AST541 Notes: Spherical Collapse, Press-Schechter Oct/Nov 2018

We spent the last two months studying cosmology. The basic goal, which was highlighted by the successful of WMAP , Planck, and low-z galaxy redshift surveys, is to establish a basic world model. We now have a concordance cosmology, a LCDM universe, with large cosmological constant, low matter density dominated by CDM, a flat geometry and an almost scale invariant initial power spectrum. This model is consistent with predications from inflationary cosmology, and is consistent with all available data.

The goal for the last month is galaxy formation, to see how collapsed objects, galaxies, clusters of galaxies grow in the post recombination universe after the CMB era. We will also exam the evolution of intergalactic gas, the IGM, as it is closely tied to galaxy formation.

Our goal is to present some basic tools, and go over basic processes, while highlight some of the outstanding issues. We will not go into great details about the popular model of the day, so I hope that things discussed will have some lifetime before they become irrelevant. Two processes that we have not discussed so far, but are crucial for galaxy formation: nonlinear evolution of perturbation, which results in collapse and viralization of gravitationally bound objects; and dissipation processes, which will affect how gas, or baryons evolve and finish the formation of galaxies. We will discuss nonlinear collapse and P-S first.

1 Nonlinear collapse

We have studied linear perturbations. Once these perturbations go nonlinear (which we define precisely later), the collapse proceeds rapidly via gravitational instability. This leads to galaxies, large-scale structure, etc.

The successful "concordance cosmology" postulates that we live in an adiabatic Λ cold dark matter Universe.

Cold dark matter models are characterized by the bottom-up assembly of halos, sometimes called "hierarchical" structure formation. Since CDM has no (known) temperature, it has no pressure, and so can collapse into halos at the smallest scales. These halos then merge into larger halos as gravity pulls them together.

Until about a decade or two ago, some debate about "top-down" vs. "bottom-up" formation. Top down formation occurs when dark matter is hot (neutrinos?), and then larger structures fragment into smaller ones, like a GMC fragments upon collapse into Jeans mass-sized lumps.

With tight limits on the amount of hot dark matter from studying the small-scale power spectrum, today nobody talks much about top-down structure formation. Everything is "hierarchical".

Aside: Warm dark matter has been advocated for solving very small scale issues with CDM (e.g. cusp problem). The "warmness" of DM can be characterized by a DM particle mass. Current limits (mostly from Lyα forest) are $m_{\rm WDM}$ >few keV. Particle theory DM candidates are typically in the GeV-TeV range.

2 Zel'dovich approximation

It is possible to gain insights into the early collapse of perturbations by extending linear perturbation theory. A perturbation in 3-D is in general triaxial. Qualitatively, collapse occurs first along pancakes (1D), then filaments (2D) and finally quasi-spherical halos (3D).

Today, computer simulations can accurately follow this procedure from the linear regime to the present day halo population. But computers are actually poor in the $\delta <\sim 1$ regime, since it involves subtracting two large numbers that nearly cancel.

Linear theory is technically only valid when $\delta \ll 1$, but we can follow evolution to $\delta \sim 1$ using the Zel'dovich (1970) approximation.

We will work in a "Lagrangian" frame. As opposed to measuring coordinates relative to a fixed (or comoving) grid ("Eulerian" frame), we will study the deformation of material around a location moving with a particle $\mathbf{q}$. A high-density peak will draw in matter from it's underdense surroundings.

Because we are studying the deformation of matter perturbations rather than the growth of those perturbation relative to a fixed grid, it is effectively a 2nd order perturbation expansion. But it also yields intuition about the nature of collapse.

Call $\mathbf{q}$ the comoving coordinate of a particle at the center of the perturbation, and $\mathbf{x}$ to be the proper coordinate of another nearby particle. We want to understand how $\mathbf{x}$ evolves in time. We can write, with full generality,

$$\mathbf{x}(t) = a(t)\mathbf{q} + b(t)\mathbf{f}(\mathbf{q}) \tag{1}$$

Here, the first term on the RHS represents the Hubble expansion of $\mathbf{x}$ relative to $\mathbf{q}$, and the second term represents the comoving deviation from Hubble flow, parameterized by some function $\mathbf{f(q)}$.

In the case of an initial configuration as an ellipsoid, Zel'dovich showed that the motion of each particle can be described by a diagonal "deformation tensor" (dx_i/dq_j; Longair 16.11).

$$D = |dx_i/dq_j| = a(t)\delta_{ij} + b(t)\partial x_i/\partial q_j$$

For a suitable choice of axes, $\mathbf{f(q)}$ can be represented by three constants related to the principal axes of the local ellipsoid: α, β, and γ. While these constants can be different for different perturbations, the Zel'dovich approximation states that $a(t)$ and $b(t)$ are the same for all particles. This is obvious for $a(t)$ in a homogeneous cosmology, but didn't have to be so for $b(t)$.

Diagonalizing the deformation tensor then yields the density evolution, described by

$$\rho(a - b\alpha)(a - b\beta)(a - b\gamma) = \bar{\rho}a^3 \tag{2}$$

which describes conservation of mass in the deforming ellipsoid.

Zel'dovich derived $b(t)$. For $\Omega = 1$, $a(t) = (t/t_0)^{2/3}$ and $b(t) = 0.4(t/t_0)^{4/3} = 0.4R^2(t)$, where $t_0 = 2/3H_0$ is the final time of collapse. Hence $b(t)$ describes the second-order perturbation to the expansion for the ellipsoidal volume. Essentially, $b(t)/a(t)$ represents the linear regime growth factor.

Depending on which of α, β, γ is largest, the density will then approach a singularity along that direction. Hence the ZA shows that collapse occurs first along one direction (into a planar configuration). These are called "Zel'dovich pancakes".

Once "shell crossing" occurs, i.e., when $a(t) - \alpha b(t) = 0$, the ZA density formally goes to infinity, and the ZA breaks down. Of course the density doesn't really go to ∞, since torques cause angular momentum that prevent singular collapse. Careful simulations have actually validated the ZA as extremely accurate until shell crossing.

Simulations of structure formation generate initial conditions using the ZA. (1) A power spectrum is generated given cosmological parameters. (2) At each k a Gaussian random number is thrown to determine the power on that scale. (3) The density perturbations are laid down (using random phases), which effectively determines α, β, γ for each particle. (4) Particles' positions ($\mathbf{x}(t)$) are evolved from a uniform grid using the ZA from $z \sim 1000$, until a time just before the first shell crossing in the volume. (5) The velocities can be easily computed from $\dot{\mathbf{x}} = \dot{b}(t)\mathbf{f}(\mathbf{q})$. Given the "initial" positions and velocities, the evolution is then followed numerically.

It is possible to extend the ZA and try to follow the particles into the weakly nonlinear ("quasi-linear") regime. This hasn't proved terribly insightful, but it has been done.

3 Spherical collapse

Let's skip over the messy pancake and filament stages and go straight to the end state: A halo which we will assume (for now) is spherical.

Intuition: The evolution of a spherical density perturbation is identical to the evolution of the Universe with a matter density equal to the density of the halo, i.e. a high-Ω universe!

If we have a spherical perturbation, Gauss's Law tells us we can ignore the matter outside the sphere, and that the mass interior is constant. So

$$\frac{d^2r}{dt^2} = \frac{-GM}{r^2}. \tag{3}$$

Integrating once we have

$$\dot{r}^2 = \frac{2GM}{r} + C, \tag{4}$$

i.e. conservation of energy. This ODE has a solution

$$r = A(1 - \cos\theta) \tag{5}$$
$$t = B(\theta - \sin\theta) \tag{6}$$
$$A^3 = GMB^2 \tag{7}$$

where $\theta = [0, 2\pi)$ is a parametric time variable, and $C = -A^2/B^2$. This is a cycloid. Since $C < 0$, the system is bound (kinetic<potential). This is identical to the solution for the evolution of the scale factor in a closed Universe.

Let's study the behavior of this system at early times. Initially, the perturbation expands with Hubble flow. For $\theta \to 0$, $r = A\theta^2/2$ and $t = B\theta^3/6$. Hence $\theta^6 = 8r^3/A^3 = 36t^2/B^2$, or $r^3 = (9/2)GMt^2$. Now $r^3 = 3M/4\pi\rho$, so we get $6\pi G\rho = t^{-2}$.

To relate this to overall cosmic expansion, recall that $H^2 = 8\pi G\rho/3$, so then $6\pi G\rho = (9/4)H^2$. Hence we get $(9/4)H^2 = t^{-2}$, or $t = 2/3H$. This is exactly the the time evolution of an $\Omega = 1$ universe! So at early times (when θ is small), our spherical model evolves like an $\Omega = 1$ universe. This is interesting, but should not be surprising.

As we move forward in time (or θ), we need to expand to higher order.

$$r = A\theta^2/2(1 - \theta^2/12) \tag{8}$$
$$t = B\theta^3/6(1 - \theta^2/20) \tag{9}$$

This takes some algebra but can be written as

$$r = \frac{A}{2}\Big(\frac{6t}{B}\Big)^{2/3}\Big[1 \mp \frac{1}{20}\Big(\frac{6t}{B}\Big)^{2/3}\Big] \tag{10}$$

The top sign is for the cycloid, bottom is for hyperbolic. The RHS first term is the first order expansion as before, and the next term represents the growth of the density enhancement.

The initial mass of the system is $M = \frac{4\pi}{3}\bar{\rho}r^3$. If the density is enhanced by an overdensity δ, the radius must shrink (by δr) in order to conserve enclosed mass:

$$M = \frac{4\pi}{3}\bar{\rho}r^3(1+\delta)(1+\delta r)^3. \tag{11}$$

Equating the initial and final masses gives $(1+\delta)(1+\delta r)^3 = 1$. Expanding to first order then gives

$$\delta \approx -3\delta r = \pm\frac{3}{20}\left(\frac{6t}{B}\right)^{2/3}. \tag{12}$$

where we take δr from the second term in the above equation. Note that once again we have $\delta \propto t^{2/3}$, as in linear theory.

We can use these formula to quantify some key events in the perturbation's history. First it is growing with Hubble expansion. Then it breaks away owing to its self-gravity, and reaches a maximum expansion at $\theta = \pi$; i.e. $r = 2A, t = \pi B$. Final collapse occurs when $\theta = 2\pi$, $r = 0, t = 2\pi B$.

We can therefore estimate the overdensity at turnaround and collapse:

$$\delta_{\rm turnaround} = (3/20)(6\pi)^{2/3} = 1.06 \tag{13}$$

$$\delta_{\rm collapse} = (3/20)(12\pi)^{2/3} = 1.69 \tag{14}$$

Hence shortly after a perturbation's overdensity exceeds unity, it turns around and begins its contraction. The collapse overdensity is 1.69, but of course this is spurious – in our perfectly spherically symmetric, pressureless model the overdensity at collapse is actually ∞, but our linear regime extrapolation yields 1.69. This number will still turn out to be useful.

Note that all the shells within our assumed tophat spherical perturbation behave homologously; there are no shell crossings, and all shells of matter turn around and collapse at the same time.

4 Virialized halos

In practice, the collapse gets halted well before singularity by a process known as "virialization". Nearby LSS torques up the matter distribution so that it contains a net angular momentum. Because dark matter can't release its gravitational potential energy, it obtains a velocity that takes it around the center, and eventually equipartitions its energy with the rest of the matter through dynamical friction, resulting in a pressure-supported (and slightly rotating) virialized halo.

Simulations show that the end state of virialization is a halo with a centrally-concentrated mass distribution.

Since galaxy rotation curves are flat, it used to be that one often postulated $\rho \propto r^{-2}$. This gives $M(r) \propto r$, and $v^2 \propto GM/r$ =const.

But simulations tend to show a different profile, first given by Navarro Frenk and White (1997):

$$\rho \propto \frac{1}{r(r+r_s)^2} \tag{15}$$

where r_s is the scale radius. The NFW profile is characterized by two parameters, an overall normalization (set by the halo mass) and the "concentration" c, roughly defined as the ratio of the virial radius to the core radius.

There is much debate in the literature regarding the NFW profile, particularly its inner slope. But CDM simulatations generically predict a cuspy inner profile, with $\rho \propto r^{1-1.5}$ as $r \to 0$, with the exact slope a matter of much bickering. Observations, at face value, indicate $\rho \propto r^{\sim 0-1}$, though there is much debate about that as well.

Also note that the NFW profile is only true for relaxed, undisturbed halos; halos with recent merger activity can deviate substantially from this profile. The reason why simulations produce an NFW-like profile is not understood.

4.1 Virial overdensity:

An important aspect of spherical collapse is the final overdensity reached by the collapse. While perturbation theory cannot give us this highly nonlinear endstate, one can still determine this from energetic arguments.

First let us determine the density at virialization. At turnaround the kinetic energy is zero, and $V = E$. The virial theorem says the final state has $V = -2T$, so $E = T + V = -T$, which gives $V = 2E$. So the potential energy is doubled from turnaround to virialization (assuming energy conservation), which means the radius of the system must be halved. Now, the turnaround radius is $r = 2A$, so the final virialized radius must be $r = A$. The density of this halo is therefore $\rho_h = 3M/4\pi A^3$.

To get the overdensity, we need the mean cosmological density at the time of collapse. For an $\Omega = 1$ universe, we previously derived $\bar{\rho} = (6\pi G t^2)^{-1}$. For t, we use the time to collapse, namely $t = 2\pi B$. Hence $\bar{\rho} = (24\pi^3 G B^2)^{-1}$. The overdensity at collapse is then

$$\delta \approx \rho_h/\bar{\rho} = \frac{72\pi^3 GMB^2}{4\pi A^3} = 18\pi^2 = 178 \tag{16}$$

So a collapsed halo will always have an average density within it that is roughly 200 times the cosmic mean density at that epoch!

The exact value depends on cosmology. For low-Ω universes it is larger, because the late-time mean cosmological density is lower than expected (i.e. more expansion).

In simulations it has been found that halo scaling properties are most easily understood when scaled to exactly 200 times the mean. So people often talk of the "virial radius" r_{200}, "virial mass" M_{200}, etc. These are all quantities that correspond to a spherical region around the halo center that encompasses a mean density of 200 times the mean density at that epoch.

4.2 Virial velocity dispersion:

Since matter in the halo is pressure supported, it has a line of sight (1D) velocity dispersion σ. The kinetic energy is then given by $T = 3M\sigma^2/2$. Now in virial equilibrium, the KE should be half the PE, or equal to the PE at turnaround. Hence $3M\sigma^2/2 = GM^2/2A$. Using $A = (GMB^2)^{1/3}$, and $B = t/2\pi$ at collapse, we get

$$\sigma^2 = \frac{1}{3}\Big(\frac{2\pi GM}{t}\Big)^{2/3} \tag{17}$$

Taking $t = 2/3H$ as the collapse time, and $M = 4\pi\rho r_0^3/3 = H^2 r_0^3/2G$, we get

$$\begin{aligned} \sigma^2 &= \frac{1}{3}\Big(\frac{\pi H^2 r_0^3}{(2/3H)}\Big)^{2/3} && (18)\\ &= \frac{1}{3}\Big(\frac{3\pi}{2}\Big)^{2/3}(Hr_0)^2 && (19)\\ &\approx (Hr_0)^2 && (20) \end{aligned}$$

Hence the 1-D velocity dispersion of a collapsed object is simply the Hubble flow velocity across the radius of the initial perturbation! Expressing it in terms of the virial radius $r_{200} = r_0/200^{1/3}$, we get

$$\sigma \approx 5.5 H r_{200} \tag{21}$$

This depends somewhat on cosmology, since in general $t \neq 2/3H$. Note also that means that for a constant mass, velocity dispersion of the halo will be different. Work out $\sigma \sim (1+z)^{1/2}$. In general, high-redshift halos are smaller, and have larger velocity dispersion, for the same mass, because the turnaround time is shorter and the universe is denser.

Recap from previous lecture on spherical collapse. Perturbation will grow non-linear and then collapse to self-gravitating, virialized objects. It goes through three steps: (1) 1-D collapse, to Zeldovich pancakes. We can develop perturbative models, i.e., Zeldovich approximation, to describe this process. Density perturbation is triaxial. One axis will collapse first. (2) 1-D pancake will collapse to filaments; (3) finally, the highest density regions will go through quasi-spherical collapse to dark matter halos.

The problem of spherical collapse is similar to the evolution of a over-critical universe when it is linear or quasi-linear. We used this to work out and density scale for turnaround and viralization. We showed that when linear perturbation is 1.06, it will turn around and begin to collapse. After twice that time, when linear perturbation grows to 1.69, it will reach complete collapse and becomes a virialized object.

We also showed that characteristic density for a viralized halo is 178, or close to 200. This gives us the virial density and virial radius of an object. The central concept here is dark matter halo. Numerical simulations show that they have a universal NFW profile. Today, we will study the distribution function of such halos.

5 Press-Schechter Theory

Halos provide our major conceptual unit for the deeply non-linear regime. These lumps of dark matter host the formation of galaxies through the condensation of baryons within them.

Now that we have studied the behavior of individual halos, we can ask the question, is it possible to determine the mass spectrum (or "mass function") of halos from cosmological considerations? Amazingly, it is. This was first done by Press & Schechter (1976)

Press-Schechter theory is an analytic model for the evolution of the halo mass function. Its derivation is far from rigorous, yet the results have been shown to be remarkably accurate (at least to within $\sim \times 2$). P-S (or its extensions) is unquestionably the most used analytic formula in cosmological galaxy formation theory. Efforts to make P-S more rigorous and/or more accurate has been a cottage industry for over 30 years, and certainly some progress has been made, but P-S still yields a great deal of insight from relatively simple considerations.

Here's the basic scheme: Imagine that we have a region of space with mass M that is collapsing. This mass can be connected with a particular comoving length scale (i.e. r or k) in the initial density field by $M = 4\pi\rho_0 r^3$ (ρ_0 is the cosmic mean density), plus $k = 2\pi/r$. Now consider the density *fluctuations* in spheres of mass M (i.e. in spherical tophats of radius r). Such fluctuations have an RMS value that we derived before:

$$\sigma_r^2 = \int \frac{k^2\, dk}{2\pi^2} P(k) \left(\frac{3\sin(kr) - 3kr\cos(kr)}{(kr)^3} \right)^2$$

The idea of Press-Schechter is that halos form out of peaks in the matter fluctuations. In the linear-theory spherical collapse model, the density at collapse is $1.69\rho_0$. Press-Schechter makes the ansatz that linear theory is correct until the density reaches this magic value, and then it suddenly collapses into a halo ($\rho \sim 200\rho_0$). Though seemingly unphysical, this turns out to be a reasonable approximation since gravitational instability operates very quickly.

So at any given time, all regions that have a density of 1.69 will have collapsed to form a halo. Hence the fraction of mass that is in halos of mass $> M$ is given by the fraction of the Gaussian distribution of RMS σ_r that exceeds 1.69:

$$f(> M) = \frac{1}{\sqrt{2\pi}} \int_{1.69/\sigma_r}^{\infty} dx\, e^{-x^2/2}$$

The fraction of mass that is in halos between mass M and $M + dM$ is given by df/dM. For convenience, define $\nu(M) \equiv 1.69/\sigma_r$. Then

$$\frac{df}{dM} = \frac{1}{\sqrt{2\pi}} \frac{dx}{dM}\, e^{-x^2/2}\Big|_{x=\nu}^{x=\infty} = \frac{1}{\sqrt{2\pi}} \frac{d\nu}{dM} e^{-\nu^2/2}$$

The number density of such halos is the number density of all halos (ρ_0/M) times the fraction of mass in halos from $M \to M + dM$:

$$\frac{dn}{dM} = \frac{\rho_0}{M}\frac{df}{dM} = \frac{\rho_0}{M}\frac{1}{\sqrt{2\pi}} e^{-\nu^2/2} \frac{d\nu}{dM}$$

Now we substitute $\frac{d\log\nu}{d\log M} = \frac{M}{\nu}\frac{d\nu}{dM}$, so

$$\frac{1}{M}\frac{dn}{dM} = \frac{dn}{d\log\, M} = \frac{\rho_0}{M}\frac{1}{\sqrt{2\pi}} \nu e^{-\nu^2/2} \left(\frac{d\log\nu}{d\log\, M} \right)$$

The term $d\log\nu/d\log\, M$ is less frightening than it looks. Over a sufficiently small range in k, $P(k)$ is roughly a power law: $P \propto k^n$ (recall $n \approx -2$ on galaxy scales). Hence

$$\sigma_r^2 = \int \frac{k^{2+n}\, dk}{2\pi^2} \left(\frac{3\sin(kr) - 3kr\cos(kr)}{(kr)^3} \right)^2$$

The term in parantheses is around unity when $kr \ll 1$ and vanishes fo $kr \gg 1$. So we can roughly estimate it as a step function: Unity up to $k_{\max} = 1/r$, and zero for larger k. In that case,

$$\sigma_r^2 = \int_0^{1/r} \frac{k^{2+n}\, dk}{2\pi^2} = \frac{1}{2\pi^2}\frac{1}{r^{n+3}}$$

So $\sigma_r^2 \propto r^{-(n+3)}$, so $\sigma_r \propto M^{-(n+3)/6}$ (using $M \propto r^3$), which makes $\nu = 1.69/\sigma_r \propto M^{(n+3)/6}$. Hence, the logarithmic derivative $d\log\nu/d\log M = (n+3)/6$ where n is the effective logarithmic slope of the power spectrum at a mass scale M.

Now for the last swindle. In purely cold dark matter, most of the mass is within halos, since even the smallest fluctuations will permit collapse. However, the P-S derivation has only half of the mass in halos, because $F(0) = 1/2$; the negative part of the Gaussian has been left out since it corresponds to underdense regions. The swindle is to simply multiply $dn/d\log M$ by a factor of 2! Hence, our final answer is

$$\frac{dn}{d\log(M)} = \frac{\rho_0}{M}\sqrt{\frac{2}{\pi}}\left(\frac{n+3}{6}\right)\nu e^{-\nu^2/2}$$

The physical reason why this swindle works is that, once things go nonlinear, the collapse is able to draw matter in from less dense regions, resulting in a log-normal matter density distribution. Hence much of the mass does end up in the overdense regions.

Since $\nu \propto M^{(n+3)/6}$, let us introduce a parameter called M_* such that $\nu = (M/M_*)^{(n+3)/6}$. Then

$$\frac{dn}{dM} = \frac{\rho_0}{M^2}\sqrt{\frac{2}{\pi}}\left(\frac{n+3}{6}\right)\left(\frac{M}{M_*}\right)^{(n+3)/6}\exp\left[-\left(\frac{M}{M_*}\right)^{(n+3)/3}\right]$$

This form should be familiar – it's a Schechter function!

The most important things to note about the P-S formula are the limits for small and large M.
(1) For $M \gg M_*$, $n > -1$, and the mass function cuts off exponentially.
(2) For $M \ll M_*$, $n \to -3$, so the mass function goes as M^{-2}. Of course if it was exactly $n = -3$, it would be zero, but in practice it never gets close enough for the $(n+3)/6$ term to matter.

Figure 16.4.

Figure 16.5. How well it worked.

6 Press-Schechter Extensions

Remarkably, P-S works over virtually all mass scales from dwarf galaxies to clusters, to better than a factor of two. However, in detail P-S tends to systematically underpredict large-mass halos and overpredict small-mass ones.

Extended Press-Schechter (EPS): Bond et al (1991) used an excursion set formalism to statistically estimate how many small halos would be subsumed into larger ones (and therefore be effectively uncountable). This reduced the number of small-mass halos in P-S and increased the number of large-mass ones, thereby bettering the agreement with simulations. Furthermore, EPS allows characterization of the merging rate of dark matter halos.

Sheth-Tormen: Assumed halos were elliptical instead of spherical, in a way that depended on the shear from the surrounding environment (Sheth, Mo, Tormen 2001). This alters the collapse a bit. People call this the Sheth-Tormen (1999) mass function because the empirical fit was presented before the analytic paper. Reed et al (2003) showed that S-T provides an excellent fit to numerical simulations (also Jenkins etal 2001). The S-T mass function is given by:

$$f(\sigma_r) = A\sqrt{\frac{2a}{\pi}}\left[1+\left(\frac{\sigma_r^2}{a\delta_c^2}\right)^p\right]\frac{\delta_c}{\sigma_r}\exp-\frac{a\delta_c^2}{2\sigma_r^2},$$

where $\delta_c = 1.686$, $A = 0.3222$, $a = 0.707$, and $p = 0.3$. Today, S-T is the most commonly used analytic halo mass function.

Figure; SDSS cluster

Figure: comparing with Jenkins

Figure 16.6.

Press-Schechter illustrates important aspects of hierarchical cluster models.

- small halo appears first. Galaxies with $M \sim 10^12$ won't show up until $z \sim 4$.
- galaxy mass objects began to form at $z \sim 10$.
- cosmological dependence.

Figure: Bahcall et al.

7 Problem with P-S mass function

It would be tidy indeed if galaxies mapped straightforwardly onto halos in such a way that L_* in the Schechter luminosity function corresponded to M_* in the P-S mass function. Unfortunately, this is not the case; L_* galaxies today have halo masses well below M_*. Furthermore, the faint-end slope of the galaxy luminosity function is significantly shallower that -2. Hence the process of galaxy formation is just a bit more complex that such a simple one-to-one mapping.

White figure.

Problems at both high and low mass end.

High-end: feedback?

The low-mass slope of the mass function M^{-2} is much steeper than what is observed in the luminosity function of galaxies. Indeed, this discrepency extends to fairly high masses, as much as the LMC in certain contexts. This could be reconciled in many interesting ways:
1) low-mass halos are inefficient in forming stars and are therefore underluminous, for many possible reasons:
1a) photoheating from external UV evaporates the gas,
1b) supernovae from internal star formation pushes the gas out,
1c) the objects get torn apart by encounters with larger galaxies,
2) the universe forms fewer low mass halos because of some deviation from the standard cold dark matter scheme: WDM, interactions, decays, quantum exclusion.

This brings us to the question of how to make real galaxies out of dark matter halo. So far we have only considered dark matter particles, which are collisionless. In contrast, the baryonic component of galaxies, will radiate. The fact that you can see it means that baryons are losing energy by radiation from stars and ISM. This is a dissipative process, in the sense that the baryonic matter can lose thermal energy, and therefore the total energy, and collapse further.

8 The Role of Dissipation

Dissipative processes play a dominant role in the formation and evolution of stars. A star can only be formed if the collapsing protostellar cloud can get rid of its binding energy. The best way is by radiative cooling. This process continues until the cloud becomes optically thick to its own radiation. The loss of binding energy is then mediated by the dust grain which will still be optically thin.

Note that there is then a key difference between normal star formation and the formation of the first stars. There is no dust grain to cool. There are other differences which we should come back later.

Dissipation will also play a key role in the formation of the entire galaxy. The theoretical frame work was first worked out by Rees, Ostriker, Silk, etc., and was highlighted in one of the most influential papers by George Blunmenthal, Sandy Faber, Joe Primack and Martin Rees, in 1984, entitled Formation of galaxies and large-scale structure with cold dark matter.

Figure 16.2 shows the cooling curve as a function of temperature and metallicity. You will visit this again in ISM class. The cooling rate is

$$dE/dt = -N^2\Lambda(T),$$

where N is the number density and Λ is the cooling function. Square dependence is that most of the cooling process are two body processes, collision or recombination. In the absence of metal, the dominant loss mechanism at high temperature is thermal bremsstrahlung, with energy losing rate $\sim N^2T^{1/2}$. At lower temperature, the main loss mechanisms are helium at $T \sim 10^5$, and f-b and b-b of atomic hydrogen at $T \sim 10^4$. Note two things:

- dependence on metallicity
- quickly drops towards low T

Note that for galaxy context, what kind of temperature we are referring too. This is basically the virial temperature of the halo. Without any heating (star formation) and cooling, this is the temperature that the gas particle is going to be. So what we just said is that for small halos, which formed early, and had very low T, and with no metal, cooling is hard. First galaxies will have a hard time forming stars, and probably the star formed there was big. Will come back later.

Now let's work out some timescales. We can define cooling time as the time it takes for the plasma to radiate away all its energy:

$$t_{cool} = \frac{E}{|dE/dt|} = \frac{3NkT}{N^2\Lambda(T)},$$

this timescale can be compared with the timescale for gravitational collapse

$$t_{dyn} \sim (G\rho)^{-1/2} \sim N^{-1/2}.$$

Figure 16.3 shows the locus of the equality $t_{cool} = t_{dyn}$ in a temperature-number density diagram. Inside this locus, the cooling time is shorter than the collapse time, so it is expected that dissipative processes are more important than dynamical processes in determining the behavior of the baryonic matter. There is also a line showing where object has dynamical time longer than Hubble time in which case it won't collapse at all. So it can be seen clearly that the range of masses which lie within the critical locus and which can cool in 10^{10} years. This is the mass range of 10^6 to 10^{12} solar masses, exactly what the mass of our normal galaxy. Note at higher mass, you have your clusters which won't be able to cool. That's why the largest galaxies have 10^{13} solar mass.

This is an important conclusion. Nothing in our dark matter analysis will give us the characteristic mass of galaxy. It is not a determined by power spectrum, but by the cooling process.