5-18-2009

Large Scale Structure and Galaxies

Ravi K. Sheth
University of Pennsylvania, shethrk@sas.upenn.edu

Large Scale Structure and Galaxies

Abstract

These notes sketch the motivation for and ingredients of the Halo Model of nonlinear and biased structures in the Universe. A key part of this approach is the relation between halo abundances and their large scale clustering. These come from the excursion set approach, so I have taken the opportunity to collect together all the formulae associated with this approach into one place. These include expressions for: the unconditional mass function, the conditional mass function, the environmental dependence of the mass function, halo bias, merger rates, creation anddestruction rates, the distribution of half-mass assembly times, masses and mass at fixed assembly time. In addition, I discuss how the approach can be used to describe voids, filaments and sheets, as well as the nonlinear counts in cells distribution, and provide analytic formulae for a number of these statistics.

Together these formulae show that, in hierarchical models: massive halos assemble their mass later than low mass halos; halos which assemble their mass abnormally late for their mass will tend to have experienced a recent major merger; if one is interested in the mass assembled in pieces which are above some mininum mass, then this happens earlier for the more massive halos; for similar reasons, the mass fraction in pieces which are between a fixed mass range reaches a maximum at higher redshifts for halos which are more massive today. The first trend may explain why the oldest stars tend to sit in massive objects; the second may be why star formation in massive objects ended earlier. This approach also shows that the mass function in dense regions should be 'top-heavy', and that more massive halos should be more strongly clustered. If galaxy properties are determined primarily by the mass of their parent halo, then many observed correlations with environment are a simple consequence of these trends.

Finally, I summarize the Halo Model of galaxy clustering. I discuss how it describes typedependent clustering, particularly dependence on luminosity and color, and sketch how to use it to build accurate mock catalogs which include information about stellar mass, dust, and star formation history.

Keywords

cosmology, dark matter, dark energy, large scale structures, galaxy formation

Disciplines

Physical Sciences and Mathematics | Physics

Comments

Suggested Citation:
Sheth, R.K. (2009). "Large Scale Structure and Galaxies." *AIP Conf. Proc.* 1132, 158 (2009).

Large Scale Structure and Galaxies

Ravi K. Sheth

Center for Particle Cosmology, University of Pennsylvania,
209 S 33rd Street, Philadelphia, PA 19104

Abstract. These notes sketch the motivation for and ingredients of the Halo Model of nonlinear and biased structures in the Universe. A key part of this approach is the relation between halo abundances and their large scale clustering. These come from the excursion set approach, so I have taken the opportunity to collect together all the formulae associated with this approach into one place. These include expressions for: the unconditional mass function, the conditional mass function, the environmental dependence of the mass function, halo bias, merger rates, creation and destruction rates, the distribution of half-mass assembly times, masses and mass at fixed assembly time. In addition, I discuss how the approach can be used to describe voids, filaments and sheets, as well as the nonlinear counts in cells distribution, and provide analytic formulae for a number of these statistics.

Together these formulae show that, in hierarchical models: massive halos assemble their mass later than low mass halos; halos which assemble their mass abnormally late for their mass will tend to have experienced a recent major merger; if one is interested in the mass assembled in pieces which are above some mininum mass, then this happens earlier for the more massive halos; for similar reasons, the mass fraction in pieces which are between a fixed mass range reaches a maximum at higher redshifts for halos which are more massive today. The first trend may explain why the oldest stars tend to sit in massive objects; the second may be why star formation in massive objects ended earlier. This approach also shows that the mass function in dense regions should be 'top-heavy', and that more massive halos should be more strongly clustered. If galaxy properties are determined primarily by the mass of their parent halo, then many observed correlations with environment are a simple consequence of these trends.

Finally, I summarize the Halo Model of galaxy clustering. I discuss how it describes type-dependent clustering, particularly dependence on luminosity and color, and sketch how to use it to build accurate mock catalogs which include information about stellar mass, dust, and star formation history.

Keywords: cosmology – dark matter – dark energy – large scale structures – galaxy formation
PACS: 98.65.Dx

OBSERVATIONS AND MOTIVATION

The next decade will be the age of precision cosmology. Much of this precision will come from surveys of objects for which gastrophysics has been important – supernovae, galaxies and galaxy clusters. These surveys follow-on from where the SDSS left-off, with the SDSS itself being the most recent in long and fruitful history of survey astronomy. The optimists argue that we will constrain cosmological models in which baryons are thought to make up less than 10 percent of the total mass-energy budget, to one percent precision, despite the fact that most of our observations are of baryons, and our understanding of the associated gastrophysics is nowhere near 10 percent. The following notes lay out the basis for this optimism.

We have known for just under a century that ours is but one of many galaxies. We

CP1132, *Cosmology and Gravitation: XIII Brazilian School on Cosmology and Gravitation*
edited by M. Novello and S. E. Perez Bergliaffa

have known for 80 years that galaxies are clustered, for 40 years that this clustering signal is almost a power law, and for about 30 years that not all galaxies cluster similarly. Departures from a power law are now routinely measured: these depend on galaxy type, and type-dependent clustering now provides important insights into galaxy formation.

Since different galaxy types are differently biased tracers of the underlying mass distribution, the question arises as to how one can develop a unified statistical language for describing different point processes which all arise from the same underying density field. This language is known as the Halo Model [13]. It provides a unified framework for relating galaxies and galaxy clusters to the underlying nonlinear dark matter – it is the language in which a nonlinear biased description of the dark matter is most easily discussed.

Before showing how this model is built, it is worth making the following point explicitly: Discussions of galaxy formation generally fall into two types, those in which smoothed density, pressure and temperature fields are thought to be important, and those in which discrete objects, so-called dark matter halos, are the fundmental units. The Halo Model, in its current implementation, makes one further assumption – that the mass of these units is the most important parameter. This approach has shown that a description based on mass rather than smoothed density is by far more efficient and effective when discussing nonlinear structures: the Halo Model can be used to predict the smoothed density, pressure and temperature fields, whereas the opposite has yet to be done. (This is analogous to the choice between coordinate and Fourier-space basis of Gaussian random fields: although both are equivalent, the effects of smoothing, quasi-linear evolution, etc. are much easier in the Fourier description.) The physical reason for this is that, at late times, the sizes of the fundamental units are small compared to their typical separations. Thus, the Halo Model has been successful for describing observations of clusters and galaxies, but there is at yet no halo model description of the Lyman-α forest.

To illustrate this point, the right-hand panel of Figure 1 shows the correlation function (the fractional excess of pairs over random) of galaxies in the SDSS. The filled circles show the full sample (for the specialists, these are all galaxies above some luminosity in a volume limited catalog). The other symbols show the result of measuring the clustering signal in subsamples of this one, made by selecting galaxies based on the number of neighbours within some fixed (projected) distance (in this case $8h^{-1}$ Mpc). The open triangles show that the 10% in the densest regions cluster more strongly than if this cut is relaxed to include the upper 30% of the objects (filled triangles). In turn, these cluster more strongly than the full sample. Except on very small scales, the 30% in the least dense regions are even less clustered (filled squares), but making a more extreme cut, so that only the densest 10% are included, results in stronger clustering (open squares). Thus, clustering is *not* a monotonic function of environment!

These trends are well reproduced in the panel on the right, which shows measurements in a dark-matter only simulation that was turned into a mock SDSS catalog using a Halo Model motivated approach. The catalog was tuned to reproduce the filled circles; all the others are predictions or tests of the approach. Figure 2 shows that, in addition to getting the non-monotonicity of the signal right, it accurately reproduces all the bumps and wiggles seen in the data.

The notes which follow are intended to show why measurements like these will soon provide excellent constraints on generic predictions of hierarchical models: the shift in

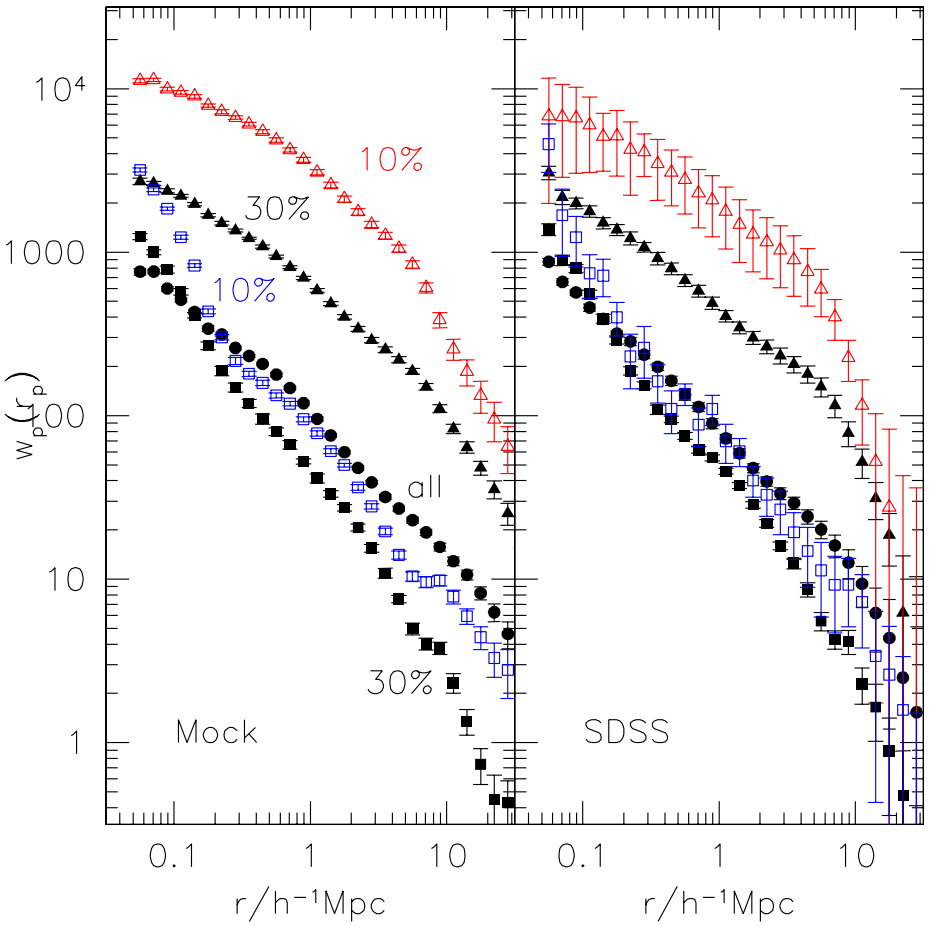

FIGURE 1. Environmental dependence of clustering in the SDSS (right) and in a mock catalog (left) which was based on ingredients from the Halo Model [from 1].

amplitude on scales above $10h^{-1}$Mpc is sensitive to the fact that the mass function of halos in dense regions is 'top-heavy'; the jump on scales $\sim 0.2h^{1}$Mpc, and the tendency for this to happen on smaller scales in the underdense regions is another manifestation of this, because the Halo Model says this feature marks the virial radii of halos (halos are expected to have the same density whatever their mass, so virial radii are predicted to increase as the one-third power of halo mass); the non-monotonic behaviour with large scale environment arises naturally if the initial conditions were Gaussian, although the small scale signal is also sensitive to halo concentrations, and hence their formation histories. And finally, the fact that the simulations appear to slightly overpredict the

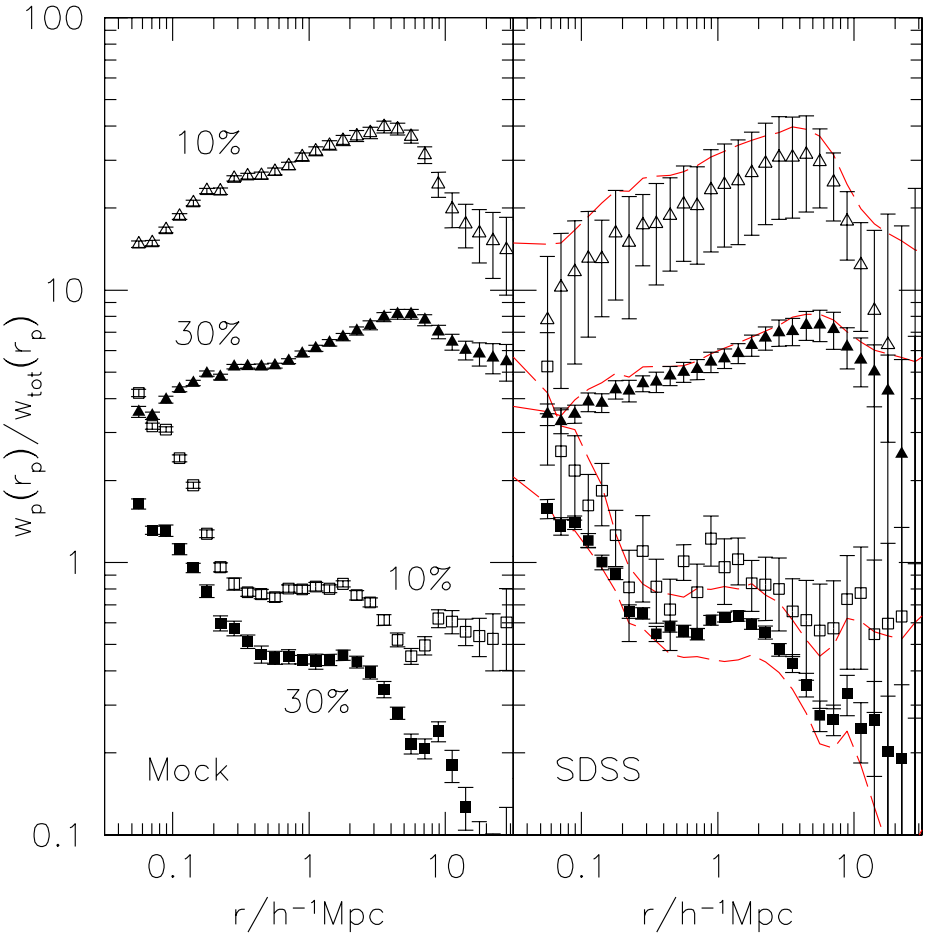

FIGURE 2. Same as previous figure, but now all symbols and curves have been normalized by those for the full sample [from 1].

trends seen in the data suggests that the simulations assumed too large a value for the amplitude of the initial fluctuations (known in the jargon as σ_8).

The notes assume that the linear theory of gravitational instability is familiar: the linear theory growth factor for the overdensity will be written as $D(t)$; that for the potential as $D(t)/a(t)$, where a is the expansion factor. Explicit expressions for $D(t)$, and of the evolution of the background cosmological parameters Ω and Λ may be found in textbooks [35, 37**?**] or review articles [6, 13]. The linear theory power spectrum of the density fluctuation fi eld at time t is $P_L(k,t)$. When the fi eld is smoothed with a fi lter

of comoving scale R, then the spectral moments play a special role. These are

$$\sigma_j^2(R,t) \equiv \int \frac{\mathrm{d}k}{k} \frac{k^3 P_{\rm L}(k,t)}{2\pi^2} k^{2j} |W(kR)|^2, \quad \text{and we define} \quad S(R,t) \equiv \sigma_0^2(R,t) \tag{1}$$

because the case $j = 0$ is particularly important.

NONLINEAR EVOLUTION

One of the standard predictions of nonlinear hierarchical structure formation models is the abundance of virialized structures [38, 54, 20]. Simulations show that this abundance depends on the large scale environment: the ratio of massive to low mass objects is larger in dense regions [e.g., 15]. Recent measurements in galaxy surveys appear to bear this out: the virial radii of objects in underdense regions are smaller, consistent with their having smaller masses (Figures 1 and 2). The following section uses the spherical evolution model to show why this happens. Although much of this is standard, I have added some discussion of what changes if the gravitational force law is modified from an inverse square.

Spherical evolution: Collapse and expansion

The spherical evolution model describes the evolution of the size R of a spherical region in an expanding universe [19, 40, 37, 35, 6]. Since realistic structures are neither spherical nor smooth, that it works at all is because it is, at heart, a statement of the constraints imposed by mass and energy conservation.

The model begins by stating that $F = ma$, so

$$\frac{\mathrm{d}^2 R}{\mathrm{d}t^2} = -\frac{GM(<R)}{R^2} + \frac{\Lambda}{3} R, \tag{2}$$

where $M = 4\pi R_i^3 \bar{\rho}(t_i)(1+\delta_i)$ is the mass enclosed by the perturbation, and Λ is the cosmological constant (which we assume is constant in space and time). Models with evolving dark energy will have $\Lambda(t)$, and if the dark energy clusters, then $\Lambda(t)[1+\lambda(t)]$, where λ is the fluctuation, but we will not consider these here.

Multiplying both sides of this expression by $2\,dR/dt$ yields

$$\frac{\mathrm{d}\,(dR/dt)^2}{\mathrm{d}t} = 2\frac{dR}{dt}\left[-\frac{GM(<R)}{R^2} + \frac{\Lambda}{3} R\right]. \tag{3}$$

Multiplying both sides by dt and then integrating once (recall $M(<R)$ is constant) yields

$$\left(\frac{\mathrm{d}R}{\mathrm{d}t}\right)^2 = \frac{2GM}{R} + \frac{\Lambda}{3} R^2 - E_i, \tag{4}$$

where E_i is the constant of integration. One way to set this constant is by requiring that the initial velocity and density perturbations satisfy linear theory:

$$\left(\frac{\mathrm{d}R}{\mathrm{d}t}\right)_i^2 = (H_i R_i)^2(1-\delta_i/3)^2 = \frac{2GM}{R_i} + \frac{\Lambda}{3}R_i^2 - E_i, \tag{5}$$

If the perturbation is sufficiently dense initially, it will reach a maximum size before turning around and collapsing. At turnaround, $dR/dt = 0$, so

$$\frac{2GM}{R_{ta}} + \frac{\Lambda}{3}R_{ta}^2 = E_i = \frac{2GM}{R_i} + \frac{\Lambda}{3}R_i^2 - (H_i R_i)^2(1-\delta_i/3)^2 \tag{6}$$

Dividing throughout by $(H_i R_i)^2$, and recalling that $M \propto R_i^3$ shows that this is a cubic equation for R_{ta}/R_i, so it can be solved analytically. Note in particular that R_i/R_{ta} depends on δ_i and the background cosmology, but that this dependence is the *same* for all R_i.

To get a feel for the solution, suppose $\Lambda = 0$. Then, because $2GM/R_i = \rho_i(1+\delta_i)(8\pi G/3H_i^2)(H_i R_i)^2 = \Omega_{mi}(1+\delta_i)(H_i R_i)^2$ this becomes

$$\frac{R_i}{R_{ta}} = 1 - \frac{(1-\delta_i/3)^2}{\Omega_{mi}(1+\delta_i)} \approx 1 - \frac{1-5\delta_i/3}{\Omega_{mi}} \tag{7}$$

Since $\Omega_{mi} \approx 1$ in most models, $R_{ta}/R_i \propto (5\delta_i/3)^{-1}$ decreases as δ_i increases. This shows that initially denser perturbations turnaround after fewer expansion factors, i.e., sooner, than less dense ones. In fact, the turnaround time can be got from the fact that

$$t_{ta} - t_i = \int_{R_i}^{R_{ta}} \frac{dR}{dR/dt} = t_i \int_1^{R_{ta}/R_i} \frac{dR/R_i}{d(R/R_i)/d(t/t_i)} \tag{8}$$

where it is good approximation to set the lower limit to zero. The subsequent collapse takes the same amount of time, so the time to final collapse and virialization is just a factor of two times larger that that at turnaround. Similarly, the physical size of the final virialized object is about a factor of two smaller than at turnaround. This comes from energy conservation: at turnaround, all the energy is potential, whereas at virialization, $-W = 2K$ so the total energy is half the potential: $GM/R_{ta} = GM/2r_{vir}$. (The presence of dark energy modifies this slightly, but not substantially.)

The density at virialization is large – much larger than linear theory would predict. For example, in an Einstein de-Sitter universe, $(a_{vir}/a_{ta}) = (t_{vir}/t_{ta})^{2/3} = 2^{2/3}$, so the comoving density at virialization is $(R_i/a_i)^3/(R_{vir}/a_{vir})^3 = 2^3(R_i/a_i)^3(R_{ta}/a_{ta})^3(a_{vir}/a_{ta})^3 = 2^3 2^2 (R_i/a_i)^3/(R_{ta}/a_{ta})^3$; it has increased by a factor of 32 relative to the comoving density at turnaround. This is substantially more than the factor of $(a_{\rm vir}/a_{ta}) = 2^{2/3}$ one would predict from linear theory. In contrast, the ratio of the nonlinear potential to that in linear theory is

$$\frac{GM/R_{\rm vir}}{(a_i/a_{vir})(GM/R_i)} = \frac{R_i/a_i}{R_{\rm vir}/a_{vir}} = 2\frac{R_i/a_i}{R_{\rm ta}/a_{vir}} = \frac{10}{3}\frac{a_{vir}}{a_i}\,\delta_i. \tag{9}$$

This suggests that a description based on the potential rather than density fi elds will lead to promising results. We will not have space to explore this further, but note that the spherical model for nonlinear structure formation has generally emphasized the density, not the potential.

In general models with dark energy, one must solve numerically for the evolution of R_i/R as a function of t. Since linear theory makes a prediction for the linear growth, it is conventional to express $R_i/R(t)$ as a function of $D(t)\delta_i/D_i$. This relation is cumbersome, even in the simplest case of an Einstein de Sitter universe. However, it is rather well approximated by

$$\left(\frac{R_i/a_i}{R(t)/a(t)}\right)^3 \equiv 1+\Delta \approx \left(1-\frac{\delta_{\rm L}(t)}{\delta_{\rm sc}(t)}\right)^{-\delta_{\rm sc}(t)}, \quad \text{where} \quad \delta_{\rm L}(t) \equiv \frac{D(t)}{D_i}\,\delta_i, \qquad (10)$$

and $\delta_{sc}(t)$ is the critical density required for collapse at t evolved from t_i to t using linear theory. It happens that δ_{sc} depends very weakly on cosmology – it is 1.686 for $\Omega = 1$ and tends to 1.5 as $\Omega \to 0$, with nonzero Λ making only a small difference – so it is a very weak function of t. Infall speeds $v_{\rm pec}$ can be got by differentiating this expression with respect to t. Note that this expression is also accurate for underdensities, i.e., when $\Delta < 0$.

Environment as effective cosmology

It is an interesting exercise to show that, in the spherical evolution model, the growth of structure in an initially over- or underdense region is just like that in a universe with a different background cosmology. To correctly estimate the background cosmology associated with, say, an underdense void, one must account not only for the lower density, but for the fact that the effective Hubble constant of the void cosmology is larger than in the background [e.g., 18]. One way of thinking about the effective Hubble constant is that it ensures that the effective cosmology has the same age as the background cosmology. (The cosmological constant is, of course, constant, but when expressed in units of the critical density in the effective model, it is modifi ed because the critical density depends on the effective Hubble constant.) Because the spherical model does this automatically, the excursion set approach in the next section incorporates this self-consistently, without having to appeal to the concept of an effective cosmology [27]. This is a direct consequence of Birkhoff's theorem.

Modified gravity and Birkhoff's theorem

There are two special features of equation (2) which are peculiar to standard inverse-square-law gravity. The fi rst is that, of the two terms in it, one scales as R^{-2}, and the other as R. These are the only two force laws which produce stable closed orbits – so one wonders if theories which modify gravity should worry about this.

The second point is that, when solving for the evolution of $R(t)$, it was enough to study the evolution of the boundary of the perturbation: an initially tophat perturbation

remains so (until it has fully collapsed). This is a consequence of Birkhoff's theorem. Modifications to gravity typically mean that Birkhoff's theorem no longer applies. This complicates the spherical model, because now each shell must be evolved separately: Generically, a top hat perturbation will not remain a tophat. In addition, because such theories often introduce a scale beyond which gravity is modified, perturbations which never cross this scale don't know the difference. As a result, δ_{sc}, which was the same for all masses in standard gravity, becomes mass dependent. See [28] for the first analysis which incorporates these subtleties.

Typically, in these theories, even linear theory is modified in a rather profound way. Whereas the linear theory growth factor is the same function of time for all k modes in standard gravity, it is k-dependent in modified theories [e.g. 58, 59]. As a result, a smooth spherical region within which the density is the same as the background universe will evolve. This qualitatively different behavior from standard gravity has not been emphasized – so it is worth showing the argument explicitly.

Consider the density field smoothed on scale R at some early time t_i. We can write this field in terms of its Fourier modes and the (Fourier Transform of the) smoothing kernel as

$$\delta_R(\mathbf{x},t_i) = \int d\mathbf{k}\, \exp(i\mathbf{k}\cdot\mathbf{x})\, \delta(\mathbf{k})\, W(kR). \tag{11}$$

The linearly evolved field is

$$\delta_R(\mathbf{x},t) = \int d\mathbf{k}\, \exp(i\mathbf{k}\cdot\mathbf{x})\, \frac{D(k,t)}{D(k,t_i)}\, \delta(\mathbf{k})\, W(kR), \tag{12}$$

where D is the linear theory growth factor. In standard gravity, D is independent of k, so if $\delta_R(\mathbf{x},t_i) = 0$ then $\delta_R(\mathbf{x},t) = 0$ also. But if D depends on k, then if $\delta_R(\mathbf{x}) = 0$ at some time t, it will, in general, be non-zero at other times (the exception being if the k-dependence of W happens to exactly cancels that of D). Thus, we are led to the rather remarkable conclusion that, when the gravitational potential has been modified, then linear theory predicts that a spherical tophat patch within which the density is the same as the background will evolve! The reason why can be traced to the fact that Birkhoff's Theorem no longer applies once the Newtonian potential has been modified. Without this Theorem, the spherical top hat filter is no longer special, and our common sense prejudice from standard gravity – that initially overdense regions become denser, underdense regions less dense, but regions within which the density is the same as the background do not evolve – must be treated with caution.

Since the argument above is true for any R, one might wonder what happens in the limit of large R. If the filter removes modes on scales of order $kR > 1$ then a uniform average density patch will not evolve only if $D(k,t)$ becomes independent of k-dependent at small k. Else, linear theory would predict that inhomogeneities would arise even from a perfectly unperturbed universe. Perhaps the requirement of large scale homogeneity can be used constrain such modified gravity theories. In any case, the equivalence between environment and effective background cosmology, which is part of standard gravity, almost certainly breaks down in these modified theories.

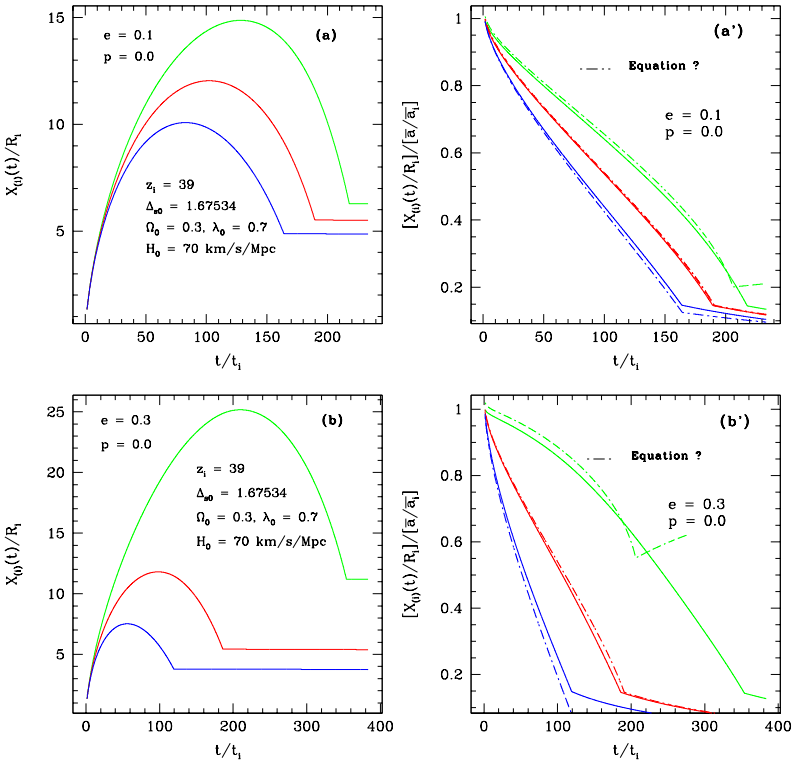

FIGURE 3. Physical (left) and comoving (right) evolution of axis lengths in the triaxial collapse model. The times at which the different axes freeze-out are determined by the initial values of (e, p, δ) and by the background cosmological model. Dot-dashed curves in the panels on the right show the simple analytic approximation of equation (15).

Triaxial evolution

The spherical cow approximation, while useful, is not realistic. To describe the evolution of non-spherical structures one needs a model of non-spherical collapse. While there is a long history of studies of triaxial collapse, the formulation of [10] is now generally adopted, because it reduces, at early times, to linear theory and the Zeldovich [69] approximation. There is also now general agreement that dark matter halos should be identifi ed with ellipsoids which have collapsed completely along all three principal axes [53].

In this framework, the time required to collapse depends on the overdensity δ of the initial patch *and* on the surrounding shear fi eld (Birkhoff's theorem is gone because the

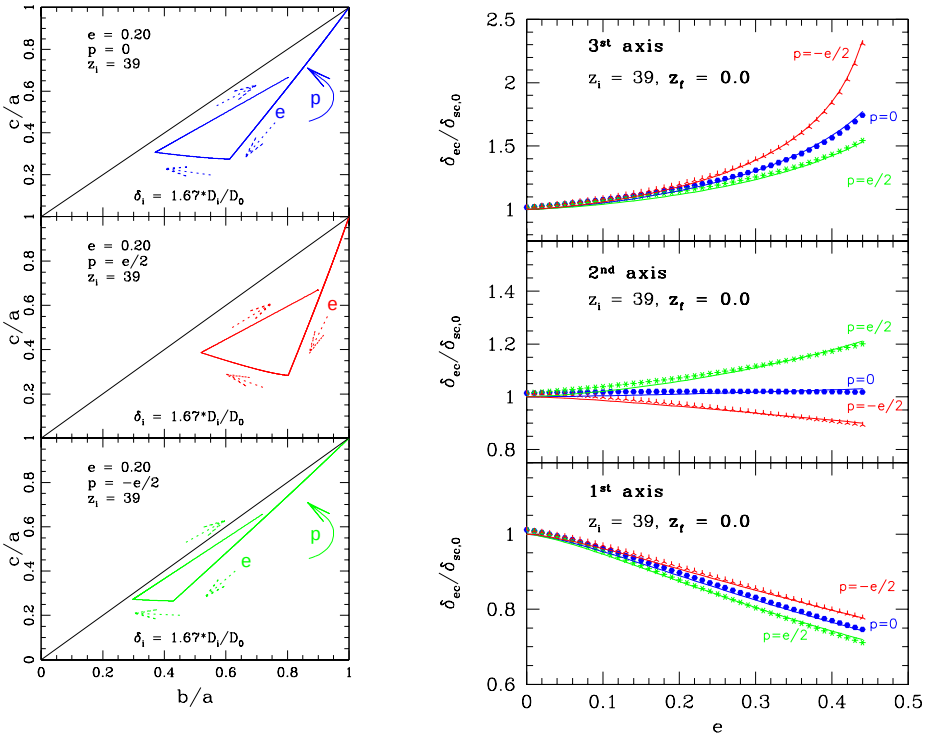

FIGURE 4. **Left:** Evolution in the axis ratio plane for a range of choices of e and p values. The value of p determines the angle of initial descent, and e determines the values of c/a and b/a when the shortest axis freezes out. The final axis ratios lie close to the line of initial descent. **Right:** Critical overdensity required for collapse along one, two, and three axes (bottom to top) at $z = 0$ in a ΛCDM model with $\Omega_0 = 0.3$ as a function of the initial shape parameters e and p.

spherical symmetry is gone!), parametrized by its ellipticity e and prolateness p. Here

$$e = \frac{\lambda_1 - \lambda_3}{2\delta_i}, \qquad p = \frac{\lambda_1 - 2\lambda_2 + \lambda_3}{2\delta_i} \qquad \text{and} \qquad \delta_i = \lambda_1 + \lambda_2 + \lambda_3, \tag{13}$$

where the λ_j are the eigenvalues of the initial deformation tensor, which itself is made up of second derivatives of the initial potential fi eld. The sum of the eigenvalues is the trace, and so the setting of $\delta_j \equiv \sum_j \lambda_j$ is really just Poisson's equation.

As happens for the spherical model, the exact evolution of a triaxial perturbation must be solved numerically, but the following approximation turns out to be quite accurate. Start by considering the nonlinear density in the Zeldovich approximation (this assumes

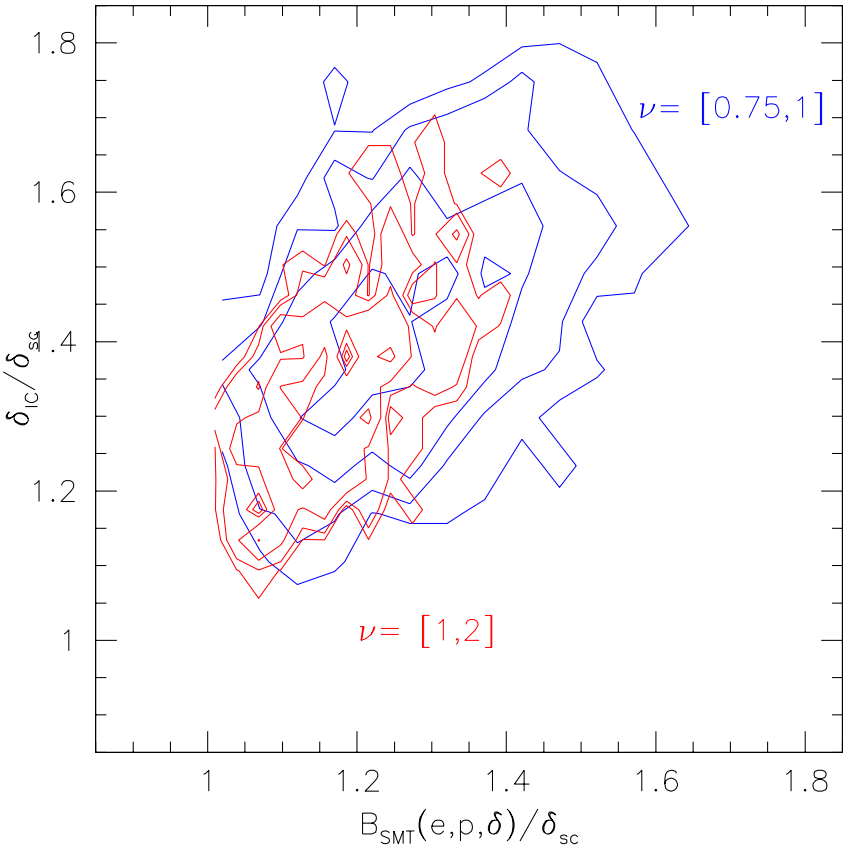

FIGURE 5. Comparison of predicted and measured dependence of critical initial density on shape parameters e and p. Note that massive halos (larger ν) tend to have initial densities which are closer to those predicted by the spherical model – which is also in agreement with the triaxial model.

that particles continue to move with their initial velocities forever.) In this case

$$1+\Delta_{\rm Zel} = \prod_{j=1}^{3}\left(1-\frac{D(t)}{D_i}\lambda_j\right)^{-1} \qquad \text{so} \qquad 1+\Delta_{\rm Zel-Sph} = \left(1-\frac{D(t)}{D_i}\frac{\delta_i}{3}\right)^{-3}. \quad (14)$$

(a sphere has all three eigenvalues equal, so each equals $\delta_i/3$). Comparison with equation (10) shows that the 'Zeldovich sphere' evolves as though $\delta_{\rm sc} = 3$. At early times ($D_t\,\delta_i/D_i \ll 1$) it matches the spherical model well, but it becomes increasingly inaccurate at later times. This suggests setting

$$1+\Delta_{\rm Ell-Coll} \approx \frac{(1+\Delta)_{\rm Sph-Coll}}{(1+\Delta_{\rm Zel-Sph})}\,(1+\Delta_{\rm Zel}), \quad (15)$$

where the two spherical models have $\delta_i = \sum_j \lambda_j$ [26]. Figure 3 shows that this approximation describes the evolution of the collapse along the first two axes reasonably well. This analytic description aids considerably in understanding many features of the collapse process, such as the overdensity required to collapse along one or two axes, and the spin and axis ratio distributions of the final collapsed objects (e.g., Figure 4).

However, for what follows, the object of most interest is $\delta_{ec}(e,p|t)$; this is the analog of $\delta_{sc}(t)$, which quantifies how the critical density for collapse at t depends on e and p. This is shown in the right hand panel of Figure 4; halos correspond to collapse along all three axes: typically, the second axis collapses at about the same time the spherical model predicts, so the third axis collapses later. As a result, the initial overdensity required for collapse today must be higher than the spherical model predicts. Figure 5 compares the predicted dependence of initial density on initial shape with that seen in simulations; there is good qualitative agreement.

This is important because similar sized patches centred on different positions in a Gaussian random field may have a range of (e,p) values. This results in stochasticity which may be the subject of similar lectures a few years from now. What is important here is that the distribution of (e,p,δ) values depends on the size of the patch: $g(e,p|\delta,R)$ [equation A3 in 53]. Since massive halos form from larger patches in the initial conditions than do less massive halos, the distribution of initial (e,p) values, and hence the distribution of final axis ratios, also depends on halo mass. Thus, the model comes with a prescription for determining halo shapes (see Figure 4). But it also means that the model predicts massive halos to have smaller values of e and p. Figure 4 suggests that, when averaged over all shapes, $\delta_{ec}(m,t)/\delta_{sc}(t)$ should be larger for small mass halos [53]. Physically, this says that to hold themselves together against the surrounding tidal field, small mass objects need to have been denser initially. Alternatively, tidal fields are more efficient at stripping away material from the outskirts of low mass halos, so the mass which remains around these objects today is the more tightly bound stuff which accreted at some earlier time, for which $D(t)\delta_{sc}(t_{early})/D(t_{early})$.

The discussion above has concentrated on what it takes to collapse along all three axes. Of course, the triaxial model provides analogous 'critical densities' for collapse along just one or two axes[53]. Convenient approximations to these are:

$$\frac{\delta_{ec}(s)}{\delta_{sc}} = 1+\beta\left(\frac{\delta_{sc}^2}{s}\right)^{\gamma} \qquad \begin{cases} (\gamma,\beta) = (0.55,-0.56) & 1-\text{axis} \\ (\gamma,\beta) = (0.28,0.012) & 2-\text{axes} \\ (\gamma,\beta) = (0.61,0.45) & 3-\text{axes} \end{cases} \tag{16}$$

[44]. In this model, tidal forces enhance collapse along the first axis and delay collapse along the last axis relative to the spherical collapse model [53]—the expressions above quantify these effects. Notice that the tidal fields may be strong enough to induce collapse of a small region (large σ), at least along one axis, even if it was *under*-dense initially!

The differences among the three critical overdensities are larger for larger values of σ, corresponding to ellipsoids of lower masses. The evolution of (initially large) high mass objects is expected to have been more nearly spherical: they have $\delta_{ec}(m,t) \approx \delta_{sc}(t)$ for all three axes. This is an important point to which we will return.

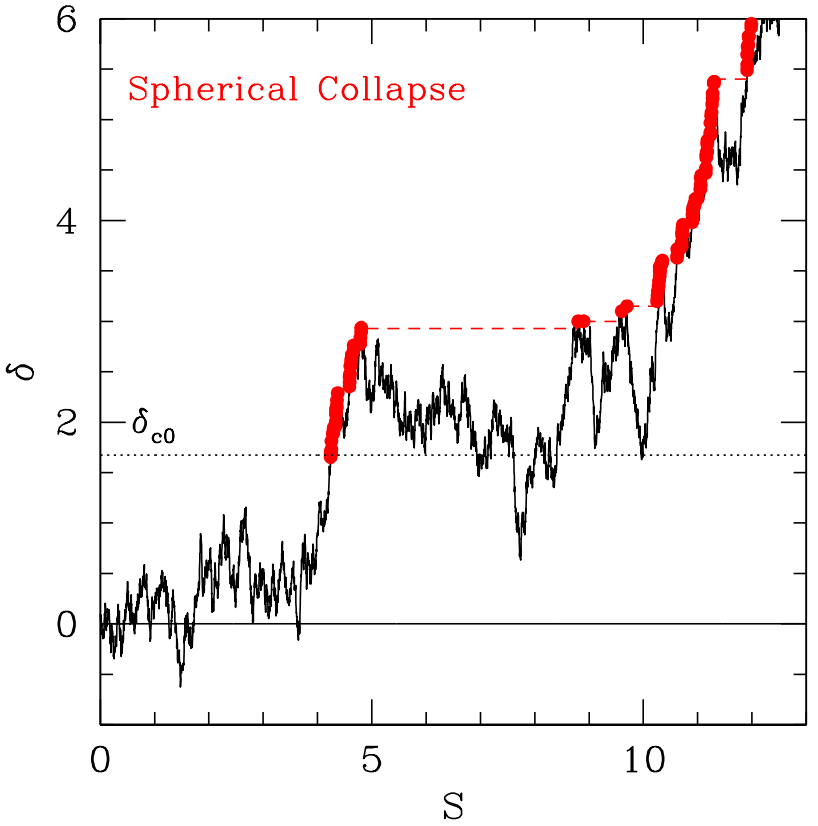

FIGURE 6. Mass history associated with a random walk (jagged line). The critical density for spherical collapse (dotted line) decreases as time increases, and mass decreases as S increases. If one imagines sliding the dotted line downwards from great height, then filled circles show the pairs (S, δ) at which the walk would first cross this line. The horizonal jumps (connected by dashed lines) show places where the mass changes dramatically – mergers [from 31].

THE EXCURSION SET APPROACH

To form, virialized objects had to fight the expansion of the Universe. This fight is more easily won if they had a head start – if they grew from large initial perturbations. Thus, given a model for gravity, the abundance of virialized objects contains information about the initial fluctuation field, and about the subsequent expansion history of the universe. The excursion set approach was developed as a method for describing how this information is encoded in the abundance and clustering of the nonlinear structures present at later times, and in their formation histories.

The key to this approach is the assumption that the nonlinear field has some memory of the initial conditions. Since the virial relation $-W = 2K$ does not care about initial conditions, it is not obvious that this is a good assumption. But arguments based on the Zeldovich approximation suggest that this should be so, at least on large scales, so it is plausible that this is also true for the abundance of objects, if not their internal structure. Comparison with simulations has shown this to be the case. In what follows I use the spherical model to illustrate the logic of the approach; the triaxial model is conceptually similar, though technically more challenging.

The ansatz

Choose a random particle in the initial density fluctuation field, and imagine smoothing the field around it with a filter of scale R. As one changes R, the overdensity within the filter will change. Imagine making a plot of the value of the smoothed density around this point as a function of R. For very large R (say, the Hubble volume), the overdensity in the smoothed filter should be negligible – the Universe is homogeneous on large scales. As R decreases, the value of the smoothed overdensity will vary, sometimes up, others down. The jagged line in Figure 6 shows that the result look like a random walk – we will discuss whether the steps in the walk are truly independent shortly. The x-axis is not quite R, but it is a monotonically decreasing function of R (see equation 1) for reasons we discuss shortly. The y-axis shows the initial overdensity multiplied by the linear theory growth factor D_0/D_i.

Although the walk starts from the origin, it will eventually reach height $\delta_{\rm sc}$ (this assumes there are fluctuations on arbitrarily small scales; while true for ΛCDM models, it may not be true in general). This first crossing of $\delta_{\rm sc}$ (it may go on to cross $\delta_{\rm sc}$ many times at still smaller R) is significant: it indicates that, when smoothed on this scale, the field was dense enough initially that it should have just collapsed and formed a virialized object today. In the spherical collapse model, shells do not cross, so the mass associated with this collapsed object is simply the mass that was originally within the smoothing filter R. Since the fluctuations are all small, this mass is $M \propto R^3$. (This also shows why the subsequent crossings of $\delta_{\rm sc}$ are not so significant – their mass is included in M. It is only the first crossing which is significant.)

Moreover, in the spherical model, the critical density required for collapse at t is independent of mass, and this critical density is a decreasing function of time. The dotted line at $\delta \approx 1.686$ in Figure 6 represents this critical value for t_0. At earlier times, this critical value was larger. The dots show the result of sliding a horizontal line downwards from great height, and recording the values of S at which the line first touches the walk. The set of (S, δ) values obtained in this way is actually a set of (M, t) values: this set can be thought of as describing the mass M of the collapsed object that this particle is in at time t. The Figure shows that, in this model, the mass increases monotonically with time, but the mass increases can sometimes be due to rather large 'instantaneous' jumps. In more picturesque language, this is a model of the mass history of objects, in which mass changes can be due to major or minor mergers, but the mass growth is hierarchical – there is no fragmentation.

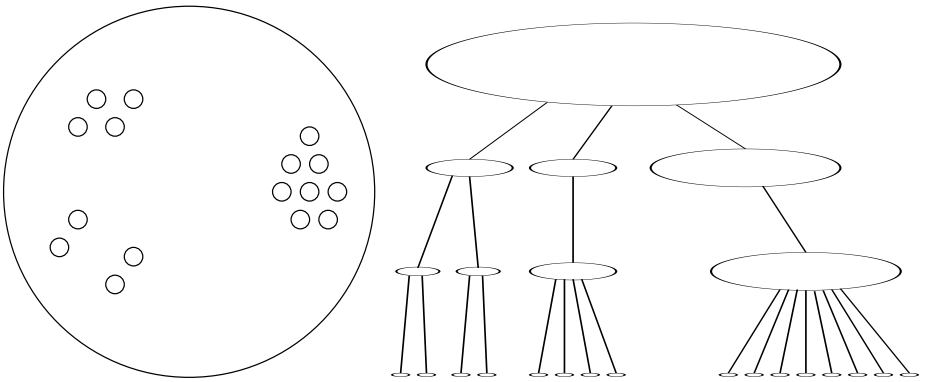

FIGURE 7. Left: Schematic drawing of the initial spatial distribution of objects that gives rise to the merger history tree shown on the right. The largest circle represents the comoving size of the initial region associated with the final collapsed bound halo. As time evolves from the initial to the final, collapse time, this comoving radius decreases. The assumption is that all the matter initially within this region remains within it always. Thus, information about how the mass of a final object was partitioned into subhaloes at a given time contains information about the halo distribution smoothed on a scale given by the radius of the larger object at that time. **Right:** Schematic drawing of the associated merger history tree. Time increases upwards: the initial time is at the bottom of the figure. The branch on the right is associated with a region that, initially, was made up of many small objects that were close to each other, but rather separated from any other objects. The branch on the left, on the other hand, is associated with a region that was initially populated rather more homogenously.

Now, clearly, the shape of the walk, and the scale R on which the walk first crosses $\delta_{\rm sc}$, and indeed, the whole set of (M,t) values, will change from one initial position or particle to another. If we imagine each object at time t as having been assembled by a sequence of mergers, then the whole set of walks associated with the various positions in the initial conditions contains information about the forest of all possible merger history trees. The excursion set ansatz is that statistical averages over this bundle of walks can provide information about various properties of this forest.

For example, in this approach, the fraction of walks that first cross $\delta_{\rm sc}$ when the smoothing scale is R or greater equals the fraction of mass that is bound up in halos of mass greater than M. Similarly, suppose one considers the subset of walks which first crossed $\delta_{\rm sc}(T)$ on scale R. For this subset, one can calculate the fraction of walks which first cross $\delta_c(t) > \delta_{\rm sc}(T)$ on scales between r and R (note r must be smaller than R). The excursion set ansatz is that this equals the fraction of the total mass in clumps hving M at time T that was in clumps of mass m or greater (of course, $m \leq M$) at the earlier time t.

There are a number of problems associated with this ansatz which we will soon discuss. Before we do so, it is worth seeing the rich variety of phenomena that this ansatz allows one to discuss.

The calculation

In practice, to compute these averages, the excursion set approach makes another assumption: that these spatial averages can be replaced with appropriate averages over an ensemble of independent walks. Although this is clearly incorrect in general, we will show what this assumption implies, and will then reconsider it later.

To estimate this fraction, it is convenient to make the following change of variables, which is motivated by the fact that the distribution of fluctuations δ in a Gaussian random field is a function of δ/σ, where σ is the rms fluctuation. We will use $S \equiv \sigma^2$ to denote the variance of the field. Our change of variables comes from noting that, at time t_i, there is a one-to-one mapping from M to S which is given by equation (1) at $t = t_i$. Thus, S, M and R are all equivalent variables. For a Gaussian field, the variables (δ, S_i) are the natural ones for the random walk. These are almost the variables shown in Figure 6.

Recall that, to estimate abundances of objects at time $t > t_i$, we are interested in the value of S_i at which the walk first exceeds some critical value. The spherical model says that, when expressed in units of the initial overdensity scaled using linear theory to time t, this value is $\delta_{\rm sc}(t)$ (although the dependence on t is weak). Therefore, had we shown the walk in units (δ_i, S_i) then the critical spherical collapse value in these units would be $\delta_{{\rm sc}i} = \delta_{\rm sc}(t)D(t_i)/D(t)$. Since $D(t)/D(t_i)$ increases with time, and $\delta_{\rm sc}(t)$ does as well (though much less strongly), this critical value decreases as t increases.

In standard gravity, the growth factor is independent of k. So if $P(k)$ is evaluated using linear theory at some time other than t_i, then, this scales S by the square of $D(t)/D(t_i)$. This means that if we show the random walk using linear theory S at t, then the height of the walk should also be scaled by one factor of $D(t)/D(t_i)$. (Note that this scales the y-axis by the square-root of the scaling applied to the x-axis – as one would expect if one thinks of the problem as a one-dimensional 'diffusion' in the y-direction, with the x-axis representing 'time'.) Figure 6 shows the walk when it has been scaled to the present time t_0. In these units, the critical overdensity for spherical collapse at the present is $\delta_{\rm sc0} = \delta_{\rm sc}(t_0)$. At some earlier time, this critical value was $\delta_{\rm sc}(t)[D(t_i)/D(t)][D(t_0)/D(t_i)] = \delta_{\rm sc}(t)[D(t_0)/D(t)]$. In these units, the critical value at earlier times was higher than it is now.

Let $f(\delta_{\rm sc}, S)\,dS$ denote the fraction of walks which first cross $\delta_{\rm sc}$ within dS of S. Next, choose some δ greater than $\delta_{\rm sc}$, and consider the probability $p(\delta, s)$ that the walk reaches δ when the smoothing scale is s. Although we know p is Gaussian, we will now rewrite this probability in a way that shows how p and f are related. This will allow us to use our knowledge of p to determine f.

Since $\delta > \delta_{\rm sc}$, all walks that reach (δ, s) *must* have crossed $\delta_{\rm sc}$ at some scale $S < s$. If we label each walk by the value of S at which it first crossed $\delta_{\rm sc}$, then it must be that

$$p(\delta, s) = \int_0^s dS\, f(\delta_{\rm sc}, S)\, p(\delta, s|\delta_{\rm sc}, S) = \int_0^s dS\, f(\delta_{\rm sc}, S)\, p(\delta - \delta_{\rm sc}, s - S), \qquad (17)$$

where $p(\delta, s|\delta_{\rm sc}, S)$ is the probability that a walk which starts from $(\delta_{\rm sc}, S)$ passes through (δ, s). The second equality is only correct if the steps are independent, so $p(\delta, s|\delta_{\rm sc}, S)$ does not depend on how the walk reached $(\delta_{\rm sc}, S)$. (This is not true in general. However, for a special choice of filter, a tophat in k-space – a choice we will

return to later – it is true for Gaussian random fields.) Therefore

$$\int_{\delta_{\rm sc}}^{\infty} d\delta\, p(\delta,s) = \int_{\delta_{\rm sc}}^{\infty} d\delta \int_0^s dS\, f(\delta_{\rm sc},S)\, p(\delta-\delta_{\rm sc}, s-S) = \int_0^s \frac{dS\, f(\delta_{\rm sc},S)}{2}. \qquad (18)$$

where the final expression – the factor of $1/2$ – uses the fact that p is symmetric about zero. (The brevity of the derivation above hides the fact that this factor of 2 has a long history, which goes by the name of the 'cloud-in-cloud' problem [38, 9].) Differentiating both sides with respect to s shows that the shape of f is related to that of p.

For a Gaussian distribution p, the left hand side can be written in terms of erfc, and hence

$$f(\delta_{\rm sc},s)\, ds = \frac{ds}{s}\frac{\delta_{\rm sc}}{\sqrt{2\pi s}} \exp\left(-\frac{\delta_{sc}^2}{2s}\right). \qquad (19)$$

To turn this estimate of the mass fraction into an estimate of halo abundances, simply set

$$\frac{dn(\delta_{\rm sc},m)}{d\ln m}\, d\ln m \equiv \frac{\rho}{m} f(\delta_{\rm sc},s)\, ds. \qquad (20)$$

This suggests defining a characteristic mass from

$$\delta_{sc}^2(z) \equiv s(m_*,z), \qquad \text{so} \qquad \frac{m_*(z)}{m_*(0)} = (1+z)^{-(3+n)/3} \qquad (21)$$

where the final expression assumes an Einstein de-Sitter universe in which $P_i(k) \propto k^n$.

Similarly, because a walk that starts at some (S,δ) other than $(0,0)$ is otherwise the same as one which starts from the origin, the associated first crossing distribution is $f(\delta_{\rm sc}-\delta|s-S)\, ds$. Thus, the conditional distribution of (m,t) objects which make up (M,T) halos is

$$\frac{m}{M}\frac{dN(m,t|M,T)}{d\ln m}\, d\ln m \equiv \frac{ds}{s-S}\frac{\delta_{\rm sc}(t)-\delta_{\rm sc}(T)}{\sqrt{2\pi(s-S)}} \exp\left(-\frac{[\delta_{sc}(t)-\delta_{\rm sc}(T)]^2}{2(s-S)}\right) \qquad (22)$$

[9, 25]. This expression can be used to quantify the tendency for massive objects to assemble later in hierarchical models. Bayes rule says $[dN(M,T|m,t)/dM]$ equals $[dN(m,t|M,T)/dm]\,[dn(M,T)/dM]/[dn(m,t)/dm]$, so taking the limit $t \to T$ provides expressions for merger rates [25].

Since an object of mass M at T can have at most one piece of mass $m \geq M/2$ at $t \leq T$, the distribution of times when half the mass has been assembled in one piece, is simply

$$\begin{aligned} \frac{\partial}{\partial t}\int_0^t dt\, p(t|M,T) &= \frac{\partial}{\partial t}\int_{m=M/2}^{M} dm\, \frac{dN(m,t|M,T)}{dm} \qquad (23)\\ &\approx 2\omega_{0.5}\,{\rm erfc}\left(\frac{\omega_{0.5}}{\sqrt{2}}\right)\frac{\partial\omega_{0.5}}{\partial\delta_{\rm sc}}\frac{\partial\delta_{\rm sc}}{\partial t} \quad \text{where } \omega_{0.5} \equiv \frac{\delta_{\rm sc}(t_{\rm f})-\delta_{\rm sc}(T)}{\sqrt{S_{0.5}-S}}, \end{aligned}$$

where $S_{0.5} \equiv S(M/2)$, and the derivative is to be evaluated at $t = t_{\rm f}$ [25]. Strictly speaking, this final expression is only correct for a white noise power spectrum. However,

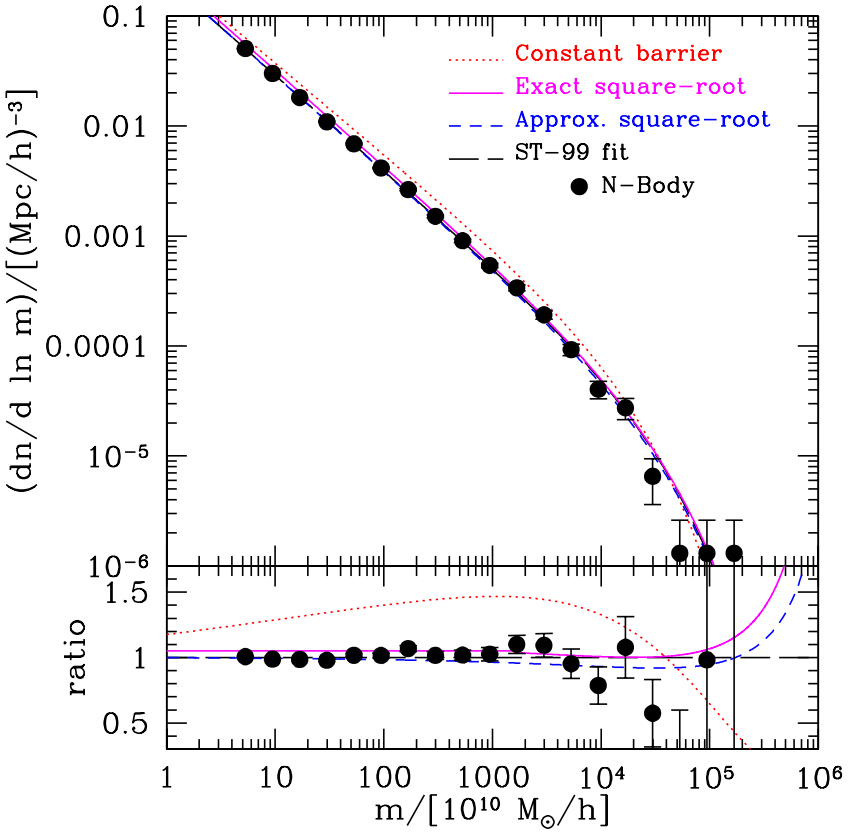

FIGURE 8. Comparison [from 30] of the mass function measured in the GIF2 simulation (symbols) with that derived from using equation (20) with the first crossing distribution of a constant barrier (equation 19, dotted); a square-root barrier (equation 16 with (β, γ) = (0.5,0.5) and lowered in height by a factor $q = 0.55$) (solid); and an analytic approximation to this first crossing problem (dashed) from [55]. Bottom panel shows the ratio of both data and theory curves to the functional form of [54].

when expressed in terms of $\omega_{0.5}$, this same formula provides a reasonable description of simulations (see Figure 9).

The mass at this time can have any value between $m/M = 1/2$ and 1. It is slightly more work to derive an expression for the distribution of this mass, so we simply state the result:

$$p(\mu)\,\mathrm{d}\mu = \frac{2}{\pi}\sqrt{\frac{1-\mu}{2\mu-1}}\,\frac{\mathrm{d}\mu}{\mu^2}, \quad \text{where} \quad \frac{1}{2} \le \mu \le 1 \quad \text{and} \quad \mu \equiv \frac{m}{M}; \tag{24}$$

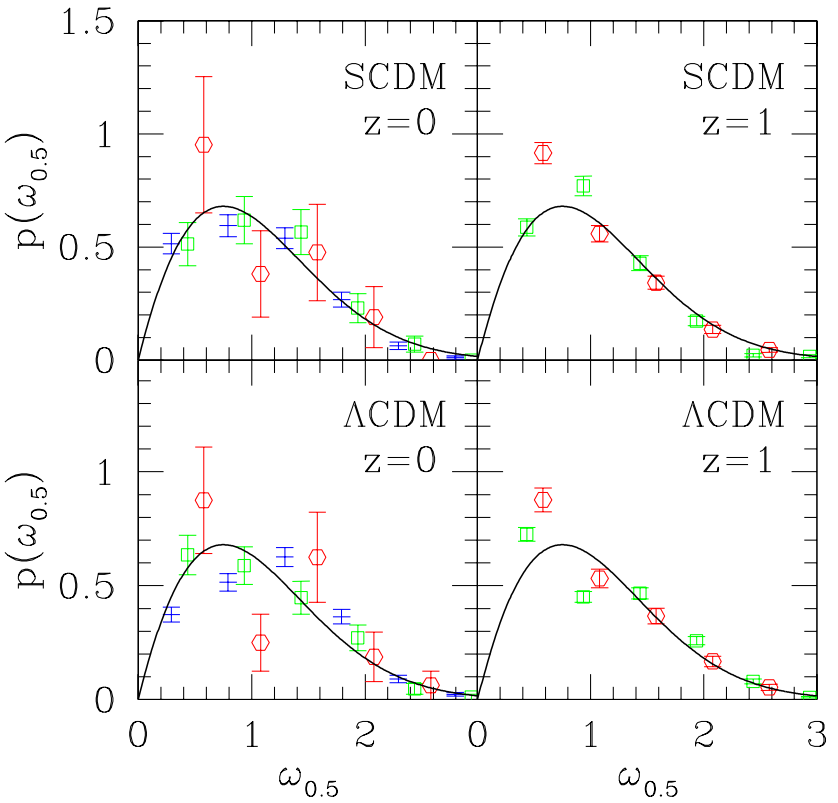

FIGURE 9. Distribution of scaled formation times in two different cosmological models, for haloes identified at two different redshifts. In these scaled units, the formation time distribution is expected to be independent of halo mass and final time. Solid curve shows the precise form which this universal formation time distribution is expected to have (equation 24). In all panels, squares and hexagons show the simulation results for parent haloes with masses in the range $4 \leq M_1/M_*(z_1) < 8$ and $16 \leq M_1/M_*(z_1) < 32$. Simple bars in the panels on the left show results for slightly lower halo masses: $M_1/M_*(z_1 = 0) \leq 2$. Error bars were estimated assuming Poisson counts. Evidently, equation (24) provides a reasonable, but not perfect description of halo formation times in the simulations (from [56]).

just prior to this time, the distribution is

$$q(\mu)\,\mathrm{d}\mu = \frac{1}{\pi(1-\mu)}\left(\sqrt{\frac{\mu}{1-2\mu}} - \sqrt{1-2\mu}\right)\frac{\mathrm{d}\mu}{\mu^2}, \quad \text{where } \frac{1}{4} \leq \mu \leq \frac{1}{2} \tag{25}$$

[34]. Figure 10 compares these distributions with measurements in simulations.

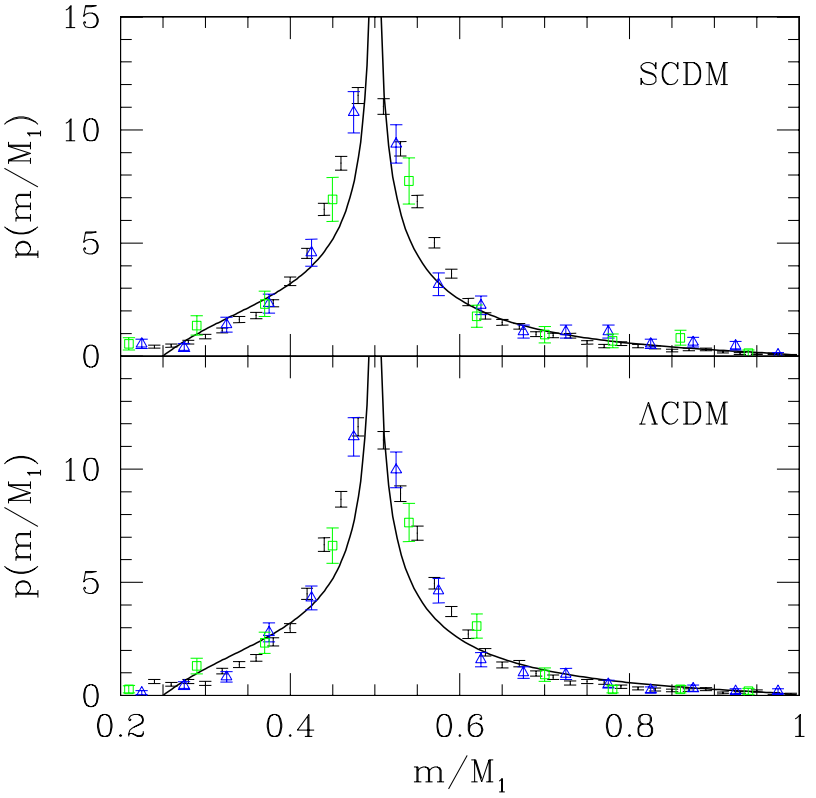

FIGURE 10. The distribution of masses m at formation, for parent haloes which have mass M_1 at $z_1 = 0$. Symbols show the simulation results for $M_1/M_*(z_1) \leq 1$ (dots), $2 \leq M_1/M_*(z_1) < 4$ (triangles), and $M_1/M_*(z_1) \geq 8$ (squares). Error bars were estimated assuming Poisson counts. Curves on the right and the left of $m/M_1 = 1/2$ show the distributions in equations (24) and (25) respectively. There is no obvious trend with M_1, although haloes in simulations appear to have $m/M_1 \approx 1/2$ slightly more frequently than the model predicts. Results for formation masses of parent haloes identified at other redshifts are similar (from [56]).

For haloes of fi xed mass M, the conditional distribution of formation masses m when it is known that the formation time was z_f is given by

$$p(\mu|z_f)\,d\mu \equiv \frac{p(\mu, z_f)\,d\mu}{p(z_f)} = \frac{p(\mu)\,d\mu}{s/S-1}\,\frac{\exp\left[-\frac{\omega_{0.5}^2}{2}\,\frac{(S_{0.5}-S)}{(s-S)}\right]}{2\,\mathrm{erfc}(\omega_{0.5}/\sqrt{2})}, \tag{26}$$

where $s \equiv \sigma^2(m)$, $S_1 \equiv \sigma^2(M_1)$, $S_f \equiv \sigma^2(M_1/2)$, and ω was defi ned in equation (24). The factor which multiplies $p(\mu)$ is largest at $s/S_1 - 1 = \omega^2$, so objects which form at

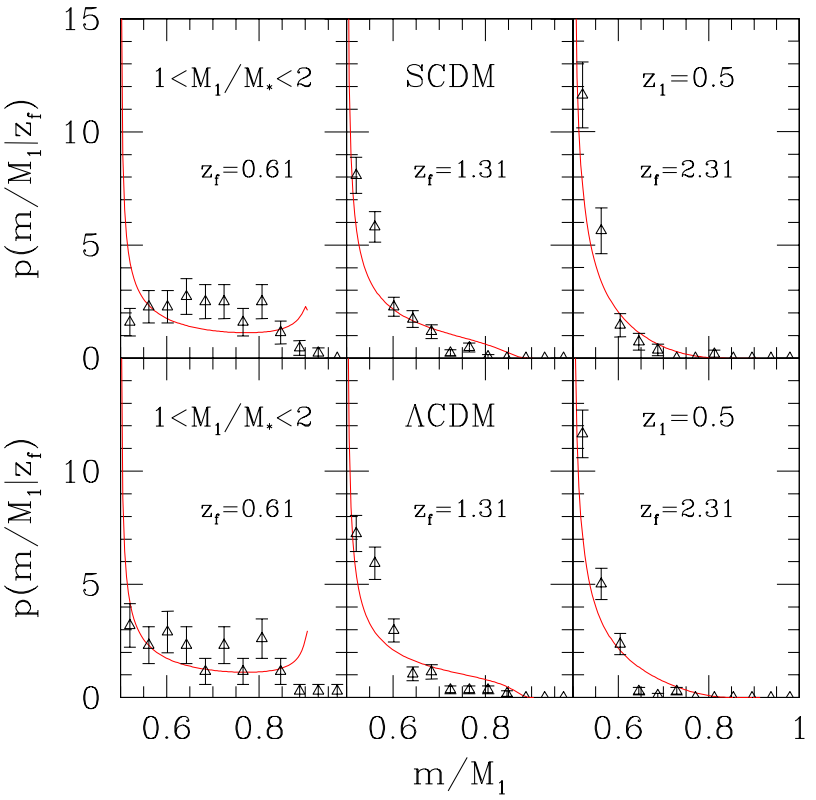

FIGURE 11. Conditional distribution of masses m at formation, given that the mass of the parent halo was in the range $1 < M_1/M_*(z_1) < 2$ at $z_1 = 0.5$, for a range of choices of the redshift of formation (labeled in the middle of each panel). Symbols show the measurements in the simulations, and curves show equation (26) (from [56]).

redshifts which are lower than the mean value for that mass (i.e., $\omega < 1$), are expected to have formation masses which are biased towards $\mu \approx 1$ (i.e., $s \approx S_1$). Conversely, objects which form at abnormally high redshifts ($\omega > 1$) are expected to have formation masses which are closer to the minimum value allowed: $\mu \approx 1/2$. Presumably, this is a consequence of the fact that, to have $\mu \approx 1$ requires two pieces each of size $\mu \approx 1/2$. In a hierarchical model, the building blocks available to form the parent halo are, on average, smaller at early times: when the probability of having an object of mass $\mu \approx 1/2$ is small, the chance of having two such objects is smaller still. In effect, our formula (26) quantifies the importance of this effect. Figure 11 shows that it provides a reasonable description of this trend in simulations.

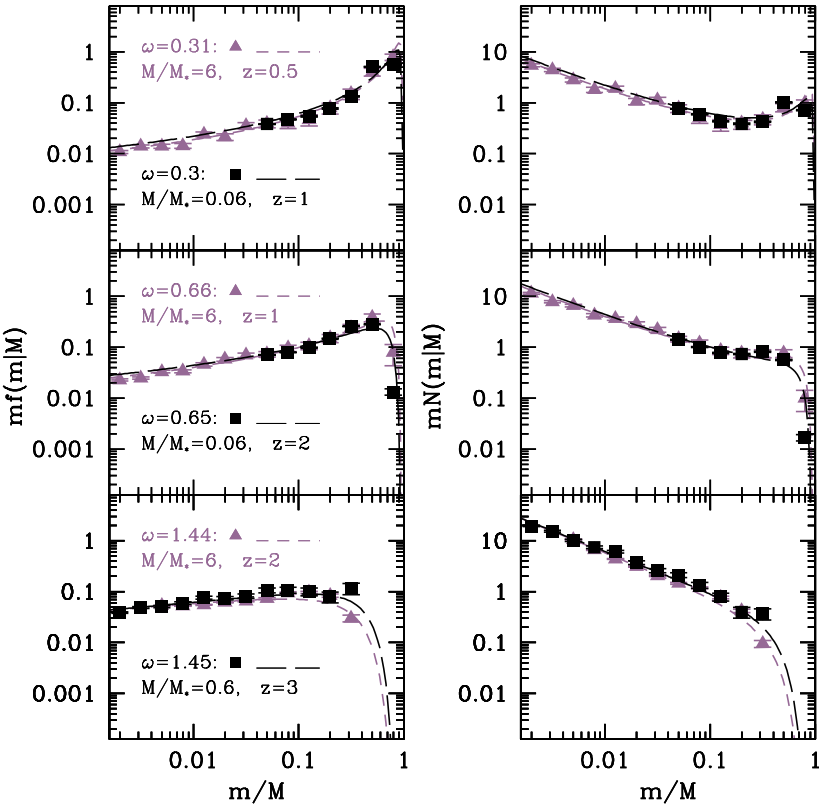

FIGURE 12. Conditional mass functions showing the mass fraction of M halos at the present that was in m halos at z [from 30]. Different combinations of M and z but similar $[\omega \equiv \delta_{sc}(z) - \delta_{sc}(z=0)]^2/S_0$ and $(s/S_0 - 1)$ yield similar conditional mass functions. Symbols show measurements in simulations and curves show the predictions of [55].

Some implications

Haloes which form at abnormally early times are more likely to have formation masses of order one-half that of the final mass of the parent, whereas haloes which form at abnormally late times are more likely to have formation masses which are closer to that of the parent (Figure 26). One consequence of this is that haloes which form late are more likely to have experienced a recent major merger. This is a generic consequence of hierarchical formation.

Suppose star formation only occurs in halos that are above a minimum mass but below a maximum mass. For argument's sake, suppose that these masses are 0.006 and

0.06 of M_* today, and that these limits do not evolve. Figure 12 shows that the mass fraction within this range increases at late times for low mass halos today (filled squares, and m/M between 0.1 and 1), but decreases for higher mass halos (filled triangles, $0.001 \leq m/M \leq 0.01$). As a result, massive halos in this model will host older stars, and the typical mass object in which star formation occurs will decrease with time. These are two aspects of the phenomenon known as 'down-sizing', and the discussion above shows how this can be accomplished in hierarchical models [47, 33].

Finally, it is a curious fact that the abundance of dark matter halos of mass $10^{12}h^{-1}M_\odot$ is almost constant from $z = 2$ to the present [e.g. Fig. 2 in 39]. This happens to be the mass of our Galaxy, it is approximately the transition scale from early to late type galaxies, and it is also the value adopted in most current models of AGN activity at $z \sim 2$.

Correlations with environment: Bias and the peak-background split

The approach above allows a straightforward estimate of how halo abundances correlate with their large scale environment [29]. In the spherical model, the environment on some scale V is described by its density. In triaxial collapse models, two other numbers also matter: these may be related to filaments and sheets. We show shortly that this provides a framework for discussing how halos populate the 'cosmic web'. Here, we explore the simpler definition of environment as 'density', without regard to 'morphology'.

The mean number of halos of mass m in a cell depends on the mass M in the cell and its volume V. In the spherical model, this volume was initially different, although the mass was not. The factor $1+\Delta \equiv M/\bar{\rho}V$ describes how much the volume has changed. In turn, $1+\Delta$ depends on the initial overdensity of the patch (equation 10). This means that we can estimate

$$\frac{dN(m,\delta_{sc}|M,V)}{dm} = \frac{dN[m,\delta_{sc}|M,\delta_{\rm L}(M/V)]}{dm}, \text{ with } \delta_{sc}-\delta_{\rm L}(M/V) \approx \frac{\delta_{sc}}{(1+\Delta)^{1/\delta_{sc}}}, \quad (27)$$

where the right hand side of the first equality is given by equation (22) for the conditional mass function, and, instead of writing the time variables which appear on the left hand side of that expression, we have written the linear theory quantities which appear on its right hand side. The result depends on $\delta_{sc}-\delta_{\rm L}$; the second expression above uses equation (10) to show that this means that the environment acts like an effective growth factor. (See [27] for an explicit demonstration that this is consistent with the picture in which the environment acts like an effective cosmology.) In terms of the excursion set description, $1+\Delta \ll 1$ in underdense regions, so the 'barrier' is higher than δ_{sc}; as a result, the typical halo masses are expected to be smaller. Conversely, the mass function is expected to be top-heavy in dense regions. In particular,

$$\frac{dN(m,\delta_{sc}|M,V)}{dm} \neq (1+\Delta)\,\frac{dn(m,\delta_{sc})}{dm}\,V; \quad (28)$$

the shape of the mass function depends on M and V.

Figure 13 shows this explicitly: the simulation volume was divided up into into cubes, each $10h^{-1}$Mpc on a side, and three subsets of cubes were chosen: the densest, and

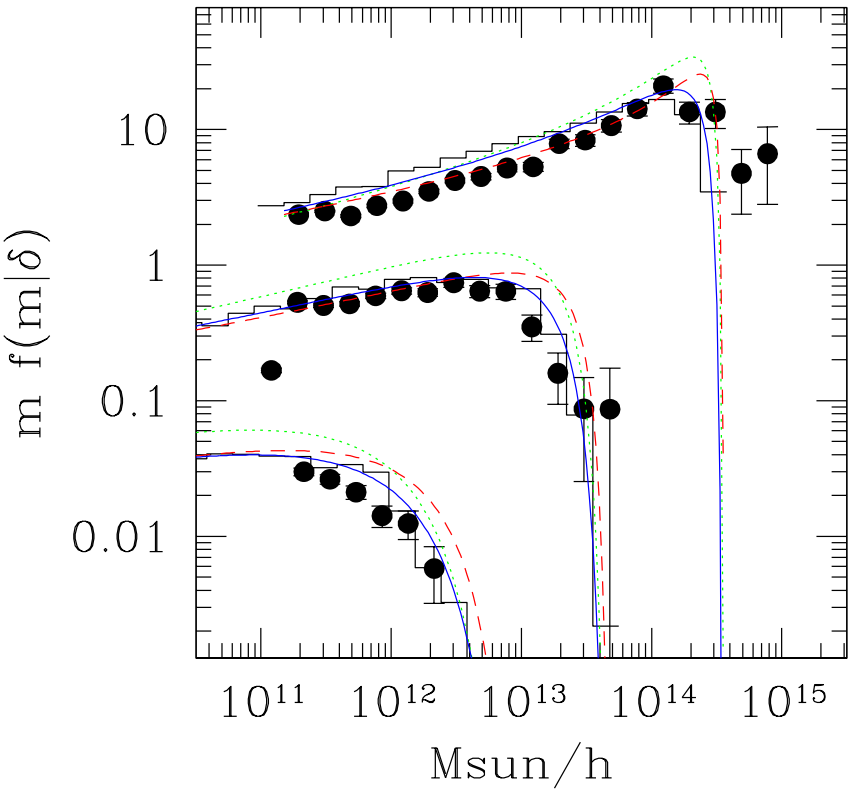

FIGURE 13. Mass functions as a function of local density in ΛCDM simulations (symbols with error bars). Dotted curves show the spherical collapse prediction, and solid curves show the prediction associated with ellipsoidal collapse [from 55, which also describes the dashed curves]. The curves have been offset upwards by a factor of ten and a hundred, in the case of the middle and topmost curves, respectively. The upper most curves show the densest cells.

least dense ten percent of the cells, and the ten percent around the median density. The symbols show the halo abundances in these subsets. They clearly have different shapes; while the spherical model describes the qualitative differences (dotted curves), the ellipsoidal collapse model is more accurate (solid curves).

The approach above simplifies when V is large, since then $\Delta \ll 1$, and so $m \ll M$ for all cells. In this case the other quantity in this expression $s(m) - S(M) \approx s(m)$, so $dN(m, \delta_{sc}|M,V)/dm \approx dn(m, \delta_{sc} - \delta_{\rm L})/dm$. Thus $dN(m, \delta_{sc}|M,V)/dm$ can be got from Taylor expanding $dn(m, \delta_{sc})/dm$ around $\Delta = 0$. This is an extremely powerful result; when written in terms of dn/dm, it is is known as the peak-background split [3]. It says

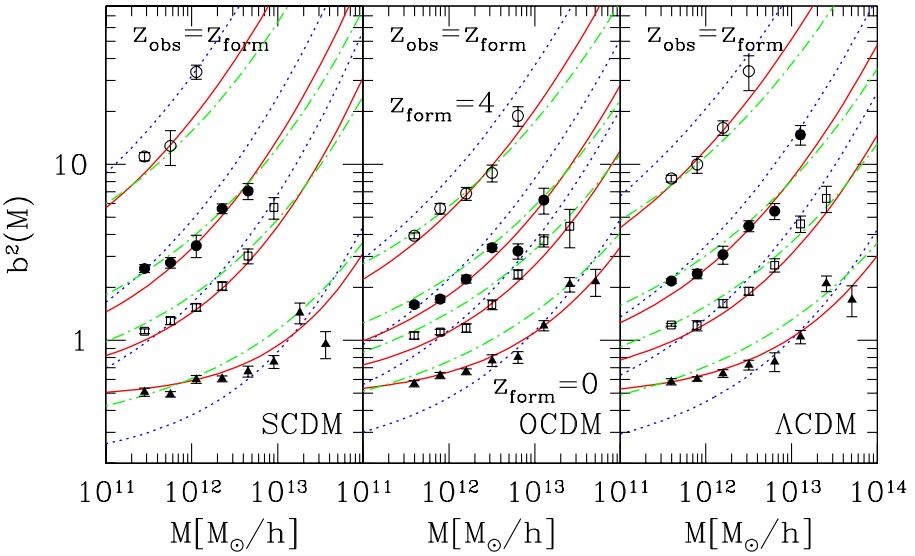

FIGURE 14. The large scale bias relation at $z_{obs} = z_{form}$ between haloes which are identified at z_{form}, and the mass at that time. Dotted curves show the relations which follow from the constant barrier model (equation 30), and solid curves show that associated with a moving barrier [from 54].

that

$$\frac{\langle dN(m,\delta_c|M,V)/dm\rangle}{V\,dn(m,\delta_c)/dm} \equiv 1+\langle\delta_h(m|M,V)\rangle = 1+\sum_{k>0}\frac{b_k(m,\delta_{sc})}{k!}\left(\Delta^k-\langle\Delta^k\rangle\right), \quad (29)$$

where the b_k are the coeffi cients of the Taylor series expansion, and the $\langle\Delta^k\rangle$ terms are required if one wishes to truncate the expansion at fi nite k but still enforce $\langle\delta_h(m)|\Delta\rangle = 0$. This expansion connects the excursion approach with what is known as the local deterministic bias model [16]. But note that because the approach provides an analytic formula, it can be used on scales where the Taylor expansion is no longer useful [52].

This expansion says that, if the halo mass function is given by equation (19), then

$$\frac{\langle\delta_m\delta_h\rangle}{\langle\delta_m^2\rangle} \approx b_1(m,\delta_{sc}) = 1+\frac{\delta_{sc}^2/s(m)-1}{\delta_{sc}} \qquad \text{and} \qquad \frac{\langle\delta_h^2\rangle}{\langle\delta_m^2\rangle} \approx b_1^2(m,\delta_{sc}); \quad (30)$$

the clustering of halos should be different from that of the mass. Figure 14 compares measurements of the ratio of the halo and mass power spectra in simulations, P_{hh}/P_{mm} at $k \ll 1$, with the predicted linear bias factor, b_1^2. The near future will test if ξ_{hh}/ξ_{mm} has the same value to percent precision, and if $P_{hm}/P_{mm} = \xi_{hm}/\xi_{mm} = \ldots = b_1$. Notice that b_1 increases strongly at large m; this is the fundamental reason for most observed correlations between galaxies and their environments (e.g., Figure 1). Formulae for the other b_k are in [41].

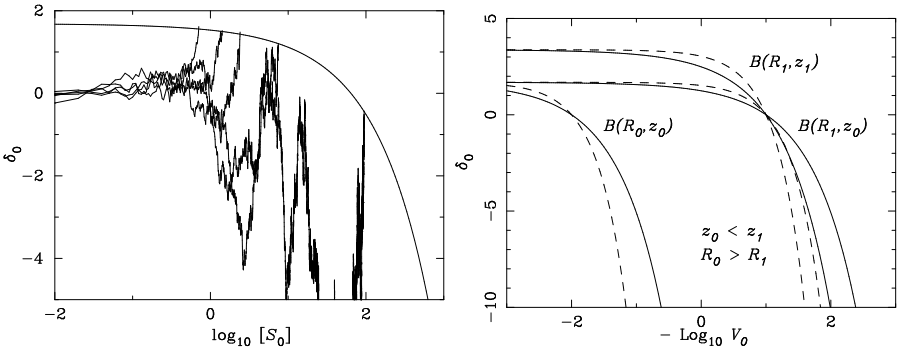

FIGURE 15. Left: Examples of trajectories (thin jagged curves) traced out by the Lagrangian overdensity, δ_0, as a function of linear variance, S_0. The trajectories are absorbed at the barrier (thick solid line). Here, the barrier shape is given by the spherical collapse model (equation 10), and $S_0 \propto 1/V_0$ as it is for white noise. **Right:** Dependence of the barrier shape on comoving Eulerian size R and redshift z. Solid curves show $B = \delta_0(R_0|R,z)$ of equation (10), and dashed curves show what would happen if $\delta_{\rm sc} = 1$ in this expression. For white noise, $S_0 \propto 1/V_0 \propto 1/R_0^3$ [from 46].

Counts in cells

The discussion above suggests that Figure 6, which shows the initial overdensity δ as a function of (Lagrangian) smoothing scale $S(M)$, can be combined with the spherical evolution model to infer how the evolved density $1+\Delta = M/\rho V$ around the same point depends on (Eulerian) smoothing scale V at the later time t. This is because equation (10) provides a relation between M, V, δ and t. To see this, suppose we fix t. Then the spherical model describes a family of curves in the space (δ, M), which are parametrized by V. But there is a one-to-one mapping from M to S, so (δ, M) can be mapped to the coordinates (δ, S) which are shown in Figure 6. Figure 15 shows this explicitly, using slightly different notation: $B(R,z)$ denotes $\delta_{\rm L}$ of equation (10), when the cell size is $V = 4\pi R^3/3$ and the time variable is the redshift z.

Note that, in the limit $V \to 0$, $B \to \delta_{sc}$, so it is the same as the 'barrier' associated with the halo mass function. Thus, for fixed t, by recording the values of S (hence M) at which the walk pierces the curves labeled by V, one obtains an estimate of the density run (on scales larger than the virial radius) surrounding each halo. If one fixes R instead, and allows t to vary, then this allows one to quantify the way in which matter flows in and out of (comoving) Eulerian cells. Finally, for a given V and t, the fraction of walks which first cross the spherical evolution curve on scale S provides an estimate of the probability that a cell of size V in the evolved distribution contains mass M: the Eulerian counts in cells distribution [46]. A good approximation to the relation implied by this model is

$$(1+\Delta)^2 p(\Delta|V) \;\approx\; \exp\left[-\frac{B(S|V)^2}{2S}\right] \sqrt{\frac{B(S|V)^2}{2\pi S}} \frac{d\ln S}{d\ln(1+\Delta)} \left|1 - \frac{\partial \ln B(S|V)}{\partial \ln S}\right|$$

$$= \exp\left[-\frac{\delta_L^2}{2S}\right]\sqrt{\frac{\delta_L^2}{2\pi S}}\left|\frac{d\ln(\delta_L/S)}{d\ln(1+\Delta)}\right| \quad (31)$$

where δ_L is given by equation (10) (see [26] for details).

More standard perturbation theory methods for the nonlinear counts in cells distribution [6] provide what is, in effect, a monotonic, deterministic mapping between the initial and final overdensities. Because the final overdensity at a specified position in space is determined solely by the initial value at that position, this is sometimes also called a 'local' mapping, since values of the initial fluctuation field at other positions are assumed to not affect the mapping. The excursion set approach outlined here accounts for the fact that the evolution of a given region may actually be determined by less local surroundings.

For example, consider the evolution of an underdense region which is surrounded by a dense shell. If the shell is sufficiently dense, then it will eventually collapse, crushing the smaller region within it. The local approximation would have predicted expansion rather than collapse for the smaller underdense region. Figure 17 below illustrates this 'void-in-cloud' problem. Clearly, in such cases, the mapping between initial and final overdensities is not as 'local' as perturbation theory assumes, and accounting for this 'cloud-in-cloud' problem is likely to be more important for small 'clouds'. If not accounted for, this effect will manifest both as stochasticity (since the same initial overdensity may map to many different final densities depending on the surroundings) and, perhaps, as a bias. In this respect, the excursion set approach provides an algorithm which accounts for this source of non-locality; of course, once the correct large scale has been chosen, the mapping (equation 10) is assumed to be deterministic.

In triaxial collapse models, the nonlinear density is a deterministic function of three quantities associated with the initial fluctuation field. In the context of perturbation theory models for the pdf, the mapping from initial density to final density will appear to be stochastic if the influence of the two other variables is not accounted for. In the excursion set approach, this stochasticity is in addition to that which derives from the cloud-in-cloud problem, which is now associated with all three variables. See [26] for a discussion of how ellipsoidal collapse models can be incorporated into this approach.

Voids

So far, we have developed a description of structure formation that was based on where the mass is. However, because the virialized structures are of order 100 times denser than the background, they occupy one percent of the volume. Therefore, one might wonder if an equivalent description of structure formation could be built by studying where the mass is not. The resulting picture is one in which the matter in the Universe accumulates in halos which populate sheets and filaments whose spatial arrangement is dictated by the growing underdense expanses which approximately fill space.

Figure 16 shows the time evolution of an initially underdense region. As time evolves, the region clearly builds up a dense and compact bounding "wall". The right hand panel

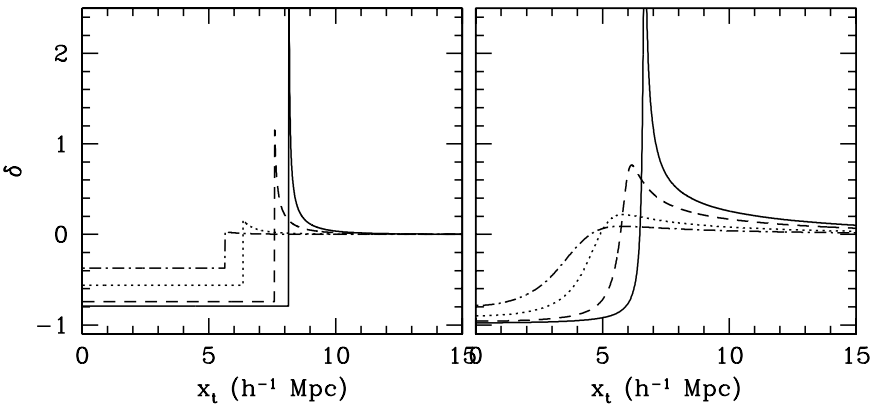

FIGURE 16. Spherical model for the evolution of voids in an Einstein de-Sitter universe. **Left:** a pure (uncompensated) tophat void evolving up to the epoch of shell-crossing. Initial (linearly extrapolated) density deficit was $\delta_{\rm L} = -10.0$, initial (comoving) radius $R_i = 5.0h^{-1}$Mpc. Timesteps shown are at $a = 0.05, 0.1, 0.2$ and 0.3. **Right:** Evolution of a void with the same $\delta_{\rm L}$ and R_i, but with initial profile given by [eq.7.10 in 3]. At late times, both profiles look similar, both having formed an obvious ridge (this happens only if the initial profile is sufficiently steep) [from 57].

shows that the development of a ridge at the boundary is fairly generic; it is not restricted to tophats. For a perfectly spherical void with a perfect tophat profile this ridge forms when the linearly extrapolated underdensity reaches a critical value: $\delta_{\rm v} = -2.81$ in an $\Omega_0 = 1$ Universe. At this time, the comoving size of the patch is 1.7 times larger than initially, so the density within the void is 0.2 times that of the background universe.

With these values in hand, one might have thought that one could estimate the abundance of voids simply by inserting this value into the excursion set approach. However, Figure 17 shows that there is an important difference between voids and clusters: voids which happen to be surrounded by an overdensity will be squeezed to vanishingly small size as the region surrounding them shrinks. Thus, in addition to accounting for the 'void-in-void' problem (the analogue of the 'cloud-in-cloud' problem for halos) one must also account for the 'void-in-cloud' problem. This can be done by solving for the fraction $f(S, \delta_{\rm v}, \delta_{\rm c})$ of walks which first cross $\delta_{\rm v}$ at S and have not crossed $\delta_{\rm c}$ at any $s \le S$. The exact solution is complicated [57], but for $\delta_{\rm c}/|\delta_{\rm v}| \ge 1/4$ or so, it is well approximated by

$$\nu_v f(\nu_v) \approx \sqrt{\frac{\nu_v}{2\pi}} \exp\left(-\frac{\nu_v}{2}\right) \exp\left(-\frac{|\delta_{\rm v}|}{\delta_{\rm c}}\frac{\mathscr{D}^2}{4\nu_v} - 2\frac{\mathscr{D}^4}{\nu_v^2}\right), \text{ where } \mathscr{D} \equiv \frac{|\delta_{\rm v}|}{(\delta_{\rm c} + |\delta_{\rm v}|)} \quad (32)$$

and $\nu_v \equiv \delta_{\rm v}^2/S$ [57]. This shows that $f(\nu_v)$ cuts-off sharply at both small and large values of ν: the distribution of void masses is reasonably well peaked about $\nu_v \approx 1$, corresponding to a characteristic mass of order $S(m) \approx \delta_{\rm v}^2$.

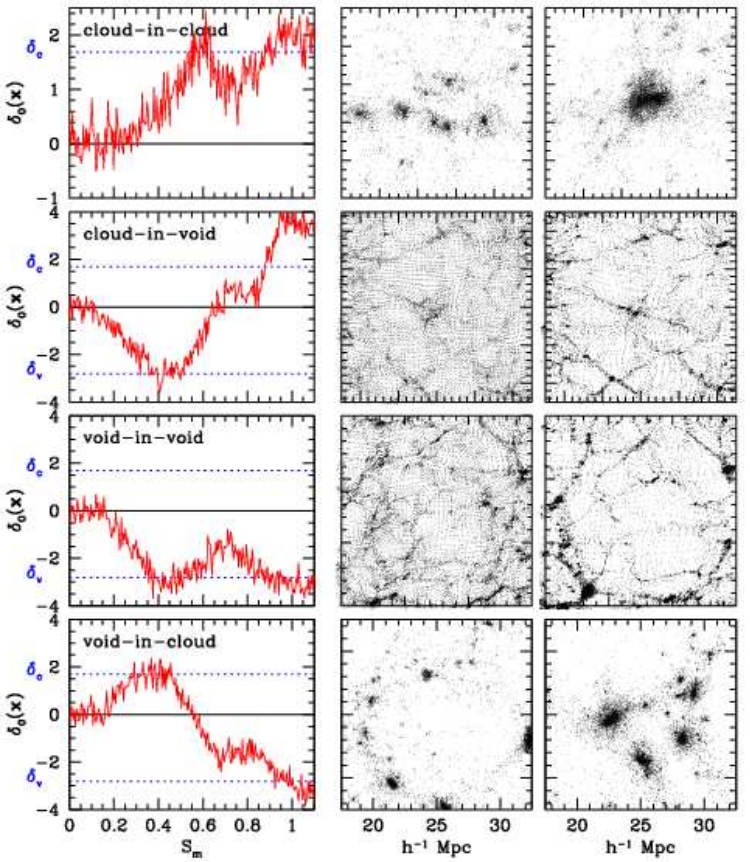

FIGURE 17. Four basic modes of hierarchical clustering: the cloud-in-cloud, cloud-in-void, void-in-void and void-in-cloud processes (from top to bottom). Each mode is illustrated using three frames. Leftmost panels show the 'random walk', δ_L vs S, associated with the particle at the center, and dotted horizontal lines show δ_{sc} and δ_v. The two frames on the right show the associated particle distribution at early (middle) and later (right) times. Whereas halos within voids may be observable (second row), voids within collapsed halos are not (bottom row shows a small void which will be squeezed to small size as the surrounding halo collapses). This is what makes the calculation of void sizes qualitatively different from that for halos [from 57].

When $\delta_c \gg |\delta_v|$, then $\mathscr{D} \to 0$, and the second exponential tends to unity. In this limit, the two-barrier distribution reduces to that associated with a single barrier at δ_v. This shows explicitly that when the void-in-cloud process is unimportant ($\mathscr{D} \to 0$), then the abundance of voids is given by accounting correctly for the void-in-void process. The quantity $\int d\nu_v f(\nu_v) \approx 1 - \mathscr{D}$ is the mass fraction in voids (this expression is exact for the exact solution), so the volume fraction in voids is $1.7^3\,(1 - \mathscr{D})$. For $\delta_v = -2.81$ and $\delta_c = 1.686$, this ratio is larger than unity, indicating that the voids fi ll the universe. Thus,

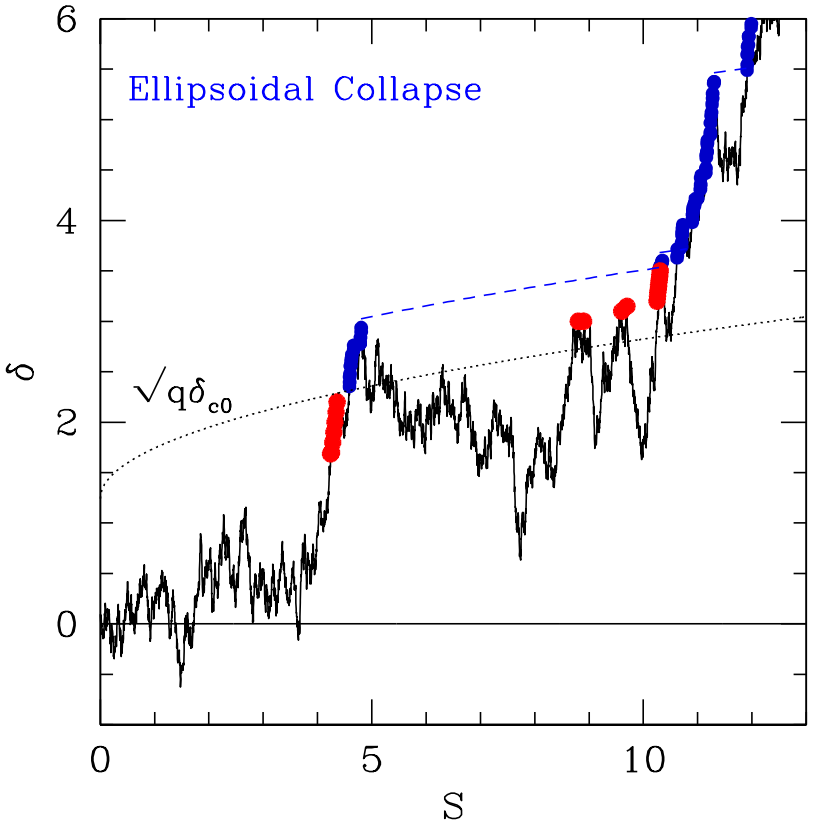

FIGURE 18. Same walk as in Figure 6, but now the barriers which must be crossed (by the same random walk) have heights which increase as the square root of S (this should be a good approximation to ellipsoidal collapse models). Some of the filled circles, which were part of the mass history in the spherical model, are not part of the history associated with ellipsoidal collapse, illustrating that the forest of merger histories depends on the details of the critical collapse threshold – i.e., on the model one uses to describe nonlinear collapse.

we have a model in which about one third of the mass of the universe is associated with voids which occupy most of the volume. The remaining seventy percent of the mass is in between the voids, and occupies negligible volume (these are most of the halos!).

Moving barrier models

The discussion above has focussed on the spherical evolution model, for which δ_{sc} is independent of m. This simplifies the excursion set approach. When the height of the barrier depends on s (e.g. equation 16), then the expression for the first crossing distribution is more complicated. However, the logic of the entire approach is not. This is illustrated in Figure 18. Unfortunately, analytic solutions to the first crossing problem are only known for special cases: $\delta_{\rm sc}(s) = \delta_{\rm c0} + \beta\, s^{\gamma}$, with $\gamma = 0, 1/2, 1$ or 2. For more complicated cases, the solution must be obtained numerically using standard methods, or by Monte-Carloing the walks. (But see [55] for a simple analytic approximation that works reasonably well for a wide range of barrier shapes.)

In general, these barriers permit fewer symmetries than the constant barrier – essentially because the equation for a straight line is only trivially modified when one shifts the origin, but the change in origin is more significant for a curve. As a result, whereas the solution to problems having two constant barriers can be written in terms of the scaling variable $\nu_{10} = (\delta_{c1} - \delta_{c0})^2/(s - S)$, for a barrier which increases as the square-root of S, such expressions have $(\delta_{c1} - \delta_{c0})^2/S$ and $(s/S - 1)$ appearing separately [30].

We noted earlier that, in the large mass (small S) limit, $\delta_{ec} \approx \delta_{sc}$. This means that, generically, the triaxial model should predict approximately the same number of massive objects as the spherical model. However, comparisons between the spherical and ellipsoidal collapse models are usually presented (e.g. Figure 8!) by lowering all factors of $\delta_{\rm sc}$ in equation (16) by a factor of about $\sqrt{0.75}$. Only when this is done does this model provide a good description of simulations [53]. As we describe below, the origin of this factor may have little to do with the physics of the collapse: a fair comparison of the models would include (essentially) the same factor to both δ_{sc} and δ_{ec}.

Of course, in the triaxial collapse model, one should really solve a three dimensional walk – in (δ, e, p) rather than just δ. This is the subject of [55]; a moving barrier, one whose height increases with s, is a convenient approximation for reducing (a little!) the complexity of the problem (see equation 16 and associated discussion). If one only studies walks in δ, and the true collapse is triaxial, then it may be more appropriate to treat the barrier which must be crossed as being stochastic. This problem has not been studied using the excursion set approach, although Figure 7 in [53] shows the 'fuzziness' of the barrier that one might expect [also see discussion in 55].

The cosmic web

The key output from the triaxial collapse models is an estimate of the typical overdensity required for collapse along one, two and three axes by redshift z. The dotted curves in Figure 19 show how these three 'barriers' depend on mass. From bottom to top, the curves show δ_{ec1}, δ_{ec2} and δ_{ec3} of equation (16).

Notice that these barrier shapes depend both on $\sigma(m)$ and on $\delta_{sc}(z)$. The presence of these two terms reflects the fact that the collapse depends on the expansion history of the universe, and on the initial spectrum of fluctuations. See [44] for the mass functions of sheets, filaments and halos at any given time, in any given cosmology, and for any given

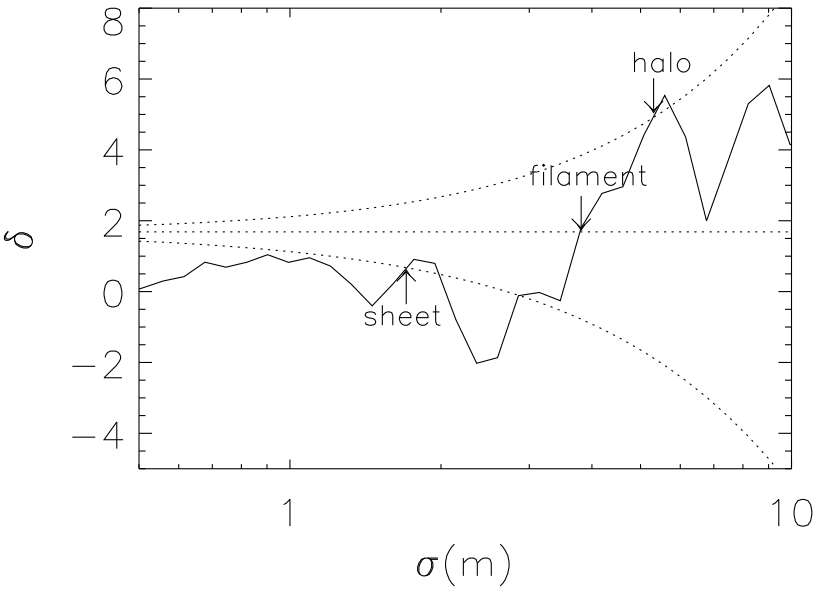

FIGURE 19. An example of a random walk (solid line) crossing the barriers (dotted lines) associated with sheets, filaments and halos (bottom to top). The fraction of walks which first cross the lowest barrier at $\sigma(m_s)$, then first cross the second barrier at $\sigma(m_f)$ and finally cross the highest barrier at $\sigma(m_h)$ represents the mass fraction in halos of mass m_h which are embedded in filaments of mass $m_f > m_h$, which themselves populate sheets of mass $m_s > m_f$ (recall that σ is a decreasing function of m). The precise barrier shapes depend on the collapse model; the dotted curves show the barriers in equation (16) [from 44].

initial fluctuation spectrum, that this model implies.

Application to modified gravity models

In standard gravity, the linear growth factor is independent of k. However, in modified gravity models, the linear growth factor is k-dependent. As a result, although one can still use the excursion set logic, one must be slightly more careful. This is because $D(k,t)/D(k,t_i)$ cannot be taken out of the integral which defines the mapping between S and M. In addition, the height of the walk at one time is not related to that at a different time by the same multiplicative factor for all S. And, typically, $\delta_{\rm sc}$ becomes a function of smoothing scale R and hence mass scale M. In such cases, it is conceptually easier to work with the random walk in the initial field units (rather than linearly extrapolated units). This means that one must convert from M to S using the initial power spectrum

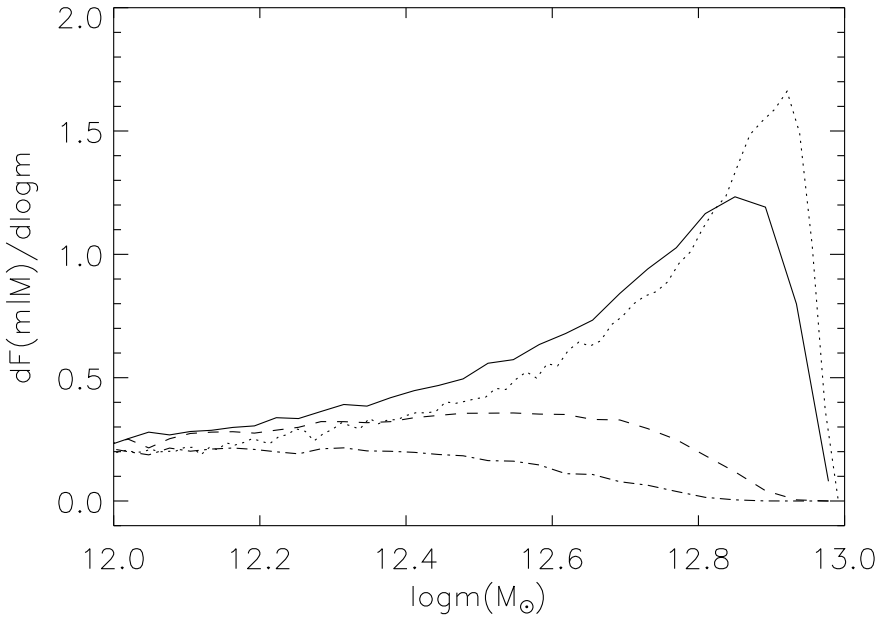

FIGURE 20. Mass fraction of $10^{13}M_{\odot}$ sheets that is in filaments (solid) and halos (dashed) of mass m all at $z = 0$. Dotted curve shows the mass fraction of $10^{13}M_{\odot}$ filaments at $z = 0$ that is in halos. Dot-dashed curve shows the mass fraction contained in halos of mass m within an average volume of the universe of the same mass ($10^{13}M_{\odot}$). The differences between the dotted and dashed curves indicate that, at fixed large-scale overdensity, the halo population is expected to be correlated with the morphology of the surrounding large-scale structure [from 44].

$S_i(M)$, and one must use the critical density $\delta_{\rm sc}(S_i)$ for the initial fi eld, not the linearly evolved one. But otherwise, the logic is the same [27].

Problems, approximations and progress

The excursion set approach cheats in two ways. First, in assuming that the different steps in the walk are uncorrelated with previous ones. Second, in using the ergodic hypothesis to replace spatial averages with ones over an ensemble – in effect, this is an assumption that walks (rather than steps in each walk) are uncorrelated with one another. (There is, of course, a third cheat, which is the assumption that the barrier to be crossed is a well-defi ned deterministic function of δ or (δ, e, p). Some fuzziness is expected – this is what causes some of the scatter in Figure 5 – because even the triaxial model represents a simple approximation to the full dynamics if collapse. We will not address this below.)

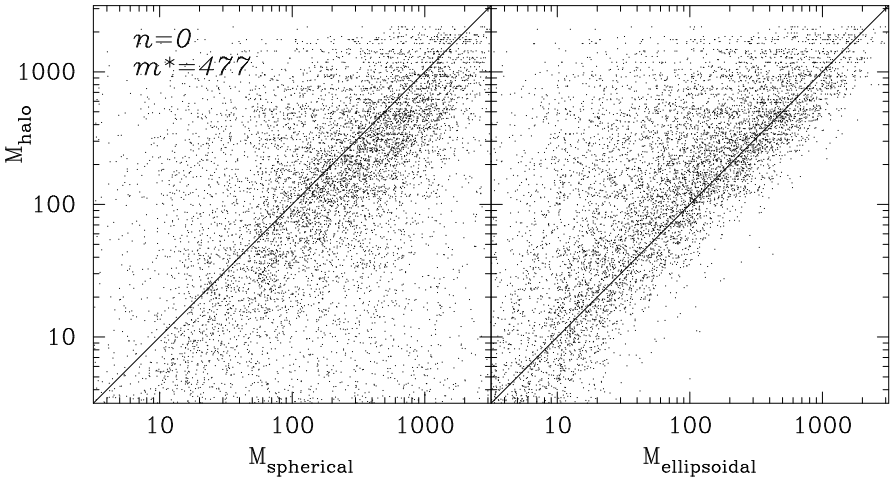

FIGURE 21. The mass of the halo in which a randomly chosen particle is, $M_{\rm halo}$, is plotted versus the mass predicted by the spherical (left panel) and ellipsoidal collapse (right panel) models. The predicted mass for ellipsoidal collapse is smaller, because δ_{ec} increases as m decreases. A randomly chosen 10^4 of the 10^6 particles in a simulation of an Einstein-de Sitter universe with white noise initial conditions were used to make the plot; bjects with $M \gg m_*$ are 'massive' [from 53].

The first has been the subject of some study, because it is relatively straightforward to generate walks with correctly correlated steps, and to then simulate the first crossing distribution. This was done in [9], who showed that one generally predicts a high mass tail which is a factor of two smaller than associated with the analysis above (this has been confirmed by subsequent workers). These mass functions provide worse descriptions of halo abundances in simulations, and so, because they have not yet been described using a simple functional form, they have been largely ignored.

The second has been less studied, but is almost certainly more important. To see why, imagine generating the walk associated with each grid point in the initial conditions. The distribution of first crossing times associated with this bundle of walks yields a prediction for the mass function. This is unlikely to be that different from the approach in which one uses an ensemble of independent walks, because correlations in the initial conditions are not that long-ranged. The real problem induced by spatial correlations is somewhat different.

Suppose one places a particle at each initial grid point, and one has used the excursion set to predict the mass of the object it will be in at some later time. If one plots this predicted mass versus the mass of the object in which it actually ends-up, the result (left hand panel of Figure 21) is a scatter plot – something which was highlighted in [66]. Using the triaxial collapse model instead leads to smaller predicted masses (right hand panel); this reduces some of the scatter, but the tendency for the particle to be in a more massive halo than predicted by its walk remains.

What causes this? Consider two neighbouring particles in the initial conditions which ended up in the same object, but for which the predicted mass is different. Which of

the predictions came closer to the correct answer? [53] argued that it is likely to be the larger of the two, for the same reason that the first crossing distribution in the excursion set approach is so special. Indeed, if the predicted mass is m, and it is correct, then this means that the predictions of *all* the walks within R_m of it should be discarded, because they are *guaranteed* to underestimate the correct mass, and so have no business playing any further role in the determination of $dn(m)/d\ln m$! Comparison with simulations showed that, indeed, of the bundle of masses predicted for an object, the largest was in general much closer to the actual mass. This brings us to an important realization: the true halo abundance distribution must be shifted to larger masses than one predicts when simply inserting equation (19) into (20). This is almost certainly the reason why halo abundances in simulations are better described by a value for δ_{sc} that is about $\sqrt{0.75}$ lower than expected.

To account for this effect, the discussion above shows that one must insert one more step between the first crossing distribution and the mass function (between equations 19 and 20). For instance, consider a walk whose first crossing distribution predicts mass m. At the very least, one would like to ensure that

- all the other walks that are within R_m of this one predict smaller masses;
- and that this walk itself is further than R_M from all walks for which the predicted mass is $M > m$.

Incorporating this effect is a tough but interesting open problem, for which a crude estimate can be got as follows.

Let $\phi(m)$ denote the quantity which should be on the right hand side of equation (20). If $p(M|m)$ denotes the probability that a walk which was predicted to have mass m actually ends up in a halo of mass M, then

$$\phi(M) = f(M) + \int_0^M dm\, f(m)\, p(M|m) - f(M) \int_M^\infty dM'\, p(M'|M); \tag{33}$$

the second term counts the increase in the abundance of M because of this effect, and the third counts the decrease, as, for similar reasons, objects originally predicted to have mass M are assigned to more massive objects. Rearranging the order of the integrals in the second term shows that

$$\phi(>M) = f(>M) + \int_0^M dm\, f(m) \int_M^\infty dM'\, p(M'|m). \tag{34}$$

Since all quantities in the final term on the right hand side are positive, ϕ will be shifted towards higher mass scales than f.

To proceed further, we require an estimate of $p(M|m)$ which incorporates both the effects itemized above. Of the two, the first is easier to estimate, because it deals only with the sphere of radius R_m, which we know has overdensity δ_{sc} when smoothed with a filter of scale R_m. The overdensity smoothed with the same filter, but displaced r from the center of this one, will have a distribution given by

$$p_m(\delta, r|\delta_{sc}, 0)\, d\delta \approx \frac{\exp(-y^2/2)}{\sqrt{2\pi}}\, dy, \qquad \text{where} \quad y = \frac{\delta - \rho(m,r)\delta_{sc}}{\sqrt{S(m)\,[1-\rho^2(m,r)]}}, \tag{35}$$

and

$$S(m)\,\rho(m,r) \equiv \int \frac{\mathrm{d}k}{k}\,\frac{k^3 P_{\rm L}(k,t)}{2\pi^2}\,|W(kR_m)|^2\,\frac{\sin kr}{kr}. \tag{36}$$

The approximation in using the conditional Gaussian distribution assumes that the fact that $\delta_M(0) < \delta_{sc}$ for all larger smoothing scales centered on the origin matters little. The expression above shows that there is some chance that $\delta(r)$ will exceed δ_{sc}. If it does, then the spherical collapse model suggests that we should associate this patch, not with mass m, but with some $M > m$. The chance that this happens somewhere within m is given by integrating p_m over the volume of m, so a first estimate of $p(M|m)$ comes from setting

$$\int_m^\infty dM\,p(M|m) \approx 3\int_0^{R_m}\frac{dr\,r^2}{R_m^3}\int_{y_{min}}^\infty d\delta\,\frac{\exp(-y^2/2)}{\sqrt{2\pi}} \quad \text{where} \quad y_{min} = \frac{\delta_{sc}}{\sqrt{S}}\sqrt{\frac{1-\rho}{1+\rho}}. \tag{37}$$

and differentiating with respect to m. It will be interesting to see how accounting for this effect changes the expected functional form (and the near universal scaling behaviour) of the halo mass function.

THE HALO MODEL

The Halo Model [see 13, for a review] provides an easy way to see how different point processes can all be related to the same underlying dark matter density field. This makes it a useful language for discussing how galaxy clustering depends on galaxy type: galaxy bias. This approach represents the following shift in paradigm. Whereas previous work (typically based on perturbation theory) used the dark matter density field as the fundamental quantity of interest, in the Halo Model, it is halo mass which is fundamental. Thus, in the Halo Model, one predicts environmental trends for the galaxy population because different galaxy types populate different mass halos, and the halo mass function is top heavy in dense regions. As a result, one does *not* attempt to explain correlations with environment (or measurements such as those shown in Figure 1) by modeling the physical effects of the large scale density, pressure or temperature fields on smaller scale galaxies. Rather, an extreme statement of this shift in paradigm is that progress is best made by trying to model how the formation history of a halo determines the properties of the galaxies it hosts, and that, since halo mass and formation history are tightly correlated, one should think of any given galaxy population as a weighted sum over the halo distribution.

Two-point statistics

The Halo Model is simplest in Fourier space, where real-space convolutions become multiplications. The real-space two-point correlation function $\xi(r)$ is obtained by Fourier transforming the power spectrum $P(k)$, which, in the halo model, is written as the sum of two terms. One arises from galaxies within the same halo and the other from

galaxies in different halos. Because halos are small compared to the separations between them, the first term, the 1-halo term, dominates on small scales, whereas the other, the 2-halo term, dominates on larger scales. The key insight gained from this approach is that what is true for the statistics is also true of the physics. Thus, the 1-halo term incorporates the nonlinear physics associated with virialized structures, whereas the 2-halo term exploits decades of work on perturbation theory. This also means that we expect to see a feature on the scale where the signal changes from being dominated by the 1-halo term to the other; this scale is related to the virial radii of the halos producing the signal. Thus, we can begin to interpret physically the bumps and wiggles in Figures 1 and 2.

For galaxies, the halo model distinguishes between the central galaxy in a halo and all the others, which are sometimes called satellites. (The galaxy is assumed to sit at the halo center in halos which contain only one galaxy.) This is because, in semi-analytic and SPH and galaxy formation models, central and satellite galaxies are rather different populations [22, 49**?**] Thus,

$$P(k) = P_{1h}(k) + P_{2h}(k), \tag{38}$$

where

$$\begin{aligned} P_{1h}(k) &= \int dm \frac{dn(m)}{dm} \langle N_{\rm cen}|m\rangle \left[\frac{2\langle N_{\rm sat}|m\rangle u_{\rm gal}(k|m)}{\bar{n}_{\rm gal}^2} + \frac{\langle N_{\rm sat}(N_{\rm sat}-1)|m\rangle u_{\rm gal}(k|m)^2}{\bar{n}_{\rm gal}^2} \right], \\ P_{2h}(k) &= \left[\int dm \frac{dn(m)}{dm} \langle N_{\rm cen}|m\rangle \frac{1+\langle N_{\rm sat}|m\rangle u_{\rm gal}(k|m)}{\bar{n}_{\rm gal}} b_1(m) \right]^2 P_{\rm Lin}(k), \end{aligned} \tag{39}$$

where the number density of galaxies $\bar{n}_{\rm gal}$ is

$$\bar{n}_{\rm gal} = \int dm \frac{dn(m)}{dm} \langle N_{\rm cen}|m\rangle \left[1 + \langle N_{\rm sat}|m\rangle\right] \tag{40}$$

and $u_{\rm gal}(k|m)$ is the Fourier transform of the galaxy density profile. (It is standard to assume this has the same form as for the dark matter, for which there is a good fitting function [32, 41] but no complete theory!). The other inputs to these expressions are the halo mass function dn/dm and halo bias factors $b(m)$ (for which we developed models in the previous sections), a prescription for how galaxies populate these halos (only the first and second moments of $p(N_{\rm gal}|m)$ matter for two-point statistics; n-point moments matter for n-point statistics), and the linear perturbation theory power spectrum.

The two parts of the 1-halo term in equation (39) can be thought of as the 'center-satellite term' and the 'satellite-satellite term'. The distribution $p_{\rm sat}(N_{\rm sat}|m)$ is expected to be approximately Poisson [24] so $\langle N_{\rm sat}(N_{\rm sat}-1)|m\rangle = \langle N_{\rm sat}|m\rangle^2$, and the entire model is specified by how $\langle N_{\rm cen}|m\rangle$ and $\langle N_{\rm sat}|m\rangle$ depend on halo mass.

Weights in the halo model

Galaxies have a range of luminosities, colors, environments, etc. The Halo Model was originally formulated to describe the point process which is associated with selecting a

galaxy sample based on one or more of these properties, and then treating all galaxies in the sample as being equivalent. However, it is sometimes desirable (e.g. when sample sizes are small) to include all galaxies, but weight each according to one or more of these properties when computing the clustering signal. The halo model can also be used to describe such measurements, which are known in the point-process literature as Mark Statistics [48].

Use $W(k)$ to denote the Fourier transform of the weighted correlation function. Like the power spectrum, write this as the sum of 1- and 2-halo terms: $W(k) = W_{1h}(k) + W_{2h}(k)$. Since central and satellite galaxies have different properties, central and satellite galaxies are weighted separately by their mean mass-dependent marks: $\langle w|m\rangle_{\rm cen}$ and $\langle w|m\rangle_{\rm sat}$. Then

$$W_{1h}(k) = \int dM \frac{dn(M)}{dM} \langle N_{\rm cen}|M\rangle \left[\frac{2\, w_{\rm cen}(M) \langle w_{\rm sat}|M, L_{\rm min}\rangle \langle N_{\rm sat}|M\rangle\, u_{\rm gal}(k|M)}{\bar{n}_{\rm gal}^2\, \bar{w}^2} + \frac{\langle N_{\rm sat}|M\rangle^2 \langle w_{\rm sat}|M, L_{\rm min}\rangle^2\, u_{\rm gal}^2(k|M)}{\bar{n}_{\rm gal}^2\, \bar{w}^2} \right], \quad (41)$$

$$\frac{W_{2h}(k)}{P_{\rm Lin}(k)} = \left[\int dM \frac{dn(M)}{dM} \langle N_{\rm cen}|M\rangle\, b(M) \frac{w_{\rm cen}(M) + \langle N_{\rm sat}|M\rangle \langle w_{\rm sat}|M, L_{\rm min}\rangle\, u_{\rm gal}(k|M)}{\bar{n}_{\rm gal}\, \bar{w}} \right]^2,$$

where the mean mark is

$$\bar{w} = \int dM \frac{dn(M)}{dM} \langle N_{\rm cen}|M\rangle \frac{w_{\rm cen}(M) + \langle N_{\rm sat}|M\rangle \langle w_{\rm sat}|M, L_{\rm min}\rangle}{\bar{n}_{\rm gal}}. \quad (42)$$

Implementation: HODs, CLFs, SHAMs

To date, the Halo Model it has been used to provide a useful framework for modeling the luminosity dependence of galaxy clustering, and the dependence of clustering on environment. The first is usually done in three rather different ways, which have come to be known as the 'halo occupation distribution' (HOD; [HOD; 21, 4, 43, 41, 7, 70] the 'conditional luminosity function' [CLF; 36, 68, 12, 65], and the 'subhalo abundance matching' [SHAM; 23, 24, 64, 11] methods.

The HOD approach uses the abundance and spatial distribution of a given galaxy population (typically, just the two-point clustering statistics) to determine how the number of galaxies depends on the mass of the parent halo. This is done by studying a sequence of volume limited galaxy catalogs, each containing galaxies more luminous than some threshold luminosity. The CLF method attempts, instead, to match the observed luminosity function by specifying how the luminosity distribution in halos changes as a function of halo mass. One can infer the CLF from the HOD approach, and vice-versa, so the question arises as to which is the more efficient description. For a given catalog, the HOD method requires the fitting of just two free parameters, so it is relatively straightforward. The CLF method requires many more parameters to be fit simultaneously, but uses fewer volume limited catalogs. SHAMs first identify the subhalos within virialized

halos in simulations, and then use subhalo properties to match the subhalo abundances to the observed distribution of luminosities. Once this has been done, CLFs or HODs can be measured in the simulations.

In the HOD and CLF approaches to the halo model, the central galaxy in a halo is assumed to be very different all the others, which are called satellites. For example, the CLF approach must provide a description of how the central and satellite luminosity functions vary as a function of halo mass. The HOD-based analyses predict that the satellite galaxy luminosity function should be approximately independent of halo mass, and hence of group and/or cluster properties [60]. [61] present evidence from the SDSS in support of this surprising and unexpected prediction. This independence can reduce the required number of free parameters in CLF-based analyses. In contrast, CLF-based approaches have yet to inform HOD-analyses.

The HOD-based approach also provides a rather simple way to understand how galaxy clustering depends on color [62]. In essence, it provides a simple algorithm for specifying how the joint CLF (i.e., the luminosity distribution in two different bands) varies with halo mass. The method exploits the fact that, to a good approximation, galaxies appear to be bimodal in their properties [8], and, in particular, the distribution of colors at fixed luminosity is bimodal [2, 67]. This is an important step towards the ultimate goal of providing a description of how the properties of a galaxy, its morphology and spectral energy distribution, are correlated with those of its neighbors.

Implicit assumptions and mock catalogs

The Halo Model description above makes three simplifying assumptions which are worth discussing explicitly. First, although we assume halos are spherical and smooth, the density run of satellites around halo centers is almost certainly neither. Generating triaxial distributions is straightforward once prescriptions for how the triaxiality depends on halo mass and how it correlates with environment are available. Once these are known, they can be incorporated into the analytic halo-model description [63]. Similarly, parametrizations of halo substructure can also be incorporated into the description [51]. Of course, both these types of correlations can be included in a 'mock catalog', if one identifies halos in a simulation, and then simply selects the appropriate number of particles (satellites from a Poisson distribution with mean which depends on halo mass) from the halo itself.

Second, note that the number of galaxies in a halo, the spatial distribution of galaxies within a halo, and the assignment of luminosities all depend only on halo mass. None of these depend on the surrounding large-scale structure. Therefore, a mock catalog constructed in this way includes only those environmental effects which arise from the environmental dependence of halo abundances.

Third, halos of the same mass will have had a variety of formation histories. Some will have assembled their mass and their galaxy populations more recently than others. Recent assembly means less time for dynamical friction, and, possibly, a younger stellar population. So, at fixed halo mass, one might expect to find a correlation between the age of a halo and the galaxy population within it. In particular, the number of galaxies

in a halo, their luminosities and their colors may all be correlated with the formation history. Our halo model description (and associated mock catalog) ignores all such correlations. Had we used a SHAM to assign luminosities, then some of correlation between formation history and the galaxy population will have been included. If one is already carrying along the particle distribution from the simulation to construct the mock, then the next level of complication is to also include additional information about the merger history in the simulation, for use when making the mock.

Current implementations also assign colors to satellite galaxies without explicit consideration of the color of the central galaxy, and they make no effort to incorporate color gradients within a halo into our model. This is mainly because the measurements to date are on large enough scales that gradients matter little (see [50, 42] for more discussion and [48] for simple prescriptions for incorporating gradients.) These are all interesting problems for the future (and they are almost certainly not independent problems!), but the measurements to date, such as those shown in Figure 1, do not require these refinements. In the future, the Halo Model approach will be used to understand the evolution of galaxy clustering. This can be done because the evolution of the halo abundances, halo bias, and halo profiles, are known, so the only required new ingredient is how $p(N_{\rm gal}|m)$ evolves. See [50] for an alternative approach in which the halo abundance is kept fixed.

ACKNOWLEDGMENTS

I thank J. Colberg, T. Y. Lam, M. Martino, J. Moreno, G. Rossi and G. Tormen; many of the plots here were made in collaboration with them. Thanks also to Mario Novello for the invitation to attend this school, M. Makler for being an outstanding host, R. Rosenfeld for discussions about the future of Dark Energy and the financial market, V. Miranda for raising so many interesting questions, M. Calvao for keeping him under control, S. Bergliaffa for his patience, and the other lecturers for an outstanding series of talks from which I learned much.

REFERENCES

1. Abbas U., Sheth R. K., 2007, MNRAS, 378, 641
2. Baldry I. K., Glazebrook K., Brinkmann J., et al., 2004, ApJ, 600, 681
3. Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986 ApJ, 304, 15
4. Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000, MNRAS, 311, 793
5. Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, ApJ, 599, 38
6. Bernardeau F., Colombi S., Gaztanaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1
7. Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
8. Blanton M., Hogg D. W., Bahcall N. A., et al., 2003, ApJ, 594, 186
9. Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
10. Bond J. R., Myers S. T. 1996, ApJS, 103, 1
11. Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
12. Cooray A., 2006, MNRAS, 365, 842
13. Cooray A., Sheth R. K., 2002, Phys. Rep., 372, 1
14. Doroshkevich A. G., 1970, Astrofizika, 3, 175
15. Frenk C. S., White S. D. M., Davis M., Efstathiou G., 1988, ApJ, 327, 507
16. Fry J. N., Gaztanaga E., 1993, ApJ, 413, 447

17. Giocoli C., Moreno J., Sheth R. K., 2007, MNRAS, 376, 977
18. Goldberg D. M., Vogeley M. S., 2004, ApJ, 605, 1
19. Gunn J. E., Gott J. R., 1972, ApJ, 176, 1
20. Jenkins A., Frenk C.S., White S.D.M., et al., 2001, MNRAS, 321, 372
21. Jing Y., Mo H. J., Börner G., 1998, ApJ, 494, 1
22. Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999, MNRAS, 303, 188
23. Klypin A., Gottlöber S., Kravtsov A. V., Khokhlov A. M., 1999, ApJ, 516, 530
24. Kravtsov A. V., Berlind A., Wechsler R. H., et al., 2004, ApJ, 609 35
25. Lacey C., Cole S., 1993, MNRAS, 262, 627
26. Lam T. Y., Sheth R. K., 2008, MNRAS, 389, 1249
27. Martino M. C., Sheth R. K., 2009, MNRAS, in press
28. Martino M. C., Stabenau F., Sheth R. K., 2009, PRD, accepted
29. Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
30. Moreno J., Giocoli C., Sheth R. K., 2008, MNRAS, 391, 1729
31. Moreno J., Giocoli C., Sheth R. K., 2009, MNRAS, in press
32. Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
33. Neistein E., van den Bosch F. C., Dekel A., 2006, MNRAS, 372, 933
34. Nusser A., Sheth R. K., 1999, MNRAS, 303, 685
35. Padmanabhan T., 1993, Structure formation in the Universe, Cambridge University Press, UK
36. Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
37. Peebles P. J. E., 1980, The Large-Scale Structure of the Universe. Princeton Univ. Press, Princeton
38. Press W., Schechter P., 1974, ApJ, 187, 425
39. Reed D., Gardner J., Quinn T., et al., 2003, MNRAS, 346, 565
40. Schechter P. L., 1980, AJ, 85, 801
41. Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, ApJ, 546, 20
42. Scranton R., 2002, MNRAS, 332, 697
43. Seljak U., 2000, MNRAS, 318, 203
44. Shen J., Abel T., Mo H. J., Sheth R. K., 2006, ApJ, 645, 783
45. Sheth R. K., 1996, MNRAS, 281, 1277
46. Sheth R. K., 1998, MNRAS, 300, 1057
47. Sheth R. K., 2003, MNRAS, 345, 1200
48. Sheth R. K., 2005, MNRAS, 364, 796
49. Sheth R. K., Diaferio A., 2001, MNRAS, 322, 901
50. Sheth R. K., Diaferio A., Hui L., Scoccimarro R., 2001, MNRAS, 326, 463
51. Sheth R. K., Jain B., 2003, MNRAS, 345, 529
52. Sheth R. K., Lemson G., 1999, MNRAS, 304, 767
53. Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
54. Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
55. Sheth R. K., Tormen G., 2002, MNRAS, 329, 61
56. Sheth R. K., Tormen G., 2004, MNRAS, 349, 1464
57. Sheth R. K., van de Weygaert, 2004, MNRAS, 350, 517
58. Shirata A., Shiromizu T., Yoshida N., Suto Y., 2005, PRD, 71, 064030
59. Shirata A., Suto Y., Hikage C., Shiromizu T., Yoshida N., 2007, PRD, 76, 044026
60. Skibba R. A., Sheth R. K., Connolly A. J., Scranton R., 2006, MNRAS, 369, 68
61. Skibba R. A., Sheth R. K., Martino M. C., 2007, MNRAS, 382, 1940
62. Skibba R. A., Sheth R. K., 2009, MNRAS,
63. Smith R. E., Watts P. I. R., Sheth R. K., 2006, MNRAS, 365, 214
64. Vale A., Ostriker J. P., 2006, MNRAS, 371, 1173
65. van den Bosch F. C., Yang X., Mo H. J., et al., 2007, MNRAS, 376, 841
66. White S. D. M., 1996, in Schaeffer R. et al., eds, Cosmology & Large–scale Structure, Proc. 60th Les Houches School, Elsevier, p.349
67. Willmer C. N. A., Faber S. M., Koo D. C., et al., 2006, ApJ, 647, 853
68. Yang X., Mo H. J., van den Bosch F. C., 2003, MNRAS, 339, 1057
69. Zel'dovich Ya. B., 1970, Astrofizika, 6, 319 (translated in Astrophysics, 6, 164 [1973])
70. Zehavi I., et al., 2005, ApJ, 630, 1