularly interesting, as these events should leave indelible imprints in the form and dynamics of galaxies, and produce observable burst

of star formation or nuclear activity. While collision rates are easy to estimate in the SA model, there is no clear criterion for establishing which of these events will disrupt or heat a disk. After comparing disruption rates derived assuming different criteria, I choose the criteria that encounters between satellites with total masses equal to, or greater than, the central galaxy can be taken to disrupt a galaxy completely, while encounters with satellites with 20% of the mass of the central galaxy will thicken the disk appreciably.

I show that for this criterion (or any similar one), galactic disks are old, even in SCDM. Ages of 10–12 Gyr are typical, and thus the estimated age of the disk of the Milky Way is perfectly reasonable in this cosmology, or in LCDM. On the other hand, if disk disruption is assumed to transfer most of the mass of the disk into a less flattened bulge, then the bulges typically produced in LCDM are more massive than expected. Possible solutions to this problem are that early disks are mainly gaseous and that the disruption of a gaseous disk is less efficient at transforming this gas into stars than inferred from numerical simulations of disk disruption (which consider disks with smaller gas fractions), or that observed disk-to-bulge mass ratios are overestimated. A less likely possibility is that the disk disruption criterion used here is not strict enough, and that disks can survive encounters with massive objects.

The disk ages predicted by the SA model may seem to contradict early work on disk disruption, which suggested it was more common in CDM models in general, and in SCDM in particular. I show however, that several recent estimates of disruption rates point to the general result established here, namely that disks are old on average. After the initial epoch of rapid merging 10 – 13 Gyr ago, mergers between massive subhalos and the central galaxies in Milky Way-type halos became sufficiently rare that the average galaxy has not experienced one since this time. This result stems from the fact that the halo merger rate on these mass scales decreases with time, while size of galactic halos increases, reducing the rate of central mergers, and also from the fact that tidal mass loss and the growth of the central disk galaxy increase

the mass difference between the satellite and the main system during the initial stages of infall.

The SA model also explains the origin of thick disks in the same context. The rate of less disruptive encounters can be estimated, in a similar manner to the disruption rate, using a larger mass ratio for the main criterion. I show that these encounters can produce thick-disk components with $10 - 20\%$ of the disk mass in $\simeq 30\%$ of all systems.

In the remainder of chapter 8, I consider the effects of more gradual heating on galactic disks. First I used the formalism of Tóth & Ostriker to calculate the contribution to disk heating from direct encounters with small satellites. I show that these encounters can heat the disk substantially at early times, although the net effect on mean scale-height of the disk at the present day is negligible, because most of the stellar mass of the disk has built up more recently. I then consider indirect heating produced by distant substructure throughout the halo. I find that halo substructure can also heat the old disk substantially, although once again it is ineffective at late times. Minor encounters or indirect heating may explain the origin of the thick disk and part of the age-velocity dispersion relation, although some of the latter must come from other sources such as heating by spiral arms and giant molecular clouds.

9.7 Future Work

As the scope of the model of galaxy formation developed in this thesis was fairly broad, I have not been able to investigate all of its components in detail. It would be interesting to test each component more thoroughly, using simulations designed specifically for that purpose. In particular, my assumptions about the structural evolution of subhalos, their concentration as a function of mass and redshift, and the dynamics of subgroups falling into a larger halo, all require further confirmation by

numerical results. Given suitable simulations, testing these assumptions should be straightforward.

Some components of the model, while inadequate, will be harder to improve upon. In particular, the merger trees used here have been demonstrated to show systematic mismatches with numerical results. While one main cause of this discrepancy (the assumption that the density threshold for gravitational collapse is independent of mass) has been established, there is no easy way of generating merger trees which correct for this effect.

Within galaxy halos, encounters between subhalos should be numerous, and may drive the evolution of the most massive satellites. While some consequences of subhalo encounters can be investigated approximately using the SA model, a proper treatment of these collisions requires numerical simulations. Similarly, the dynamical evolution of the largest subhalos is poorly represented by the analytic model, and should be determined from numerical simulations.

Finally, there are components that could be added to the current model, increasing its predictive power at the expense of added complexity. In particular, it would be straightforward to add the dissipative components of existing SA models to the one developed here, expanding on the toy model of star formation presented in chapter 7. This might be of interest when considering the detailed evolution of dwarf galaxies and the stellar structure of the disk, but SA studies should probably only serve as a preliminary to numerical work on these problems, given their complexity. Simpler additions to the model could be used to follow the statistics of black hole formation in the nucleus of the galaxy, and to relate merger rates to observed galactic activity.

The SA model presented here deals with individual halos in detail, but ignores structure in the universe on larger scales. Thus, to derive predictions for observed merger and interaction rates in large surveys will require combining it with more traditional semi-analytic models, which apply to whole populations of objects. An

important exception is in studies of cluster dynamics, since cluster galaxies are themselves substructure within a much larger dark halo. Although the model presented here was applied to the problem of galaxy and group formation, by changing the mass scales in the model and removing the central disk, its predictions can be extended to clusters with little effort.

Interesting as it may be as a distinct representation of galaxy or cluster formation, my ultimate goal in developing the semi-analytic model was to provide a fast and simple way of incorporating what we know, or think we know, about structure formation into many other calculations, such as disk heating or satellite evolution. Even today, these calculations are often performed in complete abstraction from their cosmological context – numerical simulations of disk heating start with present-day disks, and studies of satellite disruption consider satellites like those we see today, because there is no simple way to incorporate the changing mass, size and properties of galactic systems into such studies. I hope that with methods like those presented in this thesis, we will soon begin to improve this situation.

References

Aarseth, S. J. 1985, IAU Symp. 113: Dynamics of Star Clusters, 113, 251

Aguilar, L. A. & White, S. D. M. 1986, ApJ, 307, 97

Allen, A. J. & Richstone, D. O. 1988, ApJ, 325, 583

Babul, A. & Rees, M. J. 1992, MNRAS, 255, 346

Bahcall, J. N. & Soneira, R. M. 1980, ApJS, 44, 73

Balbi, A. 2001, *to appear in* Cosmology and Particle Physics, AIP Conf. Proc., eds. J. Garcia-Bellido, R. Durrer and M. Shaposhnikov (astro-ph/0011202)

Barkana, R. & Loeb, A. 1999, ApJ, 523, 54

Barnes, J. E. 1988, ApJ, 331, 699

Barnes, J. E. 1992, ApJ, 393, 484

Barnes, J. E. 1996, ASP Conf. Ser. 92: Formation of the Galactic Halo...Inside and Out, 415

Barnes, J. E. 2001, MNRAS, submitted
(available at http://www.ifa.hawaii.edu/ barnes/fogdimg.html)

Barnes, J. & Efstathiou, G. 1987, ApJ, 319, 575

Barnes, J. E. & Hernquist, L. 1996, ApJ, 471, 115

Baugh, C. M., Cole, S., Frenk, C. S. & Lacey, C. G. 1998, ApJ, 498, 504

Bekenstein, J. D. & Maoz, E. 1992, ApJ, 390, 79

Bekenstein, J. & Milgrom, M. 1984, ApJ, 286, 7

Bekki, K. 1998a, ApJ, 499, 635

Bekki, K. 1998b, A&Ap, 334, 814

Bekki, K. 1998c, ApJL, 502, L133

Bekki, K. 1998d, ApJ, 504, 50

Bekki, K. & Shioya, Y. 1997, ApJL, 478, L17

Bendo, G. J. & Barnes, J. E. 2000, MNRAS, 316, 315

Binggeli, B., Sandage, A. & Tammann, G. A. 1988, ARAA, 26, 509

Binney, J. & Tremaine, S. 1987, Galactic Dynamics (Princeton: Princeton University Press) (BT)

Blitz, L. & Robishaw, T. 2000, ApJ, 541, 675

Bond, J. R., Cole, S., Efstathiou, G. & Kaiser, N. 1991, ApJ, 379, 440

Bontekoe, T. R. & van Albada, T. S. 1987, MNRAS, 224, 349

Bower, R. J. 1991, MNRAS, 248, 332

Bruzual, G. A. 2000, *to appear in* Proceedings of the XI Canary Islands Winter School of Astrophysics on Galaxies at High Redshift, eds I. Perez-Fournon, M. Balcells, and F. Sanchez (astro-ph/0011094)

Bullock, J. S., Kravtsov, A. V. & Weinberg, D. H. 2000, ApJ, 539, 517

Bullock, J. S., Kravtsov, A. V. & Weinberg, D. H. 2001a, ApJ, 548, 33

Bullock, J. S., Kolatt, T. S., Sigad, Y., Somerville, R. S., Kravtsov, A. V., Klypin, A. A., Primack, J. R. & Dekel, A. 2001b, MNRAS, 321, 559

Buser, R. 2000, Science, 287, 69

Carlberg, R. G. & Couchman, H. M. P. 1989, ApJ, 340, 47

Carlberg, R. G. & Hartwick, F. D. A. 1989, ApJ, 345, 196

Carraro, G., Chiosi, C., Girardi, L., Lia, C. 2001, MNRAS, in press (astro-ph/0105026)

Cen, R. 2001, ApJL, 546, L77

Chandrasekhar, S. 1943, ApJ, 97, 255

Chiba, M. & Nath, B. B. 1994, ApJ, 436, 618

Colberg, J. M. & et al. 1998, ASSL Vol. 231: The Evolving Universe, 389

Cole, S., Aragon-Salamanca, A., Frenk, C. S., Navarro, J. F. & Zepf, S. E. 1994, MNRAS, 271, 781

Cole, S. & Lacey, C. 1996, MNRAS, 281, 716

Colpi, M., Mayer, L. & Governato, F. 1999, ApJ, 525, 720

Cora, S. A., Muzzio, J. C. & Vergne, M. M. 1997, MNRAS, 289, 253

Corsini, E. M. et al. 1999, A&Ap, 342, 671

Cretton, N., Naab, T., Rix, H. -. & Burkert, A. 2001, ASP Conf. Ser. 230: Galaxy Disks and Disk Galaxies, 413

Dehnen, W. & Binney, J. 1998, MNRAS, 294, 429

Dekel, A. & Silk, J. 1986, ApJ, 303, 39

Domínguez-Tenreiro, R. & Gómez-Flechoso, M. A. 1998, MNRAS, 294, 465

Drinkwater, M. J., Gregg, M. D., Holman, B. A. & Brown, M. J. I. 2001, MNRAS, 326, 1076

Efstathiou, G. 1992, MNRAS, 256, 43P

Efstathiou, G., Dalton, G. B., Sutherland, W. J. & Maddox, S. J. 1992, MNRAS, 257, 125

Efstathiou, G., Frenk, C. S., White, S. D. M. & Davis, M. 1988, MNRAS, 235, 715

Eke, V. R., Cole, S. & Frenk, C. S. 1996, MNRAS, 282, 263

Eke, V. R., Navarro, J. F. & Steinmetz, M. 2001, ApJ, 554, 114

Fall, S. M. & Efstathiou, G. 1980, MNRAS, 193, 189

Ferguson, A. M. N. & Clarke, C. J. 2001, MNRAS, 325, 781

Forcada-Miró, M. I. 1997, preprint (astro-ph/9712205)

Fuchs, B. 2000, Astronomische Gesellschaft Meeting Abstracts, Abstracts of Talks and Posters presented at the International Conference of the Astronomische Gesellschaft at Heidelberg, March 20-24, 2000, talk #13., 16, 13

Gelb, J. M. & Bertschinger, E. 1994, ApJ, 436, 467

Gerritsen, J. P. E. & de Blok, W. J. G. 1999, A&Ap, 342, 655

Ghigna, S., Moore, B. , Governato, F., Lake, G., Quinn, T. & Stadel, J. 1998, MNRAS, 300, 146

Gilmore, G., King, I. & van der Kruit, P. 1989, Proceedings of the 19th Advanced Course of the Swiss Society of Astronomy and Astrophysics (SSAA), Saas-Fee, Leysin, Vaud, Switzerland, 13-18 March, 1989, Geneva: Observatory, 1989, edited by Buser, Roland

Gilmore, G., Wyse, R. F. G. & Jones, J. B. 1995, AJ, 109, 1095

Gnedin, N. Y. 1996, ApJ, 456, 1

Gnedin, N. Y. 2000, ApJ, 542, 535

Gnedin, O. Y., Hernquist, L. & Ostriker, J. P. 1999, ApJ, 514, 109

Gnedin, O. Y. & Ostriker, J. P. 1997, ApJ, 474, 223

Gnedin, O. Y. & Ostriker, J. P. 1999, ApJ, 513, 626

Governato, F., Babul, A., Quinn, T., Tozzi, P., Baugh, C. M., Katz, N. & Lake, G. 1999, MNRAS, 307, 949

Gramann, M. & Hütsi, G. 2000, MNRAS, 316, 631

Gross, M. A. K., Somerville, R. S., Primack, J. R., Holtzman, J. & Klypin, A. 1998, MNRAS, 301, 81

Gunn, J. E. 1977, ApJ, 218, 592

Gunn, J. E. & Gott, J. R. III 1972, ApJ, 176, 1

Gurzadyan, V. G. & Mazure, A. 2001, New Astronomy, 6, 43

Haiman, Z. & Loeb, A. 1997, ApJ, 483, 21

Haiman, Z. & Menou, K. 2000, ApJ, 531, 42

Hannestad, S. & Scherrer, R. J. 2000, Phys.Rev. D62, 043522

Hayashi, E. 2001, Ph.D. thesis, University of Victoria

Hayashi, E. & Navarro, J. F. 2001, in preparation (HN)

Heisler, J. & White, S. D. M. 1990, MNRAS, 243, 199

Helmi, A., White, S. D. M., de Zeeuw, P. T. & Zhao, H. 1999, Nature, 402, 53

Hernquist, L. 1992, ApJ, 400, 460

Hernquist, L. & Quinn, P. J. 1988, ApJ, 331, 682

Hernquist, L. & Quinn, P. J. 1989, ApJ, 342, 1

Heyl, J. S., Hernquist, L. & Spergel, D. N. 1996, ApJ, 463, 69

Hilker, M. & Richtler, T. 2000, A&Ap, 362, 895

Huang, S. & Carlberg, R. G. 1997, ApJ, 480, 503 (HC)

Ibata, R. A., Gilmore G. & Irwin M. J. 1994, Nature, 370, 194

Ikeuchi, S. 1986, Ap&SS, 118, 509

Innanen, K. A., Harris, W. E. & Webbink, R. F. 1983, AJ, 88, 338

Ipser, J. R. & Semenzato, R. 1985, A&Ap, 149, 408

Jain, B. & Bertschinger, E. 1994, ApJ, 431, 495

Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P. & Yoshida, N. 2001, MNRAS, 321, 372

Johnston, K. V., Zhao, H., Spergel, D. N. & Hernquist, L. 1999, ApJL, 512, L109

Katz, N. 1992, ApJ, 391, 502

Kauffmann, G. 1996, MNRAS, 281, 487

Kauffmann, G., Guiderdoni, B. & White, S. D. M. 1994, MNRAS, 267, 981

Kauffmann, G. & White, S. D. M. 1993, MNRAS, 261, 921

Kauffmann, G., White, S. D. M. & Guiderdoni, B. 1993, MNRAS, 264, 201

Kepner, J. V., Babul, A. & Spergel, D. N. 1997, ApJ, 487, 61

King, I. 1962, AJ, 67, 471

Klypin, A., Gottlöber, S., Kravtsov, A. V. & Khokhlov, A. M. 1999a, ApJ, 516, 530

Klypin, A., Kravtsov, A. V., Valenzuela, O. & Prada, F. 1999b, ApJ, 522, 82

Kormendy, J. 1985, ApJ, 295, 73

Krauss, L. M. 2001, *to appear in* Proceedings of Third International Conference on the Identification of Dark Matter, York, England, Sept 2000 (astro-ph/0102305)

Kundić, T. & Ostriker, J. P. 1995, ApJ, 438, 702

Lacey, C. G. 1984, MNRAS, 208, 687

Lacey, C. & Cole, S. 1993, MNRAS, 262, 627 (LC)

Lacey, C. & Cole, S. 1994, MNRAS, 271, 676

Lacey, C. G. & Ostriker, J. P. 1985, ApJ, 299, 633

Lacey, C. & Silk, J. 1991, ApJ, 381, 14

Lee, Y.-W., Joo, J.-M., Sohn, Y.-J., Rey, S.-C., Lee, H.-C. & Walker, A. R. 1999, Nature, 402, 55

Leitherer, C. et al. 1999, ApJS, 123, 3

Lin, D. N. C. & Faber, S. M. 1983, ApJL, 266, L21

Loveday, J. 1996, MNRAS, 278, 1025

Ma, C. 1996, ApJ, 471, 13

Majewski, S. R., Ostheimer, J. C., Kunkel, W. E. & Patterson, R. J. 2000, AJ, 120, 2550

Maoz, E. 1993, MNRAS, 263, 75

Maraston, C. 1999, ASP Conf. Ser. 163: Star Formation in Early Type Galaxies, 28

Mateo, M. L. 1998, ARAA, 36, 435

Mayer, L., Governato, F., Colpi, M., Moore, B., Quinn, T., Wadsley, J., Stadel, J. & Lake, G. 2001, ApJL, 547, L123

Meylan, G., Sarajedini, A., Jablonka, P., Djorgovski, S. G., Bridges, T. & Rich, R. M. 2001, AJ, 122, 830

Mihos, J. C. 2001, ASP Conf. Ser. 230: Galaxy Disks and Disk Galaxies, 491

Mihos, J. C. & Hernquist, L. 1994, ApJL, 431, L9

Mihos, J. C. & Hernquist, L. 1996, ApJ, 464, 641

Milgrom, M. 1986, ApJ, 302, 617

Moore, B. 2000, Constructing the Universe with Clusters of Galaxies, IAP 2000 meeting, Paris, France, July 2000, Florence Durret & Daniel Gerbal Eds., E71

Moore, B., Ghigna, S., Governato, F., Lake, G., Quinn, T., Stadel, J. & Tozzi, P. 1999, ApJL, 524, L19

Moore, B., Governato, F., Quinn, T., Stadel, J. & Lake, G. 1998, ApJL, 499, L5

Moore, B., Katz, N. & Lake, G. 1996, ApJ, 457, 455

Naab, T., Burkert, A. & Hernquist, L. 1999, ApJL, 523, L133

Nagashima, M., Gouda, N. & Sugiura, N. 1999, MNRAS, 305, 449

Navarro, J., Frenk, C. S., White, S. D. M. 1994, MNRAS, 267, L1

Navarro, J., Frenk, C. S., White, S. D. M. 1996, ApJ, 462, 563

Navarro, J., Frenk, C. S., White, S. D. M. 1996, ApJ, 490, 493

Navarro, J. F. & Steinmetz, M. 2000, ApJ, 538, 477

Negroponte, J. & White, S. D. M. 1983, MNRAS, 205, 1009

Oh, K. S., Lin, D. N. C. & Aarseth, S. J. 1995, ApJ, 442, 142

Ostriker, J. P. 1990, ASP Conf. Ser. 10: Evolution of the Universe of Galaxies, 25

Ostriker, J. P., Spitzer, L. J. & Chevalier, R. A. 1972, ApJL, 176, L51

Ostriker, J. P. & Tremaine, S. D. 1975, ApJL, 202, L113

Padmanabhan, T. 1993, Structure Formation in the Universe (Cambridge: Cambridge University Press)

Peebles, P. J. E. 1980, The Large-scale Structure of the Universe (Princeton: Princeton University Press)

Pei, Y. C. 1995, ApJ, 438, 623

Petrosian, A. R., Gurzadyan, V. G., Hendry, M. A. & Nikoghossian, E. H. 1998, Astrophysics, 41, 32

Pignatelli, E. & Galletta, C. 1997, ASP Conf. Ser. 117: Dark and Visible Matter in Galaxies and Cosmological Implications, 136

Plionis, M. 2001, *to appear in* Galaxy Clusters and the High Redshift Universe Observed in X-rays, XXIth Moriond Astrophysics Meeting (March 2001), Eds. Doris Neumann et al (astro-ph/0105522)

Press, W. H. & Schechter, P. 1974, ApJ, 187, 425

Prochaska, J. X., Naumov, S. O., Carney, B. W., McWilliam, A. & Wolfe, A. M. 2000, AJ, 120, 2513

Quillen, A. C. & Garnett, D. R. 2001, ASP Conf. Ser. 230: Galaxy Disks and Disk Galaxies, 87

Quinn, P. J., Hernquist, L. & Fullagar, D. P. 1993, ApJ, 403, 74

Quinn, P. J. & Goodman, J. 1986, ApJ, 309, 472

Quinn, T., Katz, N., Stadel, J. & Lake, G. 1997, ApJ, submitted (astro-ph/9710043)

Rees, M. J. 1986, MNRAS, 218, 25P

Rees, M. J. & Ostriker, J. P. 1977, MNRAS, 179, 541

Rocha-Pinto, H. J., Scalo, J., Maciel, W. J. & Flynn, C. 2000, ApJL, 531, L115

Ryden, B. S. 1988, ApJ, 329, 589

Saha, P., Stadel, J. & Tremaine, S. 1997, AJ, 114, 409

Schmalzing, J., Sommer-Larsen, J. & Goetz, M. 2001, ApJL, submitted (astro-ph/0010063)

Shaw, M. A. & Gilmore, G. 1990, MNRAS, 242, 59

Sheth, R. K., Mo, H. J. & Tormen, G. 2001, MNRAS, 323, 1

Sheth, R. K. & Tormen, G. 1999, MNRAS, 308, 119

Silk, J. 1977, ApJ, 211, 638

Silk, J., Wyse, R. F. G. & Shields, G. A. 1987, ApJL, 322, L59

Simien, F. & de Vaucouleurs, G. 1986, ApJ, 302, 564

Smoot, G. F. et al. 1991, ApJL, 371, L1

Somerville, R. S. & Kolatt, T. S. 1999, MNRAS, 305, 1 (SK)

Somerville, R. S. & Primack, J. R. 1999, MNRAS, 310, 1087

Spergel, D. N. & Steinhardt, P. J. 2000, Physical Review Letters, 84, 3760

Spitzer, L. J. 1958, ApJ, 127, 17

Spitzer, L. J. 1987, Dynamical Evolution of Globular Clusters (Princeton: Princeton University Press)

Springel, V. 2000, MNRAS, 312, 859

Springel, V., White, S. D. M., Tormen, G. & Kauffmann, G. 2000, MNRAS, submitted (astro-ph/0012055)

Steinmetz, M. & Bartelmann, M. 1995, MNRAS, 272, 570

Steinmetz, M. & Buchner, S. 1995, LNP Vol. 463: Galaxies in the Young Universe, 215

Steinmetz, M. & Navarro, J. F. 1999, ApJ, 513, 555

Tantalo, R., Chiosi, C., Bressan, A. & Fagotto, F. 1996, A&Ap, 311, 361

Taylor, J. E. & Babul, A. 2001, ApJ, in press (astro-ph/0012305)

Taylor, J. E. & Navarro, J. F. 2001, ApJ, in press (astro-ph/0104002)

Toomre, A. & Toomre, J. 1972, ApJ, 178, 623

Tormen, G. 1997, MNRAS, 290, 411

Tormen, G. 1998, MNRAS, 297, 648

Tóth, G. & Ostriker, J. P. 1992, ApJ, 389, 5 (TO)

Tozzi, P. & Governato, F. 1998, ASP Conf. Ser. 146: The Young Universe: Galaxy Formation and Evolution at Intermediate and High Redshift, 461

Tremaine, S. D. 1976, ApJ, 203, 72

Tremaine, S. 1981, Structure and Evolution of Normal Galaxies, 67

Unavane, M., Wyse, F. G. R. & Gilmore, G. 1996, MNRAS, 278, 727

van den Bergh, S. 2000, PASP, 112, 529

van den Bosch, F. C. 2001, MNRAS, submitted (astro-ph/0105158)

van den Bosch, F. C., Lewis, G. F., Lake, G. & Stadel, J. 1999, ApJ, 515, 50

van der Kruit, P. C. & Searle, L. 1982, A&Ap, 110, 61

van Kampen, E. 2000, MNRAS, submitted (astro-ph/0008453)

Vega Beltrán, J. C., Pizzella, A., Corsini, E. M., Funes, J. G., Zeilinger, W. W., Beckman, J. E. & Bertola, F. 2001, A&Ap, 374, 394

Velázquez, H. & White, S. D. M. 1999, MNRAS, 304, 254

Viana, P. T. P. & Liddle, A. R. 1996, MNRAS, 281, 323.

Walker, I. R., Mihos, J. C. & Hernquist, L. 1996, ApJ, 460, 121 (WMH)

Warren, M. S., Quinn, P. J., Salmon, J. K. & Zurek, W. H. 1992, ApJ, 399, 405

Weinberg, M. D. 1986, ApJ, 300, 93

White, S. D. M. 1976, MNRAS, 174, 467

White, S. D. M., Efstathiou, G. & Frenk, C. S. 1993, MNRAS, 262, 1023

White, S. D. M. & Frenk, C. S. 1991, ApJ, 379, 52

White, S. D. M., Frenk, C. S. & Davis, M. 1983, ApJL, 274, L1

White, S. D. M. & Rees, M. J. 1978, MNRAS, 183, 341

Zaritsky, D. 1999, ASP Conf. Ser. 165: The Third Stromlo Symposium: The Galactic Halo, 34

Zwaan, M. A. 2001, MNRAS, 325, 1142

Vita

Surname:	Taylor
Given Names:	James Erskine
Place of Birth:	Helsinki, Finland

Educational Institutions Attended:

University of Toronto, Toronto, Ontario	1988–1994
University of Victoria, Victoria, B.C.	1996–2001

Diplomas & Degrees Awarded:

Bachelor of Science, University of Toronto	1993
Master of Science, University of Toronto	1994

Partial Copyright License

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of thesis:

The Formation and Survival of Disk Galaxies

Author:

James E. Taylor

August 30, 2001