Class #17: Hierarchical Galaxy Assembly

Structure and Dynamics of Galaxies, Ay 124, Winter 2009

March 2, 2009

Throughout this class we've treated galaxies as equilibrium gravitating systems. While this is often a reasonable approximation (since changes in their overall structure seem to mostly occur on timescales long compared to their dynamical times) we know that this assumption cannot always hold: there were no galaxies around just after the Big Bang and so they must have formed some time between then and now—and that implies a non-equilibrium process.

Galaxies are made from baryonic material (mostly hydrogen and helium) but the mass density of the universe is dominated by dark matter. As a result, the early stages of structure formation in the Universe occur via the gravitational collapse of systems of dark matter. In Ay 127 you'll develop simple models for this gravitational collapse process. For our purposes, only a few key points matter:

1. Regions of dark matter which are initially slightly overdense will become more overdense due to their self-gravity. They will eventually collapse to form a near-equilibrium, approximately spherical distribution of dark matter supported against further collapse by the random motions of the constituent dark matter particles. This process is known as "virialization".
2. These "dark matter halos" are typically around 200 times denser than the mean density of the Universe at the time of their formation.
3. In cold dark matter Universes, the first halos to form are small and low mass. These can be later subsumed into larger halos. As we'll see later, even though they're subsumed they're not entirely destroyed—so halos may contain populations of orbitting subhalos.
4. The deep gravitational potential wells of dark matter halos form a natural place in which galaxies may form.

The details of the galaxy formation process are still not all that well understood[1]. However, the basics ideas involved are now reasonably well established and we'll review the key concepts here.

[1]Due to the fact that the physics is non-linear, occurs over a very large range of mass and length scales and involves lots of messy physics such a hydrodynamics, star formation, supernovae explosions, active galactic nuclei, radiative transfer etc.

1 Cooling of gas in halos

While matter is distributed diffusely the pressure of the baryonic component doesn't significantly affect its motions and so we can expect the baryonic material to follow along with the gravitationally dominant dark matter. As a result, any dark matter halo sufficiently massive to overcome the pressure forces will contain a mass of baryons approximately equal to a fraction $\Omega_{\rm b}/\Omega_0$ of its total mass.

Suppose that the mass of the halo is $M_{\rm V}$ and that it has a radius of $R_{\rm V} = (3M_{\rm V}/4\pi\bar{\rho})^{1/3}$. Baryonic material arriving at the edge of this halo will have a specific kinetic energy of

$$\frac{1}{2}v^2 = \frac{{\rm G}M_{\rm v}}{R_{\rm v}} - \frac{{\rm G}M_{\rm v}}{2R_{\rm v}}, \tag{1}$$

where we've assumed that the gas fell in from a radius of $2R_{\rm V}$. (You'll see why this radius makes sense when you study spherical collapse models in Ay 127.) This velocity can greatly exceed the sound speed in the gas and so we may expect a strong shock to occur at some point. This will convert the kinetic enery of the gas into random, thermal motions, thereby heating the gas to a temperature of around

$$k_{\rm B}T \approx \frac{1}{2}\mu m_{\rm H} v^2. \tag{2}$$

We refer to this as the virial temperature and can express it as

$$T_{\rm V} = \frac{1}{2}\frac{\mu m_{\rm H}}{k_{\rm B}}\frac{{\rm G}M_{\rm v}}{R_{\rm v}}. \tag{3}$$

For a halo such as that of the Milky Way ($M_{\rm V} \sim 10^{12}M_\odot$, $R_{\rm V} \approx 300$kpc) this temperature is of order 10^6K. For a massive cluster o galaxies it can reach as high as 10^8K.

Once heated to this temperature, the pressure of the gas is sufficient to support itself against the gravitational pull of the dark matter[2] and so the gas will settle into a hydrostatic, pressure supported atmosphere. This is where significant differences between the behavior of dark matter and baryonic matter begin to appear. Unlike dark matter, the baryons can lose energy by radiating. As they do this, they will begin to lose pressure support and so must eventually sucumb to the gravitational pull of the dark matter and flow toward the center of the halo.

The gas has a thermal energy per unit volume of

$$\mathcal{E} = \frac{N}{2} n_{\rm t} k_{\rm B} T_{\rm v} \tag{4}$$

where $N = 3$ or 5 depending on whether you think that the cooling occurs at constant density or pressure, and $n_{\rm t}$ is the total number of particles (ions, atoms, electrons) per unit volume. We can write the rate at which the gas is radiating energy per unit volume as

$$\dot{\mathcal{E}} = n_{\rm H}^2 \Lambda(T, \mathbf{Z}) \tag{5}$$

where $n_{\rm H}$ is the density of hydrogen and $\Lambda(T, \mathbf{Z})$ is known as a the cooling function. Since the radative processes at work (at least in the regimes applicable to most forming galaxies) are two-body interactions, we can factor out the density dependence in this way, leaving $\Lambda(T, \mathbf{Z})$ a function

[2] You can convince yourself of this quite easily using order of magnitude estimates.

of only temperature and metallicity. To compute $\Lambda(T, \mathbf{Z})$ the ionization states of the various elements are computed assuming that they are at their equilibrium ratios for gas in collisional ionization equilibrium (i.e. that ionizations due to collisions with free electrons are balanced by the rate of recombination). Then the rates of cooling due to various atomic processes (recombination, collisional excitation, Bremsstrahlung etc.) are summed. An example of the resulting cooling function is shown in Fig. 1.

A characteristic timescale for cooling can then be estimated as:

$$t_{\rm cool} = \frac{\mathcal{E}}{\dot{\mathcal{E}}} = \frac{N}{2} \frac{n_{\rm t}}{n_{\rm H}} \frac{k_{\rm B} T_{\rm v}}{n_{\rm H} \Lambda(T, \mathbf{Z})} \propto 1/\rho. \tag{6}$$

As indicated above, this is inversely proportional to density. We expect the inner regions of the gaseous atmosphere in a halo to be denser than the outer regions. A simple model for the radial density variation is

$$\rho(r) \propto r^{-2}. \tag{7}$$

In such a model, we find

$$t_{\rm cool} \propto r^2. \tag{8}$$

If the atmosphere has been cooling for a time t the radius at which $t = t_{\rm cool}$ is therefore

$$r_{\rm cool} \propto t^{1/2}. \tag{9}$$

As such, we expect the radius within which gas has had sufficient time to cool to grow as the halo ages, allowing more and more mass to lose pressure support and flow to the center of the halo. In practice, things are more complicated for several reasons: the atmosphere will respond to the cooling of the inner regions; cooling may result in a multiphase atmosphere containing cool, dense clouds; and some gas may escape being shocked altogether and therefore have no need to cool down before it can flow to the center of the halo.

2 Angular momentum and sizes of disks

If the cooling gas in a dark matter halo cooled to zero temperature and has no angular momentum, it would fall to zero radius. However, it turns out that the gas does have some angular momentum. If this angular momentum is conserved during the post-cooling collapse then the gas must spin more and more rapidly as it shrinks to smaller and smaller radii. Eventually, it must be spinning fast enough to become rotationally supported against further collapse. At this point, we may expect it to form a rotationally supported disk. This is good, since such disks are a major component of galaxies.

Where does this angular momentum originate? During the gravitational collapse of a dark matter halo the halo can experience gravitational torques due to the non-isotropic distribution of mass surrounding it[3]. These torques transfer angular momentum to the collapsing halo, causing it to spin up. Since the baryonic material is behaving more-or-less like collisionless dark matter at this point it will gain angular momentum and spin up also.

[3] The distribution will be isotropic on average of course, but statistical variations mean that any given collapsing region will see a slightly non-isotropic mass distribution surrounding it.

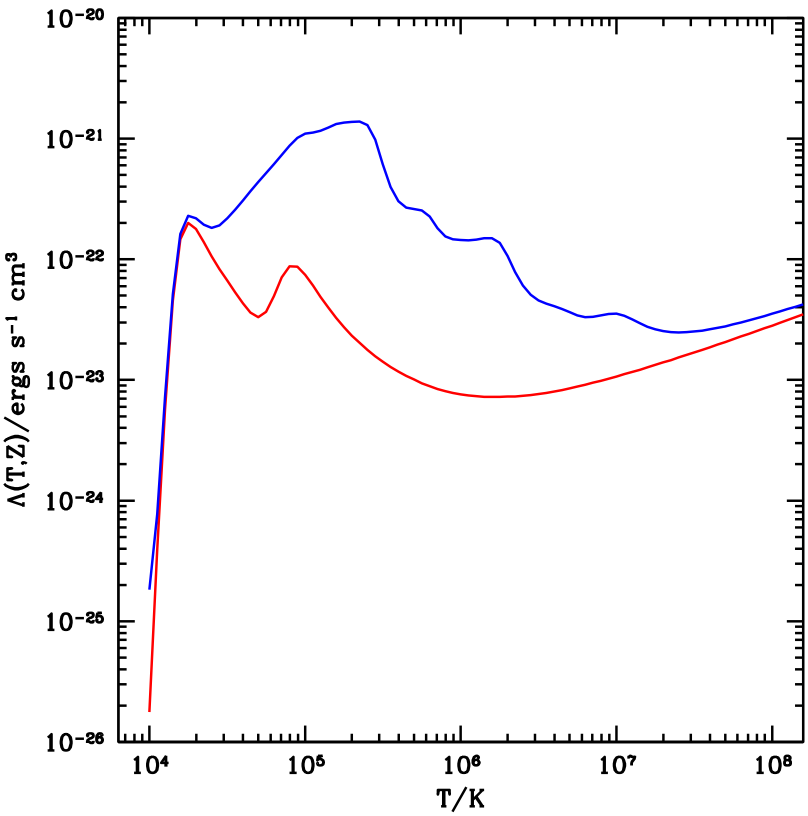

Figure 1: The cooling function, $\Lambda(T, Z)$, as a function of temperature for gas in collisional ionization equilibrium. Theh red curve is for primordial gas, while the blue curve is for gas of Solar metallicity. The two peaks in the primordial curve correspond to cooling by collisional excitation of hyrdrogen and helium. At high temperatures both curves become dominated by Bremsstrahlung radiation ($\Lambda \propto \sqrt{T}$).

The angular momentum content of a dark matter halo can be characterized by the dimensionless spin parameter defined as:

$$\lambda = |E|^{1/2}|\mathbf{J}|/\mathrm{G}M^{5/2} \sim |J|/M_\mathrm{V}R_\mathrm{V}V_\mathrm{V}, \tag{10}$$

where E is the energy of the halo and $\mathbf{J}$ its angular momentum. The first form is the standard definition, but the second is somewhat easier to work with. The typical value of λ is difficult to calculate analytically (as it changes significantly during the non-linear parts of the gravitational collapse of the dark matter halo), but numerical studies indicate that it has a typical value of $\lambda = 0.03$–0.04 with significant dispersion.

Suppose that some mass of gas M_d with angular momentum J in the halo is able to cool. If it achieves rotational supprt against the pull of the dark matter at some radius R_d then we must have that

$$\left(\frac{J}{M_\mathrm{d}R_\mathrm{d}}\right)^2 = V_\mathrm{c}^2(R_\mathrm{d}) = \frac{\mathrm{G}M(R_\mathrm{d})}{R_\mathrm{d}}, \tag{11}$$

where $M(R)$ is the mass of dark matter within radius R. Assuming that angular momentum is conserved during the collapse of this gas then J is just equal to its initial value of $J = \lambda M_\mathrm{d}R_\mathrm{V}V_\mathrm{V}$. Further assuming that the dark matter has a $\rho \propto r^{-2}$ (not a bad approximation) such that $M(R) \propto R$ we find

$$\begin{aligned} \lambda^2\left(\frac{R_\mathrm{V}}{R_\mathrm{d}}\right)^2 V_\mathrm{V}^2 &= \frac{\mathrm{G}M_\mathrm{V}}{R_\mathrm{V}} \\ \lambda^2\left(\frac{R_\mathrm{V}}{R_\mathrm{d}}\right)^2 \frac{\mathrm{G}M_\mathrm{V}}{R_\mathrm{V}} &= \frac{\mathrm{G}M_\mathrm{V}}{R_\mathrm{V}}, \end{aligned} \tag{12}$$

which simplifies to give

$$R_\mathrm{d} = \lambda R_\mathrm{v}. \tag{13}$$

So, we should expect the size of a galaxy disk to be a fraction λ of the virial radius of the dark matter halo. For the Milky Way $R_\mathrm{v} \approx 300$kpc and typical spin parameters are $\lambda = 0.035$, so we'd expect $R_\mathrm{d} \approx 10$kpc. The half-mass radius of the Milky Way's disk is around 6kpc, so this is a pretty good estimate.

Note that we've made one or two important approximations here:

1. We assumed that the baryonic component of the halo has the same λ as the dark matter. This turns out to be not exactly true.
2. We assumed that the angular momentum of the gas is conserved during collapse. Recent numerical work suggests that this is actually a good approximation.
3. We ignored the galaxy's self-gravity when computing the rotational support radius. We should really replace $M(R)$ with $M(R) + M_\mathrm{d}$ on the right-hand side of eqn. (11). This would result in a somewhat smaller radius R_d.
4. We ignored the backreaction of the baryons on the dark matter halo. The gravitational potential due to the concentration of baryons in the halo center will cause the dark matter halo to contract slightly, boosting the amount of dark matter in the inner regions. This too will result in a slightly smaller disk.

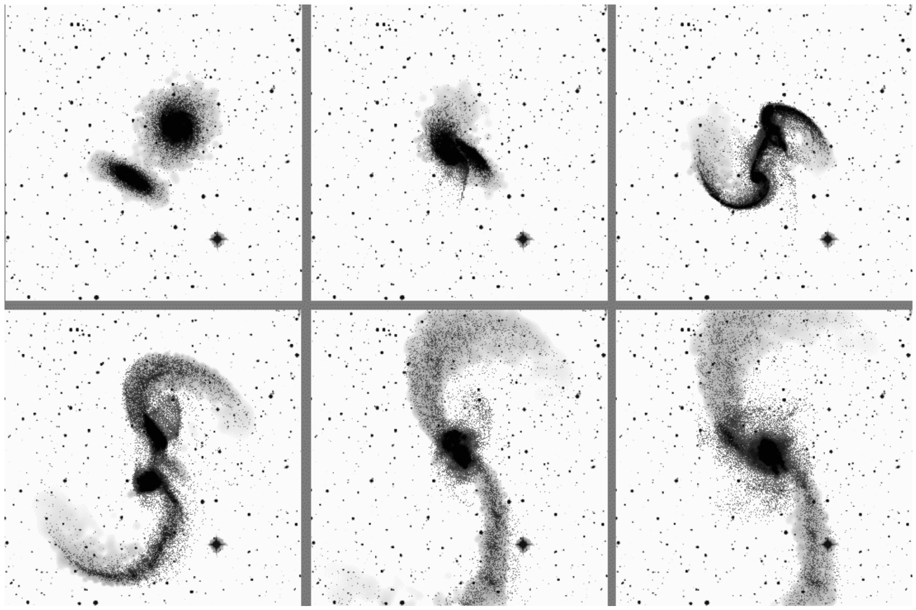

Figure 2: A sequence of snapshots from a simulation of two merging disk galaxies. The violent merging process destroys the fragile disks and leaves something which looks more or less like an elliptical galaxy (with some tidal features that will fade with time).

3 Merging/formation of spheroids

The other main class of galaxy are the ellipticals, which are not supported by rotation but instead by the random motions of their stars. Since being first suggested by Toomre & Toomre in the 1970's the idea that ellipticals form through violent mergers of pre-existing galaxies has become the widely accepted hypothesis for elliptical galaxy formation (at least for the most massive ellipticals).

The merging process is highly non-linear and can only be studied quantitatively using numerical simulations (see Fig. 2). The basic idea though is that the merging process leads to large and rapid fluctuations in the gravitational potential and that this tends to randomize the orbits of stars (a process known as "violent relaxation" and first proposed by Lynden-Bell—it's essentially a maximization of entropy argument, although, as we've seen, entropy cannot be well-defined for a gravitating system). The result is the destruction of any pre-existing disks and the formation of a pressure supported spheroidal stellar distribution. N-body simulations show that the resulting spheroid often has structural and kinematical similarities with observed ellipticals.

If merging is such an important process, it's important to ask how often it happens. In a cold dark matter Universe, a dark matter halo can contain many smaller halos from earlier generations of structure formation. These subhalos, each of which can in principle contain their own galaxy, will be orbitting around their host halo. Suppose that there are N such subhalos in a host halo, each moving with velocity of order $V_{\rm V}$ (the virial velocity of the host halo). We'd expect a galaxy to randomly run into another galaxy after a time $t_{\rm coll}$ where

$$\pi R_{\rm gal}^2 V_{\rm V} t_{\rm coll} \frac{3N}{4\pi R_{\rm V}^3} = 1. \tag{14}$$

For a galaxy cluster, $N \sim 1000$, $R_{\rm V} \approx 1$Mpc, $V_{\rm V} \approx 1000$km/s and a typical galaxy radius is 10kpc. Using these numbers we find $t_{\rm coll} \sim 13$Gyr, or about once in during the age of the Universe. In practice, most of the encounters will be highly unbound, so will be fly-bys which may disturb the galaxies but will not cause them to actually merge.

To get a significant number of mergers we actually need a dissipative process that can remove energy from the orbits of subhalos causing them to sink towards the center of the host halo and merge with any galaxy located there. This dissipation is provided by the process of *dynamical friction* which arises due to the back reaction of dark matter particles in the host halo on the orbitting subhalo.

Way back when we were considering the collisionless Boltzmann equation we derived the change in the velocity of a test particle as it passed by a mass M. We made some approximations there, but we can actually do the calculation exactly (see Binney & Tremaine for example). The result, where we consider a dark matter particle of mass m from the host halo moving by a subhalo of mass M at relative velocity at infinity V_0 and with impact parameter b is:

$$\Delta \mathbf{V}_{||} = \frac{2m}{M} \left(1 + \frac{b^2 V_0^4}{\mathrm{G}^2 M^2} \right)^{-1} \mathbf{V}_0 \tag{15}$$

and

$$\Delta \mathbf{V}_{\perp} = \frac{2m V_0^3}{\mathrm{G} M^2} \left(1 + \frac{b^2 V_0^4}{\mathrm{G}^2 M^2} \right)^{-1} \mathbf{b}. \tag{16}$$

If we now imagine a whole distribution of such dark matter particles flying by, we can derive the rate of change of the subhalo velocity:

$$\frac{\mathrm{d}\mathbf{V}_{||}}{\mathrm{d}t} = f(\mathbf{V}_0) \int_0^{b_{\rm max}} 2\pi b n \mathbf{V}_0 \Delta V_{||} \mathrm{d}b, \tag{17}$$

or

$$\frac{\frac{\mathrm{d}V_{||}}{\mathrm{d}t} = 2\pi \ln(1+\Lambda^2) \rho \mathrm{G}^2 M f(\mathbf{V}_0) \mathbf{V}_0}{V_0^3} \tag{18}$$

where $\Lambda = b_{\rm max} V_0^2 / \mathrm{G}M$. The net change in the $V_{\perp}$ component must be zero by symmetry. Thus, the acceleration of the subhalo due to this dynamical friction process is:

$$\frac{\mathrm{d}\mathbf{V}_M}{\mathrm{d}t} = 2\pi \ln(1+\Lambda^2) \rho \mathrm{G}^2 M \int f(\mathbf{V}_m) \frac{(\mathbf{V}_m - \mathbf{V}_M)}{|\mathbf{V}_m - \mathbf{V}_M|^3} \mathrm{d}^3 \mathbf{V}_m. \tag{19}$$

The integral has a useful form: replacing $\mathbf{V}_m$ by $\mathbf{x}$ and $f(\mathbf{V}_m)$ by $\rho(\mathbf{x})$ we would have the integral that gives us the gravitational force due to a mass density distribution. If we assume that the velocity distribution is isotropic, then $f(\mathbf{V}_m)$ is spherically symmetric and, since we know that the mass outside of radius r in a spherical mass distrribution does not contribute to the gravitational field at radius r we can write

$$\frac{\mathrm{d}\mathbf{V}_M}{\mathrm{d}t} = 2\pi \ln(1+\Lambda^2) \rho \frac{\mathrm{G}^2 M}{V_M^2} F(< V_M) \hat{\mathbf{V}}_M. \tag{20}$$

The rate at which this acceleration removes energy from the subhalo's orbit is then

$$\begin{aligned}\dot{E} &= M\mathbf{V}_M \cdot \frac{\mathrm{d}\mathbf{V}_M}{\mathrm{d}t} \\ &= 2\pi \ln(1+\Lambda^2)\rho \frac{\mathrm{G}^2 M^2}{V_M} F(< V_M). \end{aligned} \tag{21}$$

Since $|E| = MV_M^2/2$ we have a timescale for dynamical friction of

$$\tau_{\mathrm{DF}} = |E|/\dot{E} = \frac{V_M^3}{4\pi \ln(1+\Lambda^2)\rho \mathrm{G}^2 M F(< V_M)}. \tag{22}$$

Using $\rho = 3M_{\mathrm{V}}/4\pi R_{\mathrm{V}}^3$ and $V_{\mathrm{M}}^2 \approx V_{\mathrm{V}}^2 = \mathrm{G}M_{\mathrm{V}}/R_{\mathrm{V}}$ we can simplify this to

$$\begin{aligned}\tau_{\mathrm{DF}} &\approx \frac{R_{\mathrm{V}}}{V_{\mathrm{V}}} \frac{M_{\mathrm{V}}}{M} \frac{1}{F(< V_{\mathrm{V}}) \ln(1+\Lambda^2)} \\ &\sim \frac{t_{\mathrm{H}}}{\sqrt{\Delta}} \frac{M_{\mathrm{V}}}{M} \frac{1}{\ln(1+\Lambda^2)}. \end{aligned} \tag{23}$$

Since $\Delta \approx 200$ we expect satellites with mass $M \lesssim M_{\mathrm{V}}/\sqrt{200}$ to not be able to merge within a Hubble time. More massive subhalos though will be dragged to the center of their host halo where any galaxy they contain can then merge with any galaxy at that location.

4 Star formation and feedback

Once gas has collected into a dense stellar disk we expect that it should be able to cool further (via various metal lines initially and then through the formation of molecules) to form molecular clouds and eventually stars. If this were all that happened, each dark matter halo would eventually contain a galaxy with a mass equal to $\Omega_{\mathrm{b}}/\Omega_0$ times its total mass. We can predict the distribution of dark matter halo masses (as you'll see in Ay 127) and so this simple model allows us to compute the corresponding distribution of galaxy masses. Unfortunately, it turns out to be wildly different from what's observed—there are far too many low mass galaxies in such a model. This is a long-standing problem for galaxy formation theory to explain. The most widely accepted explanation is that star formation and the subsequent supernova explosions lead to a *negative feedback loop* which causes star formation to self-regulate and thereby reduce the fraction of a halo's material which is turned into stars significantly below $\Omega_{\mathrm{b}}/\Omega_0$.

A single supernova explosion will release an energy of around 10^{51}ergs in a form which could potentially influence the surrounding galaxy (a much larger energy is released in the form of neutrinos, but they just free-stream out of the galaxy). Since we expect on order of one supernova for every $100M_\odot$ of stars formed (given a standard IMF) we can say that roughly 10^{49}ergs of energy is deposited into the ISM per Solar mass of stars formed.

Consider a galaxy with a flat rotation curve, V_{c}. The gravitational potential energy gained in moving a mass M from radius r_1 to r_2 in such a potential is $\Delta\Phi = V_{\mathrm{c}}^2 \ln r_2/r_1$. Suppose then that we want to know how much mass supernovae could reasonably eject from such a galaxy. This would be

$$10^{49}\text{ergs}M_\odot^{-1}M_\star = M_\text{e}V_\text{c}^2 \ln r_2/r_1, \tag{24}$$

where $M_\star$ is the mass of stars formed and M_e is the mass of gas ejected. The result won't depend too strongly on the ratio r_2/r_1, so let's adopt $r_2/r_1 \sim 10$ (to get the gas far out of the galaxy). The above can then be re-written as

$$\beta \equiv \frac{M_\text{e}}{M_\star} = 5.5 \left(\frac{V_\text{c}}{200\text{km/s}}\right)^{-2}. \tag{25}$$

This implies that we could expect to see outflows even from Milky Way-like galaxies at rates comparable to the star formation rate and that outflows should be stronger for lower mass galaxies (which have smaller V_C). This assumes that all of the energy from the supernovae is efficiently coupled into an outflow, which is unlikely to be the case in reality (some of the energy could be radiated away during the initial phase of the explosion while the material is still very dense). Some "feeeback" of this type seems to be required by galaxy formation models though. Without it, far too many faint, low-mass galaxies are predicted (i.e. many more than are observed) due to the prevalence of low mass dark matter halos. Feedback significantly reduces the luminosity and mass of a galaxy that can form in a given dark matter halo. As a result, galaxies of fixed stellar mass must form in more massive (and less abundant) dark matter halos when feedback occurs. This greatly reduces their numbers, bringing them in line with osbervations.

While feedback from supernovae is effective in low mass galaxies we can see from eqn. (25) that it quickly becomes irrelevant in more massive galaxies. In a galaxy cluster for example, the circular speed is of order 1000km/s and so $\beta \approx 0.2$ leading to very weak feedback. For this reason, galaxy formation models which include only supernovae feedback tend to produce overly massive galaxies at the centers of cluster-sized halos.

This problem was traditionally "solved" by simply deciding that gas in such clusters couldn't cool. A more modern approach is to explain this lack of cooling by assuming that it is offset by heating due to energy input from a forming supermassive black hole. Osbervations suggest that all galaxies contain a black hole with a mass of about 0.1% of the stellar mass of the spheroidal component. The formation of such a black hole therefore releases an energy of $0.001\epsilon M_\odot \text{c}^2 \approx 2 \times 10^{51}\epsilon$ergs for every Solar mass of stars formed. Here ϵ is the radiative efficiency of the black hole (i.e. the fraction of the rest mass energy of accreted material that is emitted as radiation rather than being swallowed by the black hole) and is typically of order 10%. This energy release can therefore be significantly larger than that associated with supernovae (accretion onto a black hole is a much more efficient process than nuclear fusion for extracting energy from matter!) and sufficient to offset cooling in even the most massive clusters. The outstanding problems in this scenario are to explain precisely how a black hole creates and outflow and couples it to the surrounding atmosphere of hot gas and to determine whether this can really lead to self-regulation of star formation (and thereby explain the 0.1% ratio of black hole to stellar mass).