1

Pieces of the Galaxy Formation Puzzle: Where do Black Holes Fit In?

RACHEL S. SOMERVILLE
University of Michigan

Abstract

Over the past twenty years we have developed a fairly credible story of how galaxies form in the context of the hierarchical structure formation paradigm. A major challenge for the coming years is to understand the co-evolution of galaxies and the super-massive black holes that we now believe inhabit most of them. I discuss the status of our understanding of the main processes that shape galaxies, comparing where possible the results of the complementary techniques of numerical hydrodynamics and semi-analytic modeling. I suggest that recent progress in understanding the connection of the structural properties (matter density and angular momentum profiles) of dark matter halos will help us to understand the inner structure of galaxies and hence to better model black hole formation and fueling. I also discuss some of the outstanding problems in galaxy formation, and how including the formation of supermassive black holes might help alleviate them.

1.1 Introduction

In the modern hierarchical paradigm of structure formation, galaxies are molded by a complex web of interconnected processes. They are harbored within dark matter halos, which in turn are shaped by their larger scale environment and their history. They receive fuel for star formation via radiative cooling of hot gas that is trapped within the potential wells of these halos. By some mysterious process, this cold gas is eventually transformed into stars, and the most massive of these stars return heavy elements and energy to the interstellar medium (ISM) and perhaps to the hot gas halo. In some cases, supernovae may even drive large scale winds, which spew material far beyond the potential well of the galaxy, into the intergalactic medium, polluting the material that will collapse to form future generations of galaxies. While this basic picture forms a plausible story of how galaxies form and what determines their properties, and detailed calculations based upon it have yielded some non-trivial successes in predicting or interpreting various galaxy observations, there are some indications that there are some important pieces missing in our understanding. It is the growing realization that perhaps we need to understand the *co-evolution* of supermassive black holes (SMBH) and galaxies in order to fill in some of these missing pieces that has inspired this meeting, and that will no doubt be the subject of much thought in the near future.

In this article, I review the status of recent attempts to model these key processes, making wherever possible explicit comparisons between the complementary methods of "semi-analytic" modeling and numerical (or N-body) hydrodynamic simulations. While to date there have been very few attempts to explicitly model the co-evolution of SMBH and galaxies, I will attempt to highlight aspects of our understanding of galaxy formation which might help us to understand the formation of SMBH, and aspects of SMBH formation and the associated AGN phase that might impact our understanding of galaxy formation.

1.2 The Internal Structure of Dark Matter Halos

1.2.1 Halo Density Profiles and Mass Accretion History

Over the past five to ten years, the vast improvement in resolution and dynamic range accessible to N-body techniques with modern supercomputers has led to a characterization of the internal structure of dark matter (DM) halos that form in the Cold Dark Matter (CDM) paradigm. It is a now familiar result that the radial density profiles of DM halos may be well described by a universal function, which has come to be known as the "Navarro-Frenk-White" (NFW) profile (Navarro, Frenk & White 1997). The NFW profile allows halos (assumed to be spherically symmetric) to be characterized by two parameters, for example, a total mass and a characteristic radius or density. The characteristic radius or density may also be expressed as a *concentration*, which represents how much of the halo's mass lies near the center vs. the outskirts. While early work (Navarro et al. 1997) emphasized the correlation between these two parameters (lower mass halos are more concentrated), studies of larger, statistically complete samples of halos have revealed that there is also a significant scatter in halo concentration at fixed mass (Bullock et al. 2001a).

The origin of the correlation between concentration and mass, and the scatter at fixed mass, is now understood as being connected to the *mass accretion history* of the halo. By studying *structural merger trees*, in which the merger histories of dark matter halos were traced through a high-resolution N-body simulation in conjunction with the halo structural properties, Wechsler et al. (2002; W02) showed that early-forming halos have high concentrations, while late forming halos have lower concentrations. This accounts for the mass-concentration correlation, as larger mass halos form later on average than smaller mass halos. Moreover, W02 presented a recipe for modeling halo concentration based on semi-analytic merger trees, and showed that the scatter in concentration at fixed mass seen in the N-body simulations can be accounted for by the spread in formation histories for halos of a given mass. This recipe will also allow future semi-analytic models to properly include the correlations between merger history and halo structure, which will probably have important implications for predicted galaxy properties.

1.2.2 Halo Spin and Merger History

Another fundamental property of a dark matter halo is its angular momentum, often characterized by the dimensionless spin parameter, λ (Peebles 1969). It is well established that the distribution function of spins for DM halos formed in CDM simulations has the form of a log-normal (Barnes & Efstathiou 1987), and

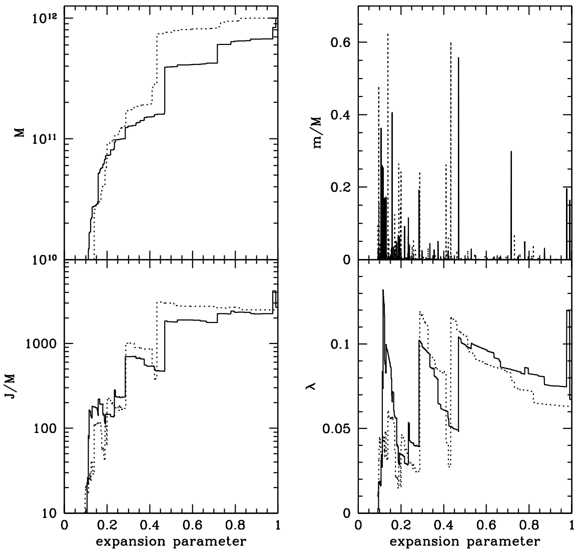

Fig. 1.1. Left panels: the build-up of mass (top) and specific angular momentum (bottom) for two different halos (solid and dotted) as a function of expansion factor $a = (1+z)^{-1}$. Right panels: Mass ratios of merging satellites (top) and spin parameter (bottom) as a function of a (reproduced from Vitvitska et al. (2002).

is insensitive to cosmic epoch, halo mass or environment, and cosmology or power spectrum (Lemson & Kauffmann 1999). The origin of this angular momentum is generally attributed to tidal torques from the large-scale density field, experienced before turnaround, while the proto-halo perturbations are in the linear regime (Peebles 1969; Doroshkevich 1970; White 1984). This picture, combined with the invariance of the *distribution* of spins with cosmic epoch, may have led to the impression that spin is a property that remains invariant over the lifetime of a halo. However, it appears that halos acquire angular momentum in *mergers*, via transfer of orbital angular momentum to internal spin. Because minor mergers (i.e., events in which one member of the merging pair is much smaller in mass than the other) are frequent and more or less isotropic, the angular momentum from these events mostly cancels out. However, major mergers (in which the lumps have comparable masses) are rare and therefore tend to result in a significant net gain of angular momentum (depending, of course, on the impact parameter of the merger). This is illustrated, based on the same structural merger trees discussed above, in figure 1.1, and has obvious significance for our understanding of the structure of galactic disks (discussed in more detail below). Based on these ideas, Vitvitska et al. (2002) and Maller, Dekel & Somerville (2002) proposed analytic recipes that could be included in semi-analytic merger trees to model halo spin, and showed that such a scheme reproduces the log-normal distribution obtained in the N-body simulations. Maller et al. (2002) further showed that this approach could reproduce the known insensitivity of halo spin to mass, redshift, and cosmology, as well as the correlation with merger history noted above.

1.3 Formation of Galactic Disks

The basic picture of disk formation in the context of CDM was outlined by Fall & Efstathiou (1980) and White & Rees (1978), which incorporated the earlier work of Mestel (1963), Crampin & Hoyle (1964), and others. The theory has been further developed in many other works (e.g. Blumenthal et al. 1986; Flores et al. 1993; Dalcanton, Spergel, & Summers 1997; Mo, Mao & White 1998; van den Bosch 2000; Avila-Reese & Firmani 2000; Klypin, Zhao, & Somerville 2002). In this picture, halos (both dark matter and diffuse gas) obtain angular momentum from tidal torques as discussed above, and this gas then conserves its angular momentum as it collapses to form a disk. This leads to a direct relationship between the angular momentum (or spin) of the halo and the size (scale length) of the disk. The analytic work showed that as long as most of the initial angular momentum is conserved during the process of disk formation, then disk galaxies are predicted to have sizes consistent with observed galaxies at $z = 0$. Put another way, if this picture is correct, then somehow cosmological tidal torques seem to impart just enough spin to the proto-galactic gas to make galaxies with about the right amount of specific angular momentum — a rather profound result.

1.3.1 The Angular Momentum Catastrophe

Unfortunately, when we try simulate disk formation in a CDM universe directly using numerical techniques the results are not so successful. When cooling is included in hydrodynamic N-body codes, coherently rotating, disk-like structures are formed — but in the first generation of simulations to attempt this calculation, the disks were deficient in specific angular momentum by about an order of magnitude compared to observed disk galaxies (Navarro & Benz 1991; Navarro & White 1994; Sommer-Larsen, Gelato & Vedel 1999; Navarro & Steinmetz 2000). Evidently, the idealized picture of disk formation outlined above does not describe what actually happens in the hydro simulations, where gas efficiently loses most of its angular momentum during cooling and collapse, resulting in disks that are too compact. This result has come to be known as the "angular momentum catastrophe".

The angular momentum catastrophe has been linked to the "bottom-up" nature of structure formation in CDM, and the very efficient early cooling in small proto-galactic clumps. Once gas is locked up in these small clumps, it tends to lose angular momentum very efficiently via dynamical friction. Indeed, it seems that if one can somehow delay cooling, the angular momentum catastrophe is much alleviated. This has been demonstrated in a variety of ways: by simply switching cooling on by hand only at a relatively late epoch (Weil, Eke, & Efstathiou 1998; Eke, Efstathiou, & Wright 2000), by implementing an inefficient star formation algorithm (Sáiz et al. 2001), by adopting an alternative cosmology with late structure formation such as Warm Dark Matter (Sommer-Larsen & Dolgov 2001) or by including strong supernova feedback (Thacker & Couchman 2001). However, it is known that our Galaxy and M31 have very old stars even in the outer parts of their disks (e.g. Ferguson &

Johnson 2001) — delaying cooling and star formation may lead to conflict with this observational fact.

It also seems that recent higher resolution simulations may suffer a less severe catastrophe than their predecessors, suggesting that insufficient resolution may remain a concern (Governato et al. 2002). A recent study by Abadi et al. (2003) suggests that perhaps the problem is not with the properties of the disks themselves, but that galaxies are too spheroid-dominated. They argue that *taken in isolation*, i.e. with the contribution of the spheroid removed, their simulated disk is not significantly deficient in angular momentum for its circular velocity or luminosity (they present results for a single galaxy only, which was selected from a larger, lower resolution cosmological box and re-simulated at high resolution). However, the galaxy they analyzed has half its stellar mass in a bulge, and resembles a very early type spiral or lenticular galaxy. Because these simulations are computationally very expensive, it is difficult to produce statistically complete samples of galaxies, and thus to make general statements about populations. These results suggest, however, that the problem may be one of the morphological mix in the simulations: too much stellar mass ends up in low-angular momentum spheroids.

To summarize, there are at least three rather different ideas prevalent in the community regarding the root cause of the angular momentum catastrophe:

- resolution or other numerical effects
- too much early merging, perhaps reflecting an excess of small scale power in the current "standard Λ" cosmological model, resulting in a too-spheroid-dominated population
- lack of understanding of astrophysical processes such as star formation and feedback

Note that the approach needed to solve the problem will be quite different depending on which of these is the most important effect. In the first case, we just need to wait for faster computers, while in the second case, we need to modify the assumed cosmology or power spectrum, and in the third case, we need to modify the treatment of star formation and/or feedback. Of course, it is entirely possible that a combination of more than one, or even of all three, of these effects is contributing to the problem. More detailed analytic and semi-analytic studies may help to disentangle these effects (e.g. Maller & Dekel 2002).

1.3.2 The Angular Momentum Profile Mismatch

The more detailed analytic models run into problems of their own, however. In order to make a connection between the structural properties of the halo and those of the galaxy that forms within it, we must know either the functional form of the initial (pre-collapse) *radial distribution* of angular momentum in the halo, *or* the form of the final (post-collapse) radial profile of the disk. In the absence of detailed knowledge of the angular momentum profiles of halos (which require very high resolution simulations to determine numerically), it became a common practice to assume that the post-collapse gas surface density profile has an exponential form (e.g. Blumenthal et al. 1986; Flores et al. 1993; Mo et al. 1998; Klypin et al. 2002). This is generally justified by arguing that the radial surface brightness profiles of observed stellar disks are typically well approximated by exponentials over most of their extent.

Recent studies have revealed that the radial angular momentum profile of material within dark matter halos, like the matter density profiles discussed earlier, have a universal form when scaled appropriately. The specific angular momentum as a function of radius $j(r)$ is found to be well fit by a power law form: $j(r) \propto r^{\alpha}$ with $\alpha = 1.1 \pm 0.3$ (Bullock et al. 2001b; see also van den Bosch et al. 2002). The value of the profile slope α has a significant scatter, and obeys a Gaussian distribution. With the usual assumption that the angular momentum is well-aligned, that mass shells do not cross and angular momentum is conserved during collapse, we can now compute the resulting radial profile of a disk that forms in such a halo — and it is not an exponential (Bullock et al. 2001b, van den Bosch 2001; 2002). There is a "cusp" of excess low angular momentum material in the inner part, and a "tail" of high angular momentum material in the outer part of the disk. This is what I refer to as the "angular momentum profile mismatch".

It has been proposed that the low-angular material may form a bulge, while the high-angular material may be inefficient at forming stars and remain gaseous. This does not work, of course, to explain bulgeless galaxies with pure exponential profiles, which while rare, do exist. An alternate explanation is that disks are not born with exponential profiles, but evolve into them via secular processes (Lin & Pringle 1987; Slyz, Devriendt & Silk 2002), but it is probably difficult for secular evolution to move low-angular momentum material *outwards*, to larger radii in the disk — perhaps this material has to be removed from the galaxy entirely.

What does any of this have to do with black holes, since black hole properties seem to correlate with spheroid, not disk, properties? First, mergers efficiently strip angular momentum from the gas and stars in a galaxy, and angular momentum loss on kpc scales is the first step in the "feeding chain" that funnels material to the accretion disk around the black hole, as discussed by K. Wada in these proceedings. If the angular momentum catastrophe is a symptom of a deeper malady, that the predominant CDM paradigm predicts too much merging, then calculations of black hole growth via mergers (e.g. Volonteri, Haardt & Madau 2003), quasar merger models such as those discussed by M. Haehnelt (these proceedings), and all related quantities will be in error. Second, if the solution to the catastrophe is strong feedback, which delays cooling until the merger rate has decreased, then this process, which regulates spheroid growth, must also regulate black hole growth in order to maintain the observed link between the two. This *may* be a key to understanding the black hole-spheroid scaling relations. Alternatively, perhaps the direction of the causality is reversed: suppose that after a merger, gas is efficiently funneled into the center of the galaxy, not only producing a spheroid but also feeding the central black hole and producing AGN activity. In turn, when the accretion-powered jets turn on, they drive the remaining gas out to large distances, where it may take a very long time to cool and fall back in. In this case, the key to solving the angular momentum catastrophe for galactic disks may lie in the observed connection between black holes and spheroids. Third, if we want to partially solve the angular momentum profile mismatch by converting the low-j material in disks into spheroids, then we have to come to grips with two separate mechanisms for forming spheroids (merging and secular evolution). Observational evidence that there are two different populations of spheroids is already mounting (see the paper by M. Carollo, these proceedings, and references therein). If this is the case, and if it is also the case that the "pseudo-bulges" produced via secular evolution have the same scaling relations as the *bona fide* ones produced by mergers (see papers by J. Kormendy and D. Richstone in these proceedings), then we need to understand *why*.

1.4 Radiative Cooling

Gas in the low-density IGM is cold and, after the epoch of reionization, photoionized. This gas may be heated to higher temperature (a few $\sim 10^6$ K) by shocks during gravitational collapse. Thus the gas in collapsed structures such as halos and even filaments is hot enough and dense enough to cool efficiently by atomic line cooling. This is the dominant process in producing the fuel for star formation in galaxies.

1.4.1 SAMs vs. hydro

Numerical techniques such as Smoothed Particle Hydrodynamics (SPH) or grid-based hydrodynamic codes treat the processes of gravitational collapse in an expanding Universe with a CDM power spectrum and the shock-heating and radiative cooling of gas described above in order to track the rate at which gas cools and becomes available for star formation. In the absence of complications such as heavy element pollution and non-gravitational heating, the physics of these processes are well-understood and the main uncertainties arise from numerical issues. An alternative approach is to use a simple spherical model (as proposed by White & Frenk 1991) applied within a semi-analytic halo merger tree to estimate the bulk rate of cooling, as is the common practice in semi-analytic models of galaxy formation (e.g. Kauffmann, White & Guiderdoni 1993; Cole et al. 1994; Kauffmann et al. 1998; Somerville & Primack 1999; Cole et al. 2000). This model relies on several assumptions:

- halos are spherical, have no substructure, and have a prescribed density profile (typically isothermal or NFW)
- all gas is shock heated to the virial temperature of the halo upon collapse
- the hot halo gas is in collisional equilibrium
- the cooling time is equal to the time needed for the gas to radiate all its thermal energy at the rate given by the radiative cooling function, for gas of the specified density and temperature.

Because the density decreases with increasing radius, this model predicts that the halo cools from the inside out. In practice, for a given density profile one can calculate a "cooling radius" (see e.g. Somerville & Primack 1999), within which the gas is assumed to be "cold", and outside of which it is assumed to be "hot". While this picture is obviously rather artificial, and several if not all of the assumptions listed above are known to be violated in the numerical simulations (for example, most of the cooling occurs in filaments, and gas is never shock-heated to the virial temperature of the halo), there is apparently quite good agreement between the results of the hydrodynamic simulations and the semi-analytic models, at least for bulk properties such as the cold gas fraction as a function of halo mass (Benson et al. 2001; Helly et al. 2002; Yoshida et al. 2002)

1.4.2 Problems: overcooling, cooling flows, and entropy floors

Unfortunately, though not surprisingly, there are several indications that reality is more complicated than the simple case described above, which we can understand and simulate readily. There are three separate but interconnected problems related to cooling, all of which hint at the need for an additional source of non-gravitational heating.

It was pointed out in early explorations of galaxy formation within the CDM paradigm (White & Rees 1978; White & Frenk 1991) that in the absence of some braking process, far too much gas is able to cool, leading to an overly large fraction of baryons in the form of stars or cold gas today. For a recent update on both theoretical and observational aspects of this problem, which is often referred to as the "overcooling problem", see Balogh et al. (2001). The overcooling problem is one compelling indication of the need for a process that either prevents gas from cooling as efficiently as expected, or that heats it up again. Supernova (SN) feedback is generally invoked to do the job, although it is not clear that the energetics of SN driven winds are up to the task (more discussion of this question follows). Another candidate, which has received less attention, is heating by AGN.

A second problem, long suspected but only recently made compelling, is the cooling flow conundrum (see the recent review by Fabian 2002). From arguments like those presented above, cooling times in cluster cores are expected to be short, and so one expects a cooling flow to form. However, XMM-Newton spectra show no evidence that cooling below 1-2 keV is taking place (e.g. Peterson et al. 2001; 2002; Kaastra et al. 2001). Chandra images of these clusters show holes coincident with radio lobes and cold fronts, indicating that these systems are complex (Fabian 2002 and references therein). The most popular solutions may be divided into two broad categories: an additional heating source, such as AGN or substructure, or heating by thermal conduction. In some (but not all) of the systems, there is direct evidence for AGN heating (e.g. observed radio jets), and the energetics are plausible, but the heating must occur over very large scales, which is a bit of a puzzle. Thermal conduction works by using the large reservoir of hotter gas in the outskirts of the cluster to heat the gas in the core. It has been considered reasonable to neglect conduction in clusters because of the conventional wisdom that magnetic fields strongly suppress conduction perpendicular to the field. Recently, however, this idea has been re-examined, and it has been pointed out that in a turbulent magnetized plasma, the field lines become tangled, allowing for more efficient conduction (Narayan & Medvedev 2001). The current situation seems to be that for some clusters, AGN heating seems to be the most plausible mechanism, while in others, conduction appears likely to be able to do the job (Fabian, Voigt, & Morris 2002; Voigt et al. 2002), suggesting that both may play a role.

The third problem is in understanding the observed scaling relations between X-ray properties of clusters, such as the X-ray luminosity-temperature (L_X–T_X) relation. The observed relation deviates from the self-similar behavior expected from gravitational heating and a constant baryon fraction, becoming steeper on the low-T end (Arnaud & Evrard 1999; Markevitch 1998; Allen & Fabian 1998). This may be expressed as the presence of an entropy "floor", a minimum entropy that is negligible in larger systems but dominates in smaller, group-sized systems (Ponman, Cannon, & Navarro 1998; Lloyd-Davies, Ponman & Cannon 2000). The question is: what produces this entropy floor? Again, feedback from SN and/or AGN are obvious candidates. It has been argued (e.g. Wu, Fabian & Nulsen 2000; Kravtsov & Yepes 2000) that the energy produced by SN is insufficient, but this is difficult to show conclusively as it depends on the density and temperature of the ICM or proto-ICM at the time at which the energy is injected. Nath & Roychowdhury (2002) show, based on a semi-analytic calculation, that the energy injected by radio jets and BAL quasars brings the predicted L_X–T_X relation into agreement with observations. See the paper by M. Begelman (these proceedings) more a discussion of much more detailed models of the impact of AGN on the cluster entropy problem.

I have discussed these three problems in some detail because all of them point out holes in our understanding of the processes that shape galaxy formation. Overcooling in clusters remains a problem in all existing semi-analytic models and hydrodynamic simulations. It is generally "cured" in semi-analytic models by invoking an ad hoc means of suppressing cooling in large halos. It is possible, even likely, that the energy released during the fueling of supermassive black holes in galaxies plays an important role in solving one or even all of these problems.

1.5 Star Formation and Feedback

1.5.1 Star Formation

Star formation in cosmological simulations remains a bit like that Far Side cartoon where, in the midst of the calculation on the blackboard it is written "and then, a miracle happens...". Once gas has cooled down to about 10^4 K, it is turned into stars with an efficiency that is determined by some sort of empirical/phenomenological recipe. In the approach most commonly used in cosmological hydrodynamic simulations, star formation is assumed to occur when some combination of the following conditions is satisfied:

- overdensity greater than threshold value
- temperature less than threshold value
- physical density greater than threshold value
- convergent flow ($\nabla \cdot v < 0$)
- Jeans unstable

and the star formation timescale is then generally assumed to be proportional to the dynamical time or the minimum of the cooling and dynamical times. An alternative method is to impose a relationship between star formation rate and local gas density, i.e. a Schmidt law $\dot{\rho}_* \propto \rho_{gas}^N$ with $N \sim 1.5$, based on the empirical relations described by e.g. Kennicutt (1989; 1998). A very nice comparison of different star formation recipes in SPH simulations has recently been presented by Kay et al. (2002). This work showed that the fraction of baryons in stars at $z = 0$ may vary by up to a factor of two for the range of recipes and parameter values appearing in the literature.

It is clear that to make real progress, we must grapple with the multi-phase nature of the ISM and with exchange between different phases; e.g., molecular hydrogen clouds form out of cold, dense, gas, while subsequent star formation may evaporate the surrounding clouds and return material to the warm, low density ISM. There have been a few recent attempts to include a multi-phase star formation algorithm in a hydrodynamical code (e.g. Yepes et al. 1997; Springel & Hernquist 2002). However, at some level, the results still rely unknown free parameters.

In semi-analytic codes, only gross information at best is available about the structure of the galaxy and the density of the cold gas, and star formation must therefore be treated even more schematically than in the numerical simulations. Typically, the star formation rate is modeled using an equation of the general form:

$$\dot{\rho}_* = \frac{m_{\rm cold}}{\tau_*} , \tag{1.1}$$

where $m_{\rm cold}$ is the mass of cold gas in the galaxy (Kauffmann, White & Guiderdoni; Cole et al. 1994; Kauffmann et al. 1998; Somerville & Primack 1999; Cole et al. 2000). A common choice for the star formation timescale $\tau_* \propto t_{\rm dyn}$, with $t_{\rm dyn}$ the dynamical time of the disk. This then resembles the Kennicutt Law and is also similar to the recipes often used in the hydro codes. However, the typical $t_{\rm dyn}$ (as well as $t_{\rm dyn}$ for a halo of a given mass) decreases rapidly with increasing redshift, because of the higher density of objects that collapsed at high redshift, when the background density was also much higher than today. This implies that the characteristic timescale for star formation is much shorter at high redshift than today. On the face of it, this appears to be in agreement with observations, but it turns out that if the Kennicutt-style scaling remains unchanging with cosmic time, star formation is *so efficient* in the very dense galaxies at high redshift that, despite higher cooling rates, nearly all available gas is consumed very rapidly. This has a number of bad effects. It leaves insufficient cold gas lying around at $z \sim 3$ to account for the observed number of Damped Lyman-α systems (Somerville, Primack & Faber 2001). The brightest Lyman-break galaxies at $z \gtrsim 2$ have star formation rates of several hundred solar masses per year, when dust extinction is accounted for (Adelberger & Steidel 2000), while the objects detected in the sub-mm by SCUBA, believed to be at similar redshifts, have implied SFR of thousands of solar masses per year (Blain et al. 2002 and references therein). It is very difficult to understand how to produce these extreme objects unless large reservoirs of cold gas are available at high redshift. With the Kennicutt-style star formation recipe, galaxy mergers tend to be fairly gas-poor at all redshifts, making it difficult to account for the steep decline of the number density of bright quasars at $z \lesssim 2$ in a model in which black hole fueling is triggered by mergers (Kauffmann & Haehnelt 2000). A modified star formation recipe, in which the star formation timescale in "normal" (i.e., non-starburst) galaxies remains *constant* with redshift, while a more efficient starburst mode is triggered by galaxy mergers, is much more successful in reproducing all of these observations (Somerville et al. 2001; Kauffmann & Haehnelt 2000). However, this implies that the empirical scaling relations that guide our understanding of star formation at low redshift are not invariant in cosmic time — if correct, quite an important conclusion.

Given the different levels of detail in the implementation of star formation algorithms in the numerical hydrodynamic simulations and in the semi-analytic models,

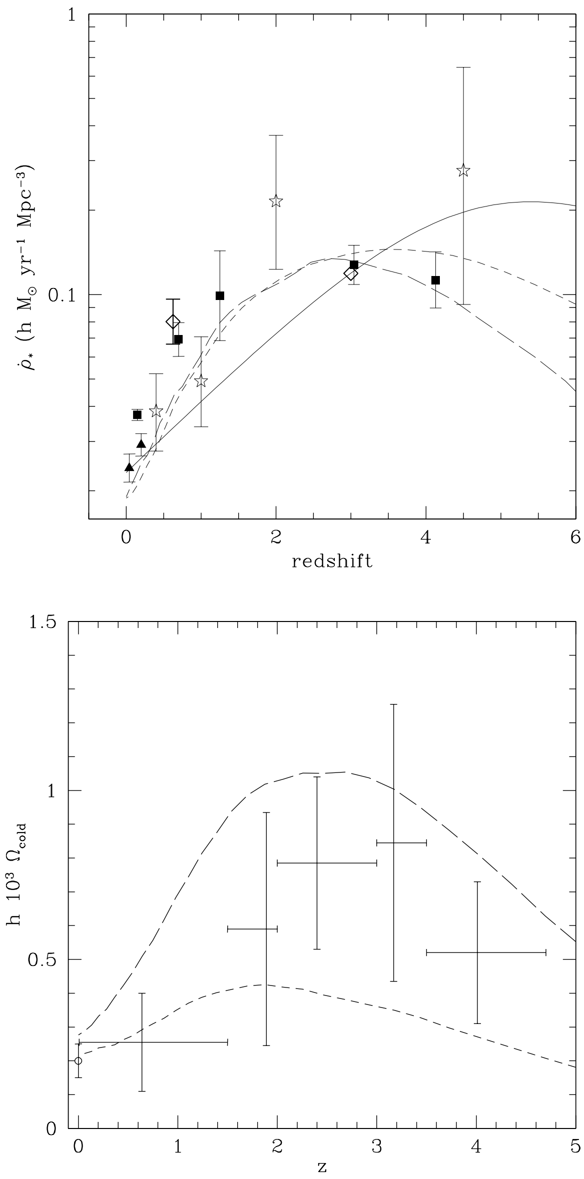

Fig. 1.2. Top: The global star formation rate density of the universe as a function of redshift. Symbols show observational estimates from the compilation in Somerville et al. (2001; SPF). The solid line shows the fit to the results of Springel & Hernquist (2002b) from hydrodynamical simulations with a multi-phase treatment of star formation and feedback. The dashed lines are results from semi-analytic models similar to those of SPF. The short dashed line shows a model using the "Kennicutt" scaling for star formation in normal galaxies, and has starbursts in major mergers. The long dashed line is a model with constant star formation efficiency ($\tau_* =$ const; see text) and bursts in both major and minor mergers, as in the "collisional starburst" model of SPF. Bottom: the global matter density in the form of cold neutral gas (in units of the critical density). Symbols show observational estimates (see SPF), while the dashed lines show the same models shown in the top panel.

one might expect the results to be quite different even though the recipes are based on similar ideas. Although no detailed comparisons have been done to date, it seems that the results are actually surprisingly similar. For example, the masses and clustering properties of galaxies at $z \sim 3$ in the hydrodynamic simulations of Weinberg et al. (2002) are quite similar to those predicted by semi-analytic models (Somerville et al. 2001; Wechsler et al. 2001). The star formation history produced by the recent SPH simulations with multi-phase star formation of Springel & Hernquist (2002) also looks reasonably similar to the one produced by semi-analytic models (Fig. 1.2.

1.5.2 Supernova Feedback

Massive stars and supernova heat the ISM and can in some cases drive large scale winds that lead to mass loss. I have already discussed the compelling need for some form of efficient feedback in CDM-based models of galaxy formation in order to avert the overcooling crisis and perhaps to solve the cooling flow conundrum, the entropy floor problem, and the angular momentum catastrophe. In addition, supernova feedback might be the best way to expel heavy elements from the dense inner parts of galaxies where they are created and strew them throughout the low-density IGM. By examining the steep slope of the CDM mass function at the low-mass end ($dN/dM \propto -2$), compared with the much flatter slope of the galaxy luminosity function, we can also quickly conclude that this feedback needs to be strongly *differential*, i.e., it needs to be more efficient in low mass halos than in high mass halos (White & Frenk 1991).

Although there is a long list of problems that we hope supernova feedback can solve, we do not know how to simulate it and there are few direct observational constraints. Some recent observational work (Martin 1999; Heckman et al. 2000) suggests that in nearby starburst galaxies, the rate at which cold gas is heated and returned to the warm diffuse ISM is approximately equal to the star formation rate. The reheating rate seems to be independent of the galaxy circular velocity, but gas may only be able to escape the potential well of the galaxy entirely (blow-away) if the circular velocity of the galaxy is less than ~ 100 km/s (Martin 1999). At high redshift, there is evidence for powerful winds and high mass loss rates in the $z \sim 3$ Lyman-break galaxies (Pettini et al. 2001) and intriguingly, the "Martin-Heckman Law" based on low redshift observations (reheating rate $\simeq$ star formation rate) seems to hold.

As with star formation, current cosmological N-body simulations simply do not have the dynamic range to directly model the effect of supernova feedback on a galaxy, and therefore again various phenomenological recipes are used. In hydrodynamical simulations, a typical approach is to return some fraction of the energy of each SN (really, of each "star particle", each of which usually weighs at least several million solar masses) to the ISM in the form of thermal energy (e.g. Katz 1992). In practice, this tends to have little effect on the galaxy, as the star formation takes place where the gas is dense and so the energy is returned to an environment where cooling times are short. The gas quickly radiates away the excess energy and very little (if any) mass is actually ejected from the galaxy, so the net effect is small. Various tricks have been used to attempt to get around this problem, including imposing a time lag by turning off cooling for some period of time following the reheating (Thacker

& Couchman 2001), and keeping the energy in a "turbulent reservoir" for some time period before letting it leak out (Springel 2000). These kinds of tricks do help to reduce the efficiency of star formation, but still do not lead to large amounts of mass loss from the galaxy.

An alternative approach involves imparting kinetic energy to the gas by giving small "kicks" to the neighboring particles (e.g. Navarro & White 1993). This is more effective in removing gas from the dense parts of the galaxy, but can lead to numerical problems, particularly in particle-based codes such as SPH — because the local density is found by smoothing over the N nearest neighbors (typically $N \simeq 32$), when particles escape into very underdense regions with very few neighboring particles it becomes impossible to calculate the density and temperature accurately. In addition, when applied within simulations of individual disk galaxies, the 'kinetic' method was found to vertically thicken the disk too much unless very low fractions ($\sim 1.0 \times 10^{-5}$) of the SN energy were assumed to be manifested as kinetic energy (e.g. Mihos & Hernquist 1994), in which case the feedback again had very little effect. Kay et al. (2002) has presented a useful comparison of several of these commonly used feedback techniques implemented within a SPH code.

Mac Low & Ferrara (1999) simulated the effect of supernova explosions on individual galaxies at very high resolution and found that significant mass loss from SN-driven winds ("blow-away") occurs only in very low mass galaxies ($M_{\rm gas} \sim 10^8 M_\odot$).

As with star formation, to make progress, probably one must consider the multi-phase nature of the ISM. Several recent analytic and semi-analytic works have approached this problem, with promising results (Efstathiou 2000; Silk 2001). Springel & Hernquist (2002) have presented a feedback scheme based on the multi-phase picture of the ISM outlined by Mc Kee & Ostriker (1977), implemented within SPH. Their results are stable to numerical resolution, unlike many previous implementations, but they still have to introduce large-scale winds essentially by hand.

To summarize, while there are many suggestions that some sort of feedback is needed that removes a significant fraction of the gas and metals from galaxies, there has been little success in simulating this process in a physical way under the assumption that the putative winds are driven by massive stars and/or supernovae. Once again, the powerful jets that are known to be associated with accretion onto black holes, which have so far been entirely neglected in detailed cosmological simulations, may have an important role to play in shaping the lives of galaxies.

1.6 Conclusions

The discovery of the tight correlation between the masses of supermassive black holes and the properties of the galactic spheroids they inhabit points to the inevitable conclusion that either galaxies know about their black holes and respond accordingly, that black hole growth is regulated by the galaxy, or that some external force shapes both the black hole and the galaxy. In this article, I have outlined ways in which things that we think we understand about galaxy formation could help us understand the formation of SMBH, and ways in which things that we *do not* understand about galaxy formation might be informed by considering the influence of the SMBH that we now believe to reside in nearly all galaxies, especially the energy released during the feeding of these beasts.

Here is a summary of the main points:

(1) We are in the process of gaining a better understanding of the internal structure of dark matter halos (their matter density and angular momentum) and the connection with the mass accretion and merger history. This will help improve modeling of the inner structure of galaxies, which will be crucial to understanding the formation and fueling of central black holes.

(2) There is a series of possibly interconnected problems in CDM-based models of galaxy formation:

- the angular momentum catastrophe and the angular momentum profile mismatch
- the overcooling crisis
- the cooling flow conundrum
- the entropy floor problem

All suggest the need for efficient non-gravitational heating and mass loss from galaxies. Supernovae and AGN are both good candidates for driving these processes, and it remains to be seen what the balance of importance will be between the two. Heavy elements may be able to provide important clues here.

(3) Modeling of star formation and supernova feedback in cosmological simulations remains extremely primitive, partly due to our lack of understanding of these processes and partly due to the numerical challenge of the very large dynamical range involved. Progress in both areas will probably only be made by explicitly treating the multi-phase nature of the ISM.

A recurring theme in this paper has been the need for some process or processes that make cooling and star formation less efficient by heating or physically removing cold gas from galaxies. So far, it seems difficult to accomplish this at the needed level with supernovae. This may reflect limitations in simulation techniques, however, it may hint that the powerful winds associated with accretion onto SMBH (e.g. manifested as radio galaxies and BAL quasars) have an important role to play in filling in some of the missing pieces in our understanding of galaxy formation. There is no doubt that pursuing the interconnection between galaxies, quasars, and the supermassive black holes they harbor will be an important topic of investigation in the years to come.

Bibliography

Abadi, M.G., Navarro, J.F., Steinmetz, M., Eke, V.R. 2002, preprint, astro-ph/0212282
Adelberger, K. L. & Steidel, Charles C., 2000, ApJ, 544, 218
Allen, S.W., Fabian, A.C. 1998, MNRAS, 297, L57
Arnaud, M., Evrard, A.E. 1999, MNRAS, 305, 631
Avila-Reese, V. & Firmani, C. 2000, RevMexAA, 36, 23
Balogh, M.L., Pearce, F.R., Bower, R.G., Kay, S.T. 2001, MNRAS, 326, 1228
Barnes, J. & Efstathiou, G. 1987, ApJ, 319, 575
Benson, A. J., Pearce, F. R., Frenk, C. S., Baugh, C. M., & Jenkins, A. 2001, MNRAS, 320, 261
Blain, A.W., Smail, I., Ivison, R.J., Kneib, J.-P., Frayer, D.T. 2003, Physics Reports, in press, astro-ph/0202228
Blumenthal, G., Faber, S., Flores, R., & Primack, J. 1986, ApJ, 301, 27
Bullock, J. S., Dekel, A., Kolatt, T. S., Kravtsov, A. V., Klypin, A. A., Porciani, C., & Primack, J. R. 2001a, ApJ, 555, 240
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
Cole, S., Aragón-Salamanca, A., Frenk, C., Navarro, J., & Zepf, S. 1994, MNRAS, 271, 781
Cole, S., Lacey, C. G., Baugh, C. M., & Frenk, C. S. 2000, MNRAS, 319, 168
Dalcanton, J., Spergel, D., & Summers, F. 1997, ApJ, 482, 659
Crampin, D.J., & Hoyle, F. 1964, ApJ, 140, 99
Doroshkevich, A.G. 1970, Astrofizika, 6, 581
Efstathiou, G. 2000, MNRAS, 317, 697
Eke, V., Efstathiou, G., & Wright, L. 2000, MNRAS, 315, L18
Fabian, A.C. 2002, "Galaxy Evolution: Theory and Observations", Eds. V. Avila-Reese, C. Firmani, C. Frenk, & C. Allen, RevMexAA SC, astro-ph/0210150
Fabian, A. C., Voigt, L. M., Morris, R. G. 2002, MNRAS, 335L, 71
Fall, S. M. & Efstathiou, G. 1980, MNRAS, 193, 189
[Ferguson & Johnson(2001)] Ferguson, A. M. N. & Johnson, R. A. 2001, ApJ, 559, L13
Flores, R., Primack, J., Blumenthal, G., & Faber, S. 1993, ApJ, 412, 443
Governato, F. et al. 2002, preprint, astro-ph/0207044
Heckman, T. M., Lehnert, M. D., Strickland, D. K., Armus, L., 2000, ApJS, 129, 493
Helly, J.C., Cole, S., Frenk, C.S., Baugh, C.M., Benson, A., Lacey, C., Pearce, F.R., MNRAS, in press, astro-ph/0202485
Kaastra, J. et al. 2001, A&A, 365, L99
Kauffmann, G. & Haehnelt, M. 2000, MNRAS, 311, 576
Kauffmann, G., Colberg, J., Diaferio, A., & White, S. D. M. 1998, MNRAS, 303, 188
Kauffmann, G., White, S., & Guiderdoni, B. 1993, MNRAS, 264, 201
Katz, N. 1992, ApJ, 391, 502
Kay, S. T., Pearce, F. R., Frenk, C.S., & Jenkins, A. 2002, MNRAS, 330, 113
Kennicutt, R. 1989, ApJ, 344, 685
—. 1998, ApJ, 498, 181
Klypin, A., Zhao, H., & Somerville, R. S. 2002, ApJ, 573, 597
Kravtsov, A. V., Yepes, G., 2000, MNRAS, 318, 227

Lemson, G., & Kauffmann, G. 1999, MNRAS, 302, 111
Lin, D. N. C. & Pringle, J. E. 1987, ApJ, 320, L87
Lloyd-Davies, E. J., Ponman, T. J., & Cannon, D. B. 2000, MNRAS, 315, 689
Mac Low, M. & Ferrara, A. 1999, ApJ, 513, 142
Maller, A. H. & Dekel, A. 2002, MNRAS, 335, 487
Maller, A. H., Dekel, A., & Somerville, R. 2002, MNRAS, 329, 423
Markevitch, M. 1998, ApJ, 504, 27
Martin, C. L. 1999, ApJ, 513, 156
Mc Kee, C.F. & Ostriker, J. P. 1977, ApJ, 218, 148
Mestel, L. 1963, MNRAS, 126, 553
Mihos, J.C. & Hernquist, L. 1994, ApJ, 437, 611
Mo, H., Mao, S., & White, S. 1998, MNRAS, 295, 319
Nath, B. & Roychowdhury, S. 2002, preprint, astro-ph/0202201
Ponman, T.J., Cannon, D.B., Navarro, J.F. 1999, Nature, 397, 135
Peterson, J. R. et al. 2001, A&A, 365, L104
Narayan, R. & Medvedev, M.V. 2001, ApJ, 562, L129
Navarro, J. F. & Benz, W. 1991, ApJ, 380, 320
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Navarro, J. F. & Steinmetz, M. 2000, ApJ, 538, 477
Navarro, J. F. & White, S. D. M. 1993, MNRAS, 265, 271
Navarro, J. F. & White, S. D. M. 1994, MNRAS, 267, 401
Peebles, P. J. E. 1969, ApJ, 155, 393
Pettini, M., Shapley, A. E., Steidel, C. C., Cuby, J.-G., Dickinson, M., Moorwood, A. F. M., Adelberger, K. L., Giavalisco, M. 2001, ApJ, 554, 981
Sáiz, A., Domínguez-Tenreiro, R., Tissera, P. B., & Courteau, S. 2001, MNRAS, 325, 119
Silk, J. 2001, MNRAS, 324, 313
Slyz, A. D., Devriendt, J. E. G., Silk, J., & Burkert, A. 2002, MNRAS, 333, 894
Springel, V. 2000, MNRAS, 312, 859
Springel, V. & Hernquist, L., 2002, preprint, astro-ph/0206393
Springel, V. & Hernquist, L., preprint, astro-ph/0206395
Somerville, R. & Primack, J. 1999, MNRAS, 310, 1087
Somerville, R. S., Primack, J. R., & Faber, S. M. 2001, MNRAS, 320, 504
Sommer-Larsen, J. & Dolgov, A. 2001, ApJ, 551, 608
Sommer-Larsen, J., Gelato, S., & Vedel, H. 1999, ApJ, 519, 501
Tamura, T. et al. 2001, A&A, 365, L87
Thacker, R. J. & Couchman, H. M. P. 2001, ApJ, 555, L17
van den Bosch, F. C. 2000, ApJ, 530, 177
—. 2001, MNRAS, 327, 1334
—. 2002, MNRAS, 332, 456
van den Bosch, F.C., Abel, T., Croft, R.A.C., Hernquist, L., White, S.D.M. 2002, ApJ, 576, 21
Voigt, L. M., Schmidt, R. W., Fabian, A. C., Allen, S. W., Johnstone, R. M. 2002, MNRAS, 335L, 7
Volonteri, M., Haardt, F., Madau, P., 2003, ApJ, in press, astro-ph/0207276
Vitvitska, M., Klypin, A., Kravtsov, A., Wechsler, R., Primack, J., & Bullock, J. 2002, ApJ, 581, 799
Weinberg, D. H., Hernquist, L., Katz, N. 2002, ApJ, 571, 15
White, S.D.M. 1984, MNRAS, 286, 38
White, S.D.M. & Frenk, C.S., 1991, ApJ, 379, 52
White, S.D.M. & Rees, M.J. 1978, MNRAS, 183, 341
Wechsler, R. H., Somerville, R. S., Bullock, J. S., Kolatt, T. S., Primack, J. R., Blumenthal, G. R., & Dekel, A. 2001, ApJ, 554, 1
Wechsler, R. H., Bullock, J. S., Primack, J. R., Kravtsov, A. V., & Dekel, A. 2002, ApJ, 568, 52
Weil, M. L., Eke, V. R., & Efstathiou, G. 1998, MNRAS, 300, 773
Wu K.K.S., Fabian A.C., Nulsen P.E.J., 2000, MNRAS, 381, 889
Yepes, G., Kates, R., Khokhlov, A., Klypin, A. 1997, MNRAS, 284, 235
Yoshida, N., Stoehr, F., Springel, S., & White, S. 2002, MNRAS, in press, (astro-ph/0202341)