Galaxy Formation

A galaxy is the environment in which stars are born and die, and distant galaxies are the luminous beacons that enable us to probe the distant universe. Our galaxy, the Milky Way, is one of billions of such systems in the observable universe. How galaxies formed represents a central theme in modern cosmology.

At first glance, the universe appears to contain two different types of galaxies: galaxies with disk-like morphologies and galaxies with spheroidal morphologies. This basic distinction breaks down, however, once individual disk and spheroidal galaxies are examined in detail, since most disk galaxies contain small spheroidal components at their centers and most spheroidal galaxies contain small disks at their centers.

Disks in and of themselves contain a wide variety of features. Most DISK GALAXIES exhibit spiral arms with a large range of winding angles and contrast. Approximately half of all disk galaxies also contain a highly elongated bar structure near their center, the bars possessing a variety of axial ratios. Some disks also show moderate deviations from planarity toward their edges (warps) while other disks show significant lanes of dust across their observed profiles. Similarly, ELLIPTICAL GALAXIES, while possessing relatively uniform profiles compared to disk galaxies, show a significant variety of substructure. At least half of ellipticals have detectable shell structure and others distinct cores. Finally, all galaxies, irrespective of type, show great variations in the amounts and spatial distribution of gas, dust, stars, and metal abundances as well as their basic surface brightnesses, luminosities, colours, and masses.

Despite their considerable diversity, galaxies show a remarkable degree of uniformity as well. The profiles of disk and ellipticals are remarkably homologous, the global structural parameters of disk and spheroidal galaxies define a tight two-dimensional plane, and the colours and apparent star formation histories of both spiral and elliptical galaxies show a striking correlation with luminosity. Of great significance is that most galaxies are very slowly evolving structures, both chemically and dynamically. Their properties were acquired long ago, at or soon after the epoch of galaxy formation.

Galaxies began as clouds of primordial gas, hydrogen and helium. Even before galaxies condensed into distinct clouds, infinitesimal density fluctuations were present in the EXPANDING UNIVERSE. These originated at very early epochs in an inflationary phase transition from a universe that initially was relatively uniform. Fluctuations grew in strength under the inexorable influence of self-gravity. Eventually, clouds developed that fragmented into stars. Much of the detailed physics in this schematic of GALAXY EVOLUTION is now understood.

This review begins with a discussion of the cosmological world model in which galaxies form, discusses the processes by which the initial perturbations are established, presents the theory for the growth of these perturbations into collapsing and eventually virialized objects, illustrates the importance of gas cooling in the formation of galaxies, outlines the processes by which galaxies acquire angular momentum, and concludes by summarizing the basic observations and theory of disk and elliptical galaxy formation.

World model

We begin by providing some background on the standard world model and the primordial fluctuations out of which galaxies are believed to have grown.

The apparent homogeneity and isotropy of the observable universe, both in terms of its large-scale structure and the cosmic infrared microwave background radiation—the almost constant 2.73 K blackbody radiation background in which the universe is immersed—motivate the assumption that the universe is both homogeneous and isotropic. By homogeneous, we mean that every point in space statistically resembles every other point in space. By isotropic, we mean there is no point in space where any direction differs statistically from any other direction in space.

Assuming universal homogeneity and isotropy, Einstein's theory of GENERAL RELATIVITY can be used to show that the evolution of the scale of the universe follows the two independent equations:

$$\frac{\dot{a}^2 + k}{a^2} = \frac{8\pi G}{3}\rho \tag{1}$$

$$\frac{2\ddot{a}}{a} + \frac{\dot{a}^2 + k}{a^2} = -8\pi G p \tag{2}$$

collectively known as Friedmann's equations, where a is a measure of the size of the universe, p is the pressure, k is the curvature, ρ is the density, and G is Newton's constant. Combined with the equation of state, these equations completely determine $a(t)$, $\rho(t)$, and $p(t)$.

These equations can be recast into the form:

$$\frac{\ddot{a}}{a} = H_0^2\left[\Omega_{\Lambda,0} - \Omega_0(1+z)^3/2\right] \tag{3}$$

$$\begin{aligned}\frac{\dot{a}}{a} &= H_0 E(z) \\ &= H_0\sqrt{\Omega_0(1+z)^3 + \Omega_{R,0}(1+z)^2 + \Omega_{\Lambda,0}}\end{aligned} \tag{4}$$

where $\Omega_0 = 8\pi G\rho_0$, $\Omega_{R,0} = 1/(H_0 a_0 R)^2$, $\Omega_{\Lambda,0} = \Lambda/3H_0^2$, ρ_0 is the matter density of the universe at the present epoch, Λ is the vacuum energy density or cosmological constant, and R is a constant with units of length. Hubble's constant, denoted by H_0, characterizes the rate at which the universe is expanding at the present epoch. It is the constant of proportionality relating an object's distance D to its rate of recession v:

$$v = H_0 D. \tag{5}$$

Note that $\Omega_{R,0} = 1 - \Omega_0 - \Omega_{\Lambda,0}$.

As the universe expands, the gravitational attraction of the mass inside it slows this expansion. This deceleration may or may not be enough to slow this expansion sufficiently so that the universe recollapses. The case where there is sufficient matter to cause such a recollapse corresponds to a universe where the universal geometry is closed, i.e., $k = +1$. The case where there is not sufficient matter to force such a recollapse corresponds to two separate geometries: one in which the universal geometry is flat ($k = 0$) and one in which the universal geometry is open ($k = -1$). A universe with a flat geometry is known as Einstein–de Sitter.

The time evolution of the universal scale length a is amenable to the following simple analytic solution in the case of an Einstein–de Sitter universe where ordinary matter dominates the energy density ($p = 0; \rho \propto (1+z)^{-3}$):

$$a = a_0(t/t_0)^{2/3} \tag{6}$$

where t_0 and a_0 are the current age and size of the universe, respectively.

For an open universe, the solution is given in terms of the following parametric equations:

$$a = \frac{\Omega_0}{(1-\Omega_0)^{3/2}} \frac{c}{2H_0} (\cosh\Theta - 1) \tag{7}$$

$$t = \frac{\Omega_0}{(1-\Omega_0)^{3/2}} \frac{1}{2H_0} (\sinh\Theta - \Theta) \tag{8}$$

while for a closed universe, the parametric equations are

$$a = \frac{\Omega_0}{(\Omega_0-1)^{3/2}} \frac{c}{2H_0} (1 - \cos\Theta) \tag{9}$$

$$t = \frac{\Omega_0}{(\Omega_0-1)^{3/2}} \frac{1}{2H_0} (\Theta - \sin\Theta). \tag{10}$$

The turn-around time t_m for this universe occurs when $\Theta = \pi$, so from equation (10), we find

$$t_\mathrm{m} = \frac{\Omega_0}{(\Omega_0-1)^{3/2}} \frac{\pi}{2H_0} \tag{11}$$

for the turn-around time.

Requiring that the curvature be identical everywhere in space-time, the most general way of expressing the concept of distance is the Friedmann–Robertson–Walker metric:

$$\mathrm{d}l^2 = a^2R^2\left[\mathrm{d}\chi^2 + f^2(\chi)(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2)\right] \tag{12}$$

where

$$f(\chi) = \begin{cases} \sin\chi, & k = +1 \\ \chi, & k = 0 \\ \sinh\chi, & k = -1. \end{cases} \tag{13}$$

The two-dimensional analogue to the Friedmann–Robertson–Walker metric is

$$\mathrm{d}l^2 = a^2\left[\mathrm{d}\chi^2 + f^2(\chi)\mathrm{d}\theta^2\right] \tag{14}$$

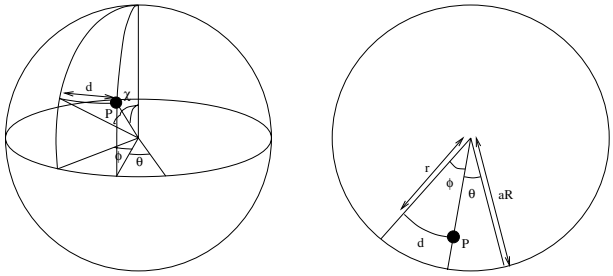

Figure 1. The topology of space in a closed ($k = +1$) two-dimensional universe. The coordinates (χ, θ) define the position of a point P in this universe. The angular width ϕ of an object of size d is illustrated.

which is more amenable to our everyday intuition, especially in the case of a closed $k = +1$ universe where

$$\mathrm{d}l^2 = a^2\left[\mathrm{d}\chi^2 + \sin^2\chi\,\mathrm{d}\theta^2\right]. \tag{15}$$

This is simply the expression for the distance on the surface of a sphere where θ is the azimuthal coordinate and where χ is the tangential coordinate (figure 1).

The wavelengths of photons expand, or REDSHIFT, along with the universe. The redshift of an object is a measure of how much the universe and, therefore, the wavelength of that object's photons have expanded until the present. A redshift of zero indicates the present. Quantitatively, we express the relationship between the size of the universe a and its redshift z as

$$a \propto \frac{1}{1+z}. \tag{16}$$

Using solutions to Friedmann equations, one may derive both the age of the universe and the effective distances to objects which existed at earlier epochs. By integrating up the infinitesimal times

$$\begin{aligned} \mathrm{d}t = \frac{1}{\dot{a}}\,da &= \frac{1}{aH_0E(z)} \frac{a\,\mathrm{d}z}{1+z} \\ &= \frac{\mathrm{d}z}{H_0E(z)(1+z)} \end{aligned} \tag{17}$$

one can compute the age of the universe:

$$t = \frac{1}{H_0} \int_{z=0}^{\infty} \frac{\mathrm{d}z}{E(z)(1+z)}. \tag{18}$$

Similarly, one may readily derive expressions for the distance though there is one subtlety. Two different measures of distance are standardly discussed in cosmology: the angular-size distance and the luminosity distance. Both distances are defined so that the standard expressions involving these quantities apply. The former, the angular-size distance, commonly denoted D_A, is defined in analogy with the expression

$$\theta = \frac{d}{D_\mathrm{A}} \tag{19}$$

where d is an object's intrinsic size and θ is the angular size of the object on the sky in radians. Similarly, the latter, the luminosity distance, commonly denoted $D_{\rm L}$, is defined in analogy with the expression

$$f = \frac{L}{4\pi D_{\rm L}^2} \tag{20}$$

where f is the observed flux and L is the intrinsic luminosity.

We now provide a heuristic derivation of the above equations. Imagine that the light from some object with size d, redshift $z_{\rm obs}$, and tangential coordinate χ converges to $\chi = 0$ and $z = 0$ on paths where $\mathrm{d}\theta = 0$. Along this path, the expression for the metric reduces to $\mathrm{d}l^2 = a^2R^2\mathrm{d}\chi^2$. The integrated coordinate distance χ is then

$$\begin{aligned}\chi &= \int_{z=0}^{z_{\rm obs}} \frac{\mathrm{d}l}{aR} = \int_{z=0}^{z_{\rm obs}} \frac{c\,\mathrm{d}t}{aR} \\ &= \frac{1}{H_0 a_0 R}\int_{z=0}^{z_{\rm obs}} \frac{c\,\mathrm{d}z}{E(z)}.\end{aligned} \tag{21}$$

The object is observed at tangential coordinate χ at some previous time, and the distance of this object (at χ and $z_{\rm obs}$) from the z-axis (see figure 1) is

$$r(z) = a(z)R\sin\chi = \frac{a_0 R}{1+z}\sin\chi. \tag{22}$$

Clearly, the angle ϕ emanating from the top of the sphere intersecting our object of size d at some distance r from the z-axis is d/r, so

$$\phi = \frac{d}{r}. \tag{23}$$

Identifying $D_{\rm A}$ in equation (19) with $r(z)$ in equation (23) yields

$$\begin{aligned}D_{\rm A} = r(z) &= \frac{a_0 R}{1+z}\sin\chi \\ &= \frac{a_0 R}{1+z}\sin\left[\frac{1}{H_0 a_0 R}\int_{z=0}^{z_{\rm obs}}\frac{c\,\mathrm{d}z}{E(z)}\right].\end{aligned} \tag{24}$$

A heuristic derivation of the luminosity distance is similarly possible. Imagine the light emitted from some object at redshift $z_{\rm obs}$ and $\chi = 0$ propagates to some χ reaching it at $z = 0$. This light would have spread out over a ring of radius $2\pi r(0)$, where $r(0)$ is again the distance from the points at χ and the z-axis. Hence, $f = L/2\pi r(0)$. The obvious three-dimensional analogue is $f = L/4\pi r(0)^2$. Multiplying the flux by a factor of $1/(1+z)$ to account for time dilation and $1/(1+z)$ to account for the energy loss due to photon redshifting, one obtains

$$f = \frac{L}{4\pi r(0)^2(1+z)^2}. \tag{25}$$

Identifying $r(0)(1+z)$ in equation (25) with $D_{\rm L}$ in equation (20) yields

$$\begin{aligned}D_{\rm L} &= r(0)(1+z) \\ &= a_0 R(1+z)\sin\left[\frac{1}{H_0 a_0 R}\int_{z=0}^{z_{\rm obs}}\frac{c\,\mathrm{d}z}{E(z)}\right].\end{aligned} \tag{26}$$

As in our derivation of the angular and luminosity distances, it is easy to see that the apparent surface brightness of objects decreases as $(1+z)^{-4}$, a factor of $(1+z)^{-1}$ due to time dilation, a factor of $(1+z)^{-1}$ due to a redshifting of the photons, and a factor of $(1+z)^{-2}$ to account for its larger angular size on the sky.

Origin of fluctuations

INFLATION is a popular scenario for producing the small matter overdensities, or seeds, on which mass accretes and galaxies later form. The popularity of inflation derives from the relatively natural explanation it provides for establishing the initial conditions from which our universe seems to have evolved. Not only does it explain the homogeneity of the universe on large scales, but it also solves the problem of the relative absence of objects like magnetic monopoles or other types of massive topological defects. It provides a natural explanation for the apparent flatness of the universe although the observational evidence for this still might be considered preliminary.

Inflation is initiated at the temperature scale corresponding to the breaking of the symmetry of grand unification ($T \sim 10^{16}$ GeV, at $t \sim 10^{-35}$ s) as the universe expands and cools through a brief period of supercooling. The associated release of scalar field energy results in an exponential increase in the scale factor, and corresponding reduction in temperature, and is followed by a period of reheating when the associated field kinetic energy thermalizes. This early epoch of exponential expansion is driven by the evolution of a scalar inflation field ϕ and potential V, the field evolving as

$$\ddot{\phi} + 3\frac{\dot{a}}{a} + \frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} = 0. \tag{27}$$

The pressure and density are

$$p_\phi = \dot{\phi}^2 + V \tag{28}$$

$$\rho_\phi = \dot{\phi}^2 - V \tag{29}$$

where the inflation potential is such that $V >> \dot{\phi}^2$, so $\rho = -p = V$ as in the case of a non-zero cosmological constant.

Zero point fluctuations of the scalar field are imprinted on the causal horizon scale, which increases exponentially. The causal horizon is greatly enlarged, to encompass all of the universe observed today, and the reheating process imprints a scale-invariant distribution of energy density fluctuations. A perturbative analysis of this scalar field's evolution predicts a power spectrum with slope close to unity: $P(k) \propto k$, where k is the wavenumber defined via the Fourier transformation of the underlying Gaussian-distributed density field $\rho(x,t)$: $\delta_k = \int \rho(x,t)\mathrm{e}^{\mathrm{i}k\cdot x}\mathrm{d}^3x$. This power spectrum is called the Harrison–Zeldovich scale-free power spectrum. The amplitude of the fluctuations is not predicted by the

inflationary theory. However the prediction of scale-invariance is generic to many, although not all, inflationary models. Scale-independent curvature fluctuations are equivalent to density fluctuations on the horizon scale of fixed amplitude. On superhorizon scales, an invariant $\delta\phi = G\,\delta M/L\,c^2 = \delta\rho/\rho)(L/ct)^2 = (\delta\rho/\rho)\,L^2_{\rm comoving}\,(a/ct)^2$ can be found which does not change with time.

The energy density of photons and other massless particles scales as $(1+z)^4$ while the energy density of nonrelativistic massive particles scales as $(1+z)^3$. Consequently, there is an epoch of matter-radiation equality before which the energy density of the universe is dominated by radiation (massless particles), and after which the energy density of the universe is dominated by matter (massive particles). The redshift at this epoch is

$$1 + z_{\rm eq} = \frac{\rho_{m_0}}{\rho_{r_0}} = 3 \times 10^4\,\Omega\,h^2 \tag{30}$$

where the matter density is $\rho_{\rm m} = \rho_{m_0}(1+z)^3$ and relativistic density is $\rho_{\rm r} = \rho_{r_0}(1+z)^4$. The horizon scale at this epoch is $L_{\rm eq} \equiv 2\,ct_{\rm eq} = 12\,h^{-1}$ Mpc, in comoving coordinates.

During the radiation-dominated era, perturbations in the primordial power spectrum $(\delta\rho/\rho)$ grow in proportion to the square of the universal scale-factor a^2. The growth of perturbations stalls once they get inside the horizon (Hubble radius) since the time scale for the growth of perturbations $(\propto (\rho_{\rm mat}G)^{-1/2})$ is large compared to the expansion rate of the universe $(\propto (\rho_{\rm tot}G)^{-1/2})$. However once the universe becomes matter-dominated, perturbations resume growth, but now in proportion to a, the universal scale-factor, instead of the square of this scale.

The stalling of growth on small scales within the Hubble radius imprints itself on the power spectrum which survives the transition to a matter-dominated universe. Subhorizon density fluctuations have a distribution with scale $\delta\rho/\rho \propto L^{-2}_{\rm comoving} \propto M^{-2/3}$, on scales $\lesssim L_{\rm eq}$, while on larger scales growth suppression implies that $\delta\rho/\rho \approx$ constant. Mathematically, we include this effect by multiplying the primordial power spectrum $P_{\rm p}$ for linearized fluctuations in a stochastic density field by the transfer function

$$P_{\rm f}(k) = T(k)\,P_{\rm p}(k) \tag{31}$$

to obtain the final processed power spectrum $P_{\rm f}$. An approximate analytical fit of a transfer function calculated numerically is given by

$$T(k) = \frac{\log(1+2.34q)}{(2.34q)} \times \left[1 + 3.89q + (16.1q)^2 + (5.46q)^3 + (6.71q)^4\right]^{-1/4} \tag{32}$$

where $q = (k/{\rm Mpc}/(\Omega h^2))$ (Peacock 1997, Bardeen *et al* 1986) and $h = H_0/(100\,{\rm km\,s^{-1}\,Mpc^{-1}})$. Here we note that the above power spectrum results in a bottom-up sequence of evolution, smaller scale fluctuations reaching large amplitude before the large-scale fluctuations.

Dissipationless non-radiative matter, known otherwise as DARK MATTER, is now a standard part of the paradigm for galaxy formation. It is important in reconciling the power spectrum observed in the COSMIC MICROWAVE BACKGROUND with the local power spectrum observed in galaxies. Nonbaryonic dark matter makes this possible because the growth of fluctuations is able to commence immediately after the era of matter-radiation equality. The growth of fluctuations in the baryonic component is, by contrast, suppressed because of its coupling to radiation background. Eventually, as the universe expands, it becomes cool enough so that radiation decouples from the baryons, an event known as last scattering. At this point, the mass scale above which fluctuations can grow suddenly drops from $\sim 10^{16} M_\odot$ to $\sim 10^6 M_\odot$, and so the fluctuations in the baryonic component are free to grow with the rest of the universe. Since the dark matter component is the densest portion of the universe, it controls the growth of fluctuations, and soon after decoupling, the baryonic component is boosted in amplitude by $\sim (1+z_{\rm eq})/(1+z_{LS})$, z_{LS} being the redshift at last scattering.

Remarkably, the acoustic modes set up in the baryonic component prior to last scattering are observable via their imprint on the cosmic microwave background radiation which provides a snapshot of fluctuations at last scattering. This may be seen as follows. The dispersion relation for fluctuation growth at rate $e^{i\omega t}$ is

$$\omega^2 = k^2 V_{\rm S}^2 - 4\pi G\rho \tag{33}$$

($V_{\rm S}$ being the sound speed) for time scales short compared to the expansion time, or equivalently for wavelengths $2\pi/h \ll ct$. Such wavelengths oscillate as sound waves with amplitude proportional to $\exp(ik\,V_{\rm S}\,t)$. Inflation specifies primordial curvature fluctuations which are time-invariant on superhorizon scales. The corresponding amplitude corresponds to the mode $\cos(k\,V_{\rm S}\,t)$ which is finite as $k \to 0$. At $t_{\rm LS}$ the maximum amplitude in $\delta T/T$ is attained by fluctuations which have entered the horizon and satisfy $kV_{\rm S}t = \pi n$, with $n = 1, 2, \ldots$. Clearly the wave that crests on the horizon at $t_{\rm LS}$ and undergoes oscillations systematically experiences both Compton drag and diffusive damping. The net result is the amplitude is diminished and only the first three or so peaks are detectable. Doppler effects contribute 90 degrees out of phase to the gravitational potential induced $\delta T/T$ peaks, and the net result is a series of peaks corresponding to the maximum in $\delta T/T = |\frac{1}{3}\delta\phi - \underline{r}\cdot\underline{v} + \frac{1}{3}\delta\rho/\rho|$. The predicted enhancement of the first peak of wavelength $2\,V_{\rm S}\,t_{\rm LS}$ is as large as a factor of three relative to the $k \to 0$ or Sachs–Wolfe potential fluctuation limit that corresponds to the inflationary imprint of primordial, nearly scale-invariant curvature fluctuations.

The observational situation is as follows. The COBE DMR experiment measured potential fluctuations on angular scales $\geq 10^\circ$. The last scattering horizon is about 1°. Several experiments report evidence for the first acoustic peak on this scale, and definitive

measurements with large sky coverage that provide adequate foreground discrimination were performed in 2000 by the Boomerang and Maxima balloon experiment collaborations (de Bernardis *et al* 2000, Hanany *et al* 2000). The situation is complicated by the contribution from the time-varying potential perturbations $\delta T/T \sim \int \dot{\phi}\, \mathrm{d}t$ which arise in models with $\Omega < 1$ at curvature-dominated epochs, i.e. $z \lesssim \Omega^{-1} - 1$ and are expected at angular scales $\gtrsim 1^\circ$. A final complication is the possible tensor mode contributions; this gravity wave background contributes to the energy density in the radiation-dominated era and only affects the low order multipoles corresponding to angular scales that are larger than the horizon angular scale at t_{eq}.

Linear evolution of density fluctuations

Gravitation magnifies the initial perturbations in the matter distribution, both dark and baryonic. While these perturbations initially grow linearly, eventually these perturbations stop expanding, break away from Hubble flow, and then collapse and virialize. Galaxies form from the gas which cools onto the center of these collapsed masses, called dark halos.

An analysis of the growth of the initial matter power spectrum begins with the basic fluid equations in terms of the density ρ, the velocity $\boldsymbol{u}$, the gravitational potential Φ, the pressure p, and the time t:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0, \quad \rho \frac{D\boldsymbol{u}}{Dt} = -\nabla p - \rho \nabla \Phi. \tag{34}$$

For convenience, we change variables to comoving coordinates $\boldsymbol{x} = \boldsymbol{r}/a$, to peculiar velocities $\boldsymbol{v} = a\mathrm{d}\boldsymbol{u}/\mathrm{d}t$, to dimensionless overdensities $\delta = \rho/\bar{\rho} - 1$, and to conformal times $\tau = t/a$. Making these substitutions, the fluid equations become

$$\boldsymbol{v} = \dot{\boldsymbol{x}} \tag{35}$$

$$\dot{\delta} + \nabla \cdot [(1+\delta)\boldsymbol{v}] = 0 \tag{36}$$

$$\dot{\boldsymbol{v}} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} + \frac{\dot{a}}{a}\boldsymbol{v} = -\frac{\nabla p}{\rho} - \nabla \Phi \tag{37}$$

while Poisson's equation reads

$$\nabla^2 \Phi = 4\pi G \bar{\rho} a^2 \delta. \tag{38}$$

The behavior of the modes can be obtained by taking the divergence of Euler's equation, eliminating $\nabla \cdot \boldsymbol{v} = 0$ by the continuity equation, taking Φ, δ, and $\boldsymbol{v}$ to be small, and then linearizing the equations. One obtains

$$\ddot{\delta} + \frac{\dot{a}}{a}\dot{\delta} = \frac{\nabla^2 p}{\rho} + 4\pi G \bar{\rho} a^2 \delta. \tag{39}$$

If we consider a pressure-free universe, this equation involves no spatial derivatives, so its solution can then be written as

$$\delta(\boldsymbol{x}, \tau) = A(\boldsymbol{x}) f_1(\tau) + B(\boldsymbol{x}) f_2(\tau). \tag{40}$$

Using the fact that $\Omega = 8\pi G\rho/3H^2 = 8\pi G\rho a^2/3\dot{a}^2$, we can rewrite equation (39) as

$$\ddot{\delta} + \frac{\dot{a}}{a}\dot{\delta} - \frac{3}{2}\Omega \left(\frac{\dot{a}}{a}\right)^2 \delta = 0. \tag{41}$$

For an Einstein–de Sitter universe ($\Omega = 1$), $a \propto \tau^2 \propto t^{2/3}$ and equation (41) becomes

$$\ddot{\delta} + 2\frac{\dot{\delta}}{\tau} - 6\frac{\delta}{\tau^2} = 0. \tag{42}$$

For solutions, one obtains a growing mode $\delta \propto D(\tau) \propto \tau^2 \propto t^{2/3}$ and a decaying mode $\delta \propto D(\tau) \propto \tau^{-1} \propto t^{-1/3}$. $D(\tau)$ is known as the linear growth factor. Given the small size of the perturbations at the era of matter–radiation equality, one ignores the decaying mode solution and simply considers the growing mode. For the more general case of a universe without a cosmological constant (i.e., $\Omega_\Lambda = 0$), it can be shown (Peebles 1980) that

$$D(\tau) = 1 + \frac{3}{x} + \frac{3(1+x)^{1/2}}{x^{3/2}} \ln\left[(1+x)^{1/2} - x^{1/2}\right] \tag{43}$$

where $x = \Omega^{-1} - 1 \propto a$.

Before finishing this discussion, we examine the evolution in position and velocity of a test particle because these quantities will prove useful later when we look at how galaxies acquire ANGULAR MOMENTUM. Since $\nabla^2 \Phi_0 = 4\pi G \bar{\rho} a^2 \delta$ is a constant in the linear regime for an Einstein–de Sitter universe, the linearized version of Euler's equation can be immediately integrated to give

$$\boldsymbol{v} = -\left(a^{-1} \int D\mathrm{d}\tau\right) \nabla \Phi_0. \tag{44}$$

Integrating the peculiar velocity gives us:

$$\boldsymbol{x} = \boldsymbol{x}_0 - \left(\int \frac{\mathrm{d}\tau}{a} \int D\mathrm{d}\tau\right) \nabla \Phi_0. \tag{45}$$

Since the growth factor $D(\tau)$ is a solution to the fluctuation growth equation $\frac{\mathrm{d}}{\mathrm{d}\tau}(a\dot{\delta}) = 4\pi G \bar{\rho} a^3 \delta$ (see equation (39)), it follows that

$$\boldsymbol{x} = \boldsymbol{x}_0 - \frac{D(\tau)}{4\pi G \bar{\rho} a^3} \nabla \Phi_0 \tag{46}$$

and

$$\boldsymbol{v} = -\frac{\dot{D}}{4\pi G \bar{\rho} a^2} \nabla \Phi_0. \tag{47}$$

This essentially Lagrangian view to the growth of perturbations is due to Zeldovich (1970).

Mean square fluctuations

It is useful to consider the evolution of the mean square mass fluctuations filtered on various spatial scales R because of the information it provides on the collapse of structure on these spatial scales. Due to the scale-free growth of initial perturbations, the power spectrum $P(k, \tau)$

$$P(k, \tau) = \left| \int \mathrm{d}^3 x \delta(\boldsymbol{x}, \tau) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \right|^2 \tag{48}$$

grows as $D(\tau)^2$ just as $\delta(\boldsymbol{x}, \tau)$ does.

Assuming that a section of the power spectrum can be written as

$$P(k, \tau) \propto D(\tau)^2 k^n \tag{49}$$

where n is the power spectrum index, we find that the mean square fluctuations filtered on a particular mass scale are

$$\begin{aligned}\sigma^2(M) &= \left\langle (dM/M)^2 \right\rangle \\ &= \int \mathrm{d}^3 k W(kR) P(k, \tau) \\ &\propto D(\tau)^2 R^{-3-n} \\ &\propto D(\tau)^2 M^{-(n+3)/3}\end{aligned} \tag{50}$$

where $W(kR)$ is a window function that filters out spatial scales of R or smaller. The top hat filter of radius R

$$W(k) = \frac{3}{kR^3}(\sin kR - kR\cos kR). \tag{51}$$

is one commonly used window function. Note that $\sigma(M, 0) = (M/M_{nl})^{-(n+3)/6}$ where $M_{nl} = 1 \times 10^{15}\sigma_8 \Omega h^{-1} M_\odot$ corresponds to mass fluctuations of amplitude unity on $8h^{-1}$ Mpc radius spheres. The factor σ_8 is $\sigma(M, 0)$ at $8h^{-1}$ Mpc for the mass density: galaxy count fluctuations have unit amplitude on this scale.

Halo scaling relations

The growth of fluctuations becomes nonlinear and start collapsing into virialized objects when

$$\left\langle (\delta M/M)^2 \right\rangle = 1. \tag{52}$$

From equation (50), this implies that

$$M \propto D(t)^{6/(3+n)}. \tag{53}$$

The density ρ, size r, and temperature T of collapsed objects then scale as

$$\rho \propto (1+z)^3 \tag{54}$$

$$r \propto (M/\rho)^{1/3} \propto (1+z) D(t)^{6/(3+n)} \tag{55}$$

$$\begin{aligned}T \propto V_\mathrm{c}^2 &\propto \frac{GM}{r} \propto M^{2/3}\rho^{1/3} \\ &\propto (1+z)^1 D(t)^{4/(3+n)}.\end{aligned} \tag{56}$$

These scaling relationships will be useful in later discussions, particularly when we examine the cooling of gas in virialized structures.

Spherical collapse model

To understand the nonlinear growth of structure, it is convenient to consider the idealized spherical 'top hat' collapse model. Here, one supposes there to be spherically symmetric regions of radius R and of uniform overdensity $\bar{\delta}$ in an otherwise uniform universe at some initial time t_i. At these early times, the universe will be approximately Einstein–de Sitter, so we can express this overdensity as the sum of the growing and decaying modes:

$$\delta = \delta_+ \left(\frac{t}{t_\mathrm{i}}\right)^{2/3} + \delta_- \left(\frac{t}{t_\mathrm{i}}\right)^{-1}. \tag{57}$$

We take the matter in this region to be expanding at the approximately the same rate as the universe, and therefore we require that the peculiar velocity be zero:

$$\tfrac{2}{3}\delta_+(t_\mathrm{i}) - \delta_-(t_\mathrm{i}) = 0. \tag{58}$$

Therefore, $\delta_+ = \frac{3}{5}\delta$. According to Birkhoff's theorem, in a spherically symmetric situation, matter external to the sphere will not influence its evolution, so it follows that

$$\frac{\mathrm{d}^2 R}{\mathrm{d}t^2} = \frac{-GM}{R^2} = \frac{-4\pi G}{3}\bar{\rho}(1+\bar{\delta})R \tag{59}$$

which is identical in form to the equation for the evolution of the cosmological scale factor a,

$$\frac{\mathrm{d}^2 a}{\mathrm{d}t^2} = \frac{-GM}{a^2} = \frac{-4\pi G}{3}\bar{\rho}a. \tag{60}$$

Therefore, the size of the region R evolves like the cosmic scale factor a but with an initial density parameter $\Omega_\mathrm{p}(t_\mathrm{i})$ given by

$$\Omega_\mathrm{p}(t_\mathrm{i}) = \frac{\rho(t_\mathrm{i})(1+\bar{\delta})}{\rho_\mathrm{c}(t_\mathrm{i})} \tag{61}$$

where $\rho_\mathrm{c}(t_\mathrm{i})$ is the critical density ($3H_\mathrm{i}^2/8\pi G$ at time t_i). By analogy with the solutions for the universe, the region will collapse if $\Omega_\mathrm{p} > 1$. By analogy with equation (4), the expansion of the region evolves according to

$$\left(\frac{\dot{a}}{a}\right)^2 = H_\mathrm{i}^2 \left[\Omega_\mathrm{p}(t_\mathrm{i})\frac{a_\mathrm{i}^3}{a^3} + (1 - \Omega_\mathrm{p}(t_\mathrm{i}))\frac{a_\mathrm{i}^2}{a^2}\right]. \tag{62}$$

Eventually, the region stops expanding, turns around, and collapses. At the turn-around time t_m, $\dot{a} = 0$, implying that $a/a_\mathrm{i} = \Omega_\mathrm{p}(t_\mathrm{i})/(1 - \Omega_\mathrm{p}(t_\mathrm{i}))$. The density at turn-around t_m is then

$$\rho_\mathrm{p}(t_\mathrm{m}) = \rho_\mathrm{c}(t_\mathrm{i})\Omega_\mathrm{p}(t_\mathrm{i})\left[\frac{\Omega_\mathrm{p}(t_\mathrm{i}) - 1}{\Omega_\mathrm{p}(t_\mathrm{i})}\right]^3. \tag{63}$$

By analogy with the solutions to the cosmological equations for closed universes (see equations (9)–(11)), we can determine t_m as

$$\begin{aligned}t_\mathrm{m} &= \frac{\pi}{2H_\mathrm{i}}\frac{\Omega_\mathrm{p}(t_\mathrm{i})}{[\Omega_\mathrm{p}(t_\mathrm{i}) - 1]^{3/2}} = \frac{\pi}{2H_\mathrm{i}}\left[\frac{\rho_\mathrm{c}(t_\mathrm{i})}{\rho_\mathrm{p}(t_\mathrm{m})}\right]^{1/2} \\ &= \left[\frac{3\pi}{32 G \rho_\mathrm{p}(t_\mathrm{m})}\right]^{1/2}.\end{aligned} \tag{64}$$

For an Einstein–de Sitter universe, the density $\rho(t_m)$ of the universe at turn-around is simply

$$\rho(t_m) = \frac{1}{6\pi G t_m^2} \tag{65}$$

so that

$$\chi = \frac{\rho_p(t_m)}{\rho(t_m)} = \left(\frac{3\pi}{4}\right)^2 \approx 5.6. \tag{66}$$

Formally, after turn-around, the region will collapse to a point at $t = 2t_m$. Of course, before that happens, shell crossings will occur. From the virial theorem, dissipationless matter collapses to a radius that is one-half the turn-around radius. Since $a \propto t^{2/3}$, the universe will expand by $2^{2/3}$. Hence, the density of the region relative to the background universe at collapse t_c is

$$\frac{\rho_p(t_c)}{\rho(t_c)} = (2^{2/3})^3 8\chi \approx 180. \tag{67}$$

An extrapolation of the linear growth estimate at time t_c yields

$$\delta_+(t_c) = \frac{3}{5}\delta_i \left(\frac{2t_m}{t_i}\right)^{2/3} = \frac{3}{5}\left(\frac{3\pi}{2}\right)^{2/3} \approx 1.68. \tag{68}$$

The assumption of a uniform spherical overdensity in an otherwise uniform universe is quite unrealistic; in fact, collapse typically proceeds toward the creation of a large number of two-dimensional pancakes. Nevertheless, numerical simulations show the basic scalings derived here to be roughly correct and useful for making simple analytic estimates.

Form of collapsed structures

For many years, the profiles of collapsed halos were taken to be that of a isothermal sphere:

$$\rho(r) = \frac{V_c^2}{4\pi G r^2} \tag{69}$$

where r is the radius and V_c is the circular velocity. Over the last few years, however, the detailed N-body simulations of Navarro *et al* (1997) have shown that the collapsed matter profile is better fitted by the double power-law profile:

$$\rho(r) = \rho_{crit} \frac{\delta_0}{(r/r_s)(1 + r/r_s)^2} \tag{70}$$

where r_s is the core radius where the slope of the power-law changes from -1 near the center to -3 at large radii and r_{200} is the virial radius of the halo. More recent high resolution simulations find a somewhat steeper innermost slope to the density profile, possibly as steep as -1.5 (Jing and Suto 2000, Moore *et al* 1999).

Press–Schechter model

A convenient and somewhat approximate description of the mass function of halos at particular redshifts is given by the Press–Schechter formalism derived heuristically by Press and Schechter (1974) using linear growth theory and the spherical 'top hat' model. The mass function of non-linear objects is computed with the aid of linear theory on the assumption that the probability distribution of density fluctuations at a given mass and redshift is Gaussian, centered on the mean $\sigma(M, z) \equiv \left\langle (\delta\rho/\rho)^2 \right\rangle^{1/2}$.

Since the set of initial perturbations is Gaussian-distributed and remains so under linear growth, we can write the distribution of mass fluctuations δ_M as

$$P(\delta_M)\mathrm{d}\delta_M = \frac{1}{\sqrt{2\pi}\sigma(M)} \exp\left(\frac{-\delta_M^2}{\sigma(M)^2}\right) \mathrm{d}\delta_M \tag{71}$$

where $\sigma(M)$ is equal to the mean square mass fluctuations $\langle (\delta M/M)^2 \rangle$ defined earlier.

The fraction of points where the mean density inside a radius R exceeds δ_c is

$$P_{>\delta_c}(M) = \int_{\delta_c}^{\infty} P(\delta_M)\mathrm{d}\delta_M. \tag{72}$$

Press and Schechter (1974), using the spherical collapse model as a heuristic guide, took all regions where the mean density exceeded the critical value needed for linear collapse, i.e., $\delta_c \approx 1.68$, to be part of collapsed structures at least as massive as $M = 4\pi\bar{\rho}a^3R^3/3$.

Then, to distinguish the fraction of structures which have just collapsed to a mass M and those that are part of a bigger structure at least as massive as $M + \mathrm{d}M$, we subtract $P_{>\delta_c}(M + \mathrm{d}M)$ from $P_{>\delta_c}(M)$. Notice that by doing this, we completely ignore the possibility that the mass scale just collapsing is contained within a larger mass that is just collapsing, a complication known as the 'cloud-in-cloud' problem. Another problem is that only half of the points are associated with an overdensity and therefore become part of any collapsed structure. Press and Schechter (1974) elected to solve this problem somewhat arbitrarily by multiplying the number of structures at a given mass scale by a factor of 2 with the vague understanding that this represents flow from underdense to overdense regions.

The number of halos $n(M)$ with masses between M and $M + \mathrm{d}M$ is

$$n(M)M\mathrm{d}M = 2\rho_M[P_{>\delta_c}(M) - P_{>\delta_c}(M + \mathrm{d}M)]\mathrm{d}M, \tag{73}$$

ρ_m being the average density at the redshift in question and the mass term M on the left-hand-side accounting for the fact that more massive halos are associated with more points of reference. Rewriting this, we get

$$\begin{aligned} n(M)\mathrm{d}M &= \frac{2\rho_M}{M}\left|\frac{\mathrm{d}P_{>\delta_c}(M)}{\mathrm{d}M}\right|\mathrm{d}M \\ &= \frac{2\rho_M}{M}\left|\frac{\mathrm{d}P_{>\delta_c}(M)}{\mathrm{d}\sigma(M)}\right|\left|\frac{\mathrm{d}\sigma(M)}{\mathrm{d}M}\right|\mathrm{d}M. \end{aligned} \tag{74}$$

Finally, since the mean square fluctuations $\sigma^2(M)$ evolve as $D(\tau)^2$ according to linear growth (see equation (50)), we can rewrite the expression as

$$n(M)\mathrm{d}M = \frac{2\rho_\mathrm{M}}{M^2}\frac{\delta_\mathrm{M}}{\sigma(M)}\frac{1}{\sqrt{2\pi}\sigma(M)} \times \exp\left(\frac{-\delta_\mathrm{c}^2}{2\sigma(M)^2}\right)\left|\frac{\mathrm{d}\sigma(M)}{\mathrm{d}M}\right|\mathrm{d}\ln M. \quad (75)$$

This is the Press–Schechter mass function at redshift z. Despite the extremely heuristic nature of this derivation and the many problems already discussed, it agrees remarkably well with numerical simulations, and thus has proven very useful in characterizing the growth of structure.

Cooling processes

Galaxies are much less massive than the mass scales ($\sim 10^{14} M_\odot$) going nonlinear in the universe today, so clearly galaxies must be more than simply virialized structures. The key seems to be the process of cooling and the time scale for the settling of baryons into the centers of their dark matter halos. There are several cases to consider. Clearly, if the cooling time of a gas is larger than the Hubble time, the gas cannot have evolved much over the history of the universe. If the cooling time is smaller than the Hubble time but larger than the dynamical time, the gas will suffer slow quasistatic collapse into the center of the virialized halo. On the other hand, if the cooling time is smaller than the dynamical time, the ambient gas will undergo runaway cooling and collapse to the center of the virialized halo. It is this case, where the cooling time is much shorter than the dynamical time scales for ACCRETION or merging (Binney 1977, Rees and Ostriker 1977, Silk 1977) that is relevant for the formation of galaxies.

There are four important processes by which gas in halos cools: Compton cooling, free–free emission (bremsstrahlung), recombination, and collision-induced de-excitation.

We begin with a consideration of the Compton cooling process. When low-energy photons pass through a gas of non-relativistic electrons, they scatter off the electrons with the Thompson cross-section σ_T:

$$\sigma_\mathrm{T} = \frac{8\pi}{3}\left(\frac{e^2}{m_\mathrm{e}c^2}\right)^2 \quad (76)$$

where m_e is the mass of the electron and e is the charge of the electron. Some photons are scattered up in energy and some are scattered down, but the net effect is to slow the electrons relative to the frame of the cosmic microwave background radiation. The mean shift in photon energy per collision is

$$h\overline{\Delta\nu} = \frac{4kT_\mathrm{e}}{m_\mathrm{e}c^2}h\nu \quad (77)$$

where k is Boltzmann's constant, h is Planck's constant, ν is the frequency of a photon, and T_e is the temperature of the electron gas. In a sea of photons with temperature T_γ, the mean rate of energy loss per electron is

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{4kT_\mathrm{e}}{m_\mathrm{e}c^2}\sigma_\mathrm{T}aT_\gamma^4 \quad (78)$$

where a is the Stefan–Boltzmann constant. The cooling time t_cool is

$$t_\mathrm{cool} = \frac{\frac{3}{2}nkT_\mathrm{e}}{n_\mathrm{e}\frac{4kT_\mathrm{e}}{m_\mathrm{e}c^2}\sigma_\mathrm{T}aT_\gamma^4} = \frac{3m_\mathrm{e}c^2}{8\sigma_\mathrm{T}aT_\gamma^4} \sim 2.1\times10^{12}(1+z)^{-4}\ \mathrm{yr}. \quad (79)$$

High temperature (10^6–10^7 K) primordial gases are almost entirely ionized. Under these circumstances, the dominant cooling mechanism is due to the acceleration of electrons off the bare H^+ and He^{2+} nuclei. This results in a cooling rate per unit rate per unit volume:

$$\frac{\mathrm{d}E}{\mathrm{d}t} \propto n_\mathrm{e}n_\mathrm{H}T^{1/2}. \quad (80)$$

The cooling time t_cool here is approximately equal to

$$t_\mathrm{cool} = \frac{\frac{3}{2}nkT}{\frac{\mathrm{d}E}{\mathrm{d}t}} = 6.6\times10^9\frac{T_6^{1/2}}{n_{-3}}\ \mathrm{yr} \quad (81)$$

where $T_6 = T/10^6\,\mathrm{K}$ and $n_{-3} = n/10^{-3}\mathrm{cm}^{-3}$.

On the other hand, low temperature (10^4–10^5 K) primordial gases are only partially ionized. Here cooling is dominated by two processes: one where electrons recombine with ions resulting in the release of a photon (recombination) and one where electrons collide with partially ionized atoms, thereby exciting them to a state which they escape by the release of a photon. The total cooling rate can be expressed as

$$\frac{\mathrm{d}E}{\mathrm{d}t} \propto n_\mathrm{e}n_\mathrm{H}f(T). \quad (82)$$

The latter process is the dominant one, and for primordial abundances, the function $f(T)$ can be approximated as $2.5(T/10^6\,\mathrm{K})^{-1/2}\,\mathrm{erg\,cm^3\,s^{-1}}$. The cooling time t_cool is then

$$t_\mathrm{cool} = \frac{\frac{3}{2}nkT}{\frac{\mathrm{d}E}{\mathrm{d}t}} = 3.0\times10^9\frac{T_6^{3/2}}{n_{-3}}\ \mathrm{yr}. \quad (83)$$

We compare these cooling time scales with the dynamical time scales $t_\mathrm{dyn} \sim \sqrt{1/G\rho}$. We consider a uniform spherical cloud with mass M in virial equilibrium with f fraction of its mass in dissipative baryonic matter and the rest in dark, dissipationless matter, so the gas mass M_g is equal to fM. For this mass configuration,

$$t_\mathrm{dyn} \sim \sqrt{\frac{1}{G\rho}} \sim \sqrt{\frac{1}{Gn/f}} \sim 6.5\times10^9 f^{1/2}n_{-3}^{-1/2}\ \mathrm{yr}. \quad (84)$$

Now, we compare this dynamical time scale with the cooling times derived for each of the cooling processes

discussed above. At early times, $z > 10$, the Compton cooling process dominates, and since

$$n_{-3} = 2.3 \times 10^{-2} f(1+\delta)(\Omega_0 h^2)(1+z)^3 \tag{85}$$

the dynamical time scale goes as

$$t_{\rm dyn} \sim 3.0 \times 10^9 (\Omega_0 h^2)^{-1/2}(1+z)^{-3/2}\ {\rm yr} \tag{86}$$

and the relative time scale τ goes as

$$\tau = \frac{t_{\rm cool}}{t_{\rm dyn}} \approx 6 \times 10^2 (1+z)^{-5/2}. \tag{87}$$

Thus, Compton cooling will be important at $z > 7$ where $\tau < 1$. Notice that the relative time scale does not depend upon mass, temperature, or density. Therefore, if galaxies formed at early times, they would have no preferred scales.

The relevant temperatures for lower mass halos ($\lesssim 10^{12} M_\odot$) are less than 10^6 K. In this case, line cooling dominates and the relative time scale τ goes as $t_{\rm cool}/t_{\rm dyn} \propto (T^{3/2}/\rho^{1/2}) \propto ((M^{2/3}\rho^{1/3})^{3/2}/\rho^{1/2}) \propto M$. Therefore, the $\tau = 1$ line runs parallel to lines of constant mass. To determine the mass limit more precisely, we look at τ:

$$\tau = \frac{t_{\rm cool}}{t_{\rm dyn}} = 0.4 T_6^{3/2} f^{-1/2} n_{-3}^{-1/2}. \tag{88}$$

We can relate this to the mass of a spherical cloud model in virial equilibrium by using the following relation from the virial theorem:

$$\frac{3kT}{2\mu} = \frac{0.3GM}{R}, \tag{89}$$

where μ, the mean molecular weight, is roughly equal to half the proton mass $m_{\rm p}$ since the medium is ionized. From this, it follows that $T^3 \propto \rho M^2 \propto n f^{-1} M^2$ and then that

$$M_{\rm g} = 1.2 \times 10^{13} T_6^{3/2} f^{3/2} n_{-3}^{-1/2} M_\odot. \tag{90}$$

Hence,

$$\tau = \frac{M_{\rm g}}{1.2 \times 10^{13} f^2 M_\odot} = \frac{M}{1.2 \times 10^{13} f M_\odot}. \tag{91}$$

This sets the mass limit below which gas can effectively cool to form structures. For $f \sim 1$, the mass limit is much larger than the typical limiting galaxy mass ($\sim 10^{12} M_\odot$), but for smaller values consistent with constraints set by big-bang NUCLEOSYNTHESIS ($f \sim 0.05 - 0.1$), the mass limit is comparable to these limits.

On the other hand, for higher mass halos ($\gtrsim 10^{12} M_\odot$), the relevant temperatures are greater than $10^6 K$. Here the dominant cooling mechanism is bremsstrahlung, and the relative time scale ratios τ go as $t_{\rm cool}/t_{\rm dyn} \propto (T^{1/2}/\rho^{1/2}) \propto ((M^{2/3}\rho^{1/3})^{1/2}/\rho^{1/2}) \propto R$, so the $\tau = 1$ line runs parallel to lines of constant radius. To determine the limiting radius more precisely, we look at τ

$$\tau = \frac{t_{\rm cool}}{t_{\rm dyn}} = T_6^{1/2} f^{-1/2} n_{-3}^{-1/2}. \tag{92}$$

Using the spherical cloud model again, we solve for R in terms of the other variables,

$$R = \sqrt{\frac{3fkT}{0.8\pi\mu^2 G n_{-3}}} = 610 f T_6^{1/2} n_{-3}^{-1/2}\ {\rm kpc} \tag{93}$$

and so

$$\tau = \frac{R}{610 f^{3/2}\ {\rm kpc}}. \tag{94}$$

Hence, for $f \sim 0.1$, massive gas clouds of radii greater than 20 kpc can efficiently cool. Since this length is smaller than the typical cluster size, cooling is not very efficient in clusters, and therefore the gas simply suffers slow quasistatic collapse.

The galaxy cluster mass function

In the previous section, we discussed two important different regimes for virialized masses, one in which the cooling time was longer than the dynamical time and one in which it was shorter than it. In the former regime, one obtains GALAXY CLUSTERS where most of the gas remains hot and in the latter regime one obtains galaxies where much of the halo gas has apparently cooled. In either case, one can use Press–Schechter theory to calculate the mass function, and with simple assumptions about the conversion of gas into stars or other luminous objects, one can convert this into a LUMINOSITY FUNCTION OF GALAXIES.

Perhaps the most direct comparison with observations is via the mass function of galaxy clusters. The shape, expected to be exponential plus a power law tail, fits the prediction remarkably well, to the extent that cluster masses are well determined. Three techniques are used to estimate cluster masses: galaxy velocity dispersion and distribution, hot gas temperature and distribution, and gravitational lensing maps. The first two methods assume virial equilibrium. All three methods give consistent results, to within a factor of 2 in mass. One can compare the characteristic cluster mass, determined by the fitting function

$$\frac{{\rm d}N}{{\rm d}M} \propto M^{-\alpha} \exp\left(-(M/M_{\rm nl})^\beta\right), \tag{95}$$

with the predicted value of $M_{\rm nl}$ taken from field galaxy counts and a bias factor that has to be empirically deduced. Indeed, $M_{\rm nl}$ corresponds to a typical observed cluster mass. The normalization of the cluster mass function depends both on the mean density and $\sigma(M, Z)$, with an exponential sensitivity to $\sigma(M, Z)$. Only five percent of galaxies are in clusters, which can therefore account for perhaps one percent of the critical density. Clusters are therefore rare objects, typically 3σ fluctuations. The number density of clusters is controlled by both the mean density and σ_8, in the combination $\sigma_8\,\Omega^{0.6} \approx 0.7 \pm 0.2$. The scale $8h^{-1}$ Mpc, corresponding to unit amplitude of the optical counts and the mass M_8 of a typical cluster, is used for normalization, and $\sigma(M, 0) = \sigma_8 (M_8/M)^{(n+3)/6}$ where $M_8 = 4\pi(8h^{-1}{\rm Mpc})^3 \Omega\bar{\rho}$.

The galaxy luminosity function

The Press–Schechter formulation can also be used to derive the galaxy luminosity function, which is described by the Schechter function

$$\frac{\mathrm{d}N}{\mathrm{d}L} \propto L^{-\alpha} \exp(-L/L_*) \tag{96}$$

where $L_* \approx 10^{10} L_\odot$ and $\alpha \approx 1$ to 1.5 depending on the galaxy selection criterion. There are two ingredients that must be incorporated in connecting mass to luminosity in order to obtain a satisfactory comparison of theory and observation. A characteristic luminosity L_* must be derived, and the formation of low mass objects into luminous objects must be inefficient, since the mass function tail has slope M^{-2} whereas the luminosity function is flatter. The characteristic luminosity follows from the fact that cooling is efficient only for masses up to $\sim 10^{12} M_\odot$. Then, assuming a typical baryon fraction, an age of $\sim 10^{11}$ yr, and a standard mass function for stars, one derives a characteristic luminosity of $\sim 10^{10} L_\odot$.

Angular momentum

As structures grow and collapse in the early universe, they exert tidal torques on each other, and this provides each collapsing mass with some angular momentum. This angular momentum, in turn, is important in determining the final properties of the disk and elliptical galaxies which form inside these collapsed structures.

The angular momentum of a collapsing halo can be expressed as

$$\boldsymbol{J} = \int_{\mathrm{V}} \mathrm{d}^3 x \bar{\rho} a^3 (a\boldsymbol{x} - a\bar{\boldsymbol{x}}) \times \boldsymbol{v} \tag{97}$$

where $\bar{\boldsymbol{x}}$ is the center of mass for the volume. Using equation (47), we express $\boldsymbol{v}$ as $-a\dot{b}\nabla\Phi_0$ where $b(\tau) = D/4\pi G\bar{\rho}a^3$. For convenience we expand $\nabla\Phi_0$ in a Taylor series around the point $\boldsymbol{x}$:

$$\begin{aligned} \nabla\Phi_0|_{\boldsymbol{x}} &= \nabla\Phi_0|_{\bar{\boldsymbol{x}}} + (\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot \frac{\partial^2 \Phi_0}{\partial \boldsymbol{x}^2}|_{\bar{\boldsymbol{x}}} \\ &= \nabla\Phi_0|_{\bar{\boldsymbol{x}}} + (\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot \underline{\underline{T}} \end{aligned} \tag{98}$$

where $T_{jl} = \partial^2\Phi_0/\partial x_{\mathrm{j}} \partial x_{\mathrm{l}}$. Rewriting this, we get

$$J_{\mathrm{i}}(\tau) = -a\dot{b}\epsilon_{ijk} T_{jl} \int_V (x_{\mathrm{l}} - \bar{x}_{\mathrm{l}})(x_{\mathrm{k}} - \bar{x}_{\mathrm{k}}) \bar{\rho} a^3 \mathrm{d}^3 x \tag{99}$$

or

$$J_{\mathrm{i}}(\tau) = -a\dot{b}\epsilon_{ijk} T_{jl} I_{lk} \tag{100}$$

where I_{lk} is the inertial tensor.

We now estimate how J_{i} scales. Since I_{lk} scales as a^2 until collapse while T_{jl} continues to scale as $\nabla\Phi/a^2 \sim (D/a)/a^2 \sim 1/a^2$, each structure effectively acquires angular momentum from its neighbors until collapse. Since the collapse of a structure occurs when $\delta \sim 1$, $b \sim D(\tau)/4\pi G\bar{\rho}a^3 \sim 1/4\pi G\bar{\rho}a^2 \sim 1/\nabla^2\Phi$ from Poisson's equation and the relation $D(\tau) \propto a$, so T_{jl} scales as $\nabla^2\Phi_0 \sim 1/b$. $\underline{\underline{I}}$ scales as $MR_0^2 \sim MR^2/a^2$. Hence,

$$\begin{aligned} J_{\mathrm{i}}(\tau) &\sim -a\dot{b} T_{jl} I_{lk} \sim a\dot{b} \frac{1}{b} \frac{MR^2}{a^2} \\ &\sim \frac{\dot{b}}{\dot{a}b} \frac{\dot{a}}{a^2} M(M/\rho)^{2/3} \\ &\sim \Omega^{0.6} H(\Omega H^2)^{-2/3} M^{5/3} \\ &\sim \Omega^{-0.07} t^{1/3} M^{5/3}. \end{aligned} \tag{101}$$

It is standard to construct a dimensionless quantity which characterizes the angular momentum that each collapsed mass has acquired via tidal torques. This quantity is called the dimensionless angular momentum λ, and it can be expressed as

$$\lambda = \frac{|\boldsymbol{J}||E|^{1/2}}{GM^{5/2}}. \tag{102}$$

Noting that $|E| \sim M^2/R \sim M^2(\rho/M)^{1/3} \sim M^{5/3}(\Omega H)^{1/3} \sim \Omega^{1/3} M^{5/3} t^{-2/3}$, we see that $\lambda \sim |\boldsymbol{J}||E|^{1/2} M^{-5/2} \sim \Omega^{-0.07} t^{1/3} M^{5/3} \Omega^{1/6} t^{-1/3} M^{5/6} M^{-5/2} \sim \Omega^{0.1}$. Therefore, the distribution of dimensionless angular momenta λ is essentially independent of a halo's mass, collapse time, or even the basic world model. N-body simulations (Warren *et al* 1992, Cole and Lacey 1996, Catelan and Theuns 1996) and analytical treatments (Steinmetz and Bartelmann 1995) find a distribution which is well fitted by the expression

$$p(\lambda) = \frac{1}{\sqrt{2\pi}\sigma_\lambda} \exp\left[-\frac{\ln(\lambda/\bar{\lambda})^2}{2\sigma_\lambda^2} \right] \frac{\mathrm{d}\lambda}{\lambda} \tag{103}$$

where $\bar{\lambda} = 0.05$ and $\sigma_\lambda = 0.5$.

Disk formation

Disk galaxies make up the dominant component of the local galaxy census. Disk galaxies are known to have exponential profiles ($I(r) \propto \exp(-r/r_{\mathrm{d}})$), to have significant fractions of dust and stars, to still be undergoing some STAR FORMATION, and to be rotationally supported. They are extremely flattened objects and can appear very elongated if viewed edge-on. They also frequently have long bars and spiral structures. It is because of this latter feature that these galaxies are often called SPIRAL GALAXIES.

Most of the global disk properties, e.g., mass, luminosity, size, and metallicity, define a two-dimensional manifold with little scatter about that manifold. It is more well-known in terms of its two-dimensional projections, in particular, the well-known TULLY–FISHER RELATION between luminosity and circular velocity. There are two main views on this tight relationship: one in which these processes as consequences of self-regulating mechanisms for star formation in disks (e.g. Silk 1997) and one in which this is simply the consequence of the cosmological equivalence of mass and circular velocity (e.g. Mo *et al* 1998).

In the past, disk galaxies were thought to have surface brightnesses tightly distributed around 21.65 b_{J} mag/arcsec^2 (Freeman 1970). Shortly after this claim

was made, arguments were made that there was a strong selection bias against low surface brightness galaxies and in reality the spread in surface brightness extended to much lower values (Disney 1976). Recently, there have been a large number of efforts to quantify the bivariate luminosity–surface brightness distribution (de Jong 1996, McGaugh 1996, Dalcanton *et al* 1997, Sprayberry *et al* 1997, de Jong and Lacey 1999). While the results are somewhat different in terms of their details, they suggest that the surface brightness distribution of galaxies peaks around 22 b_J mag/arcsec^2 with a spread of $\sim 1 - 1.5$ mag/arcsec^2. The luminosity function of spiral galaxies is also nicely described by the Schechter function.

The typical values of the dimensionless angular momentum of collapsed halos (~ 0.05) are considerably smaller than that of the largely flattened centrifugally-supported disk galaxies we observe in our universe today ($\sim 0.4 - 0.5$), so considerable dissipation must occur to produce these disks. Without the presence of dissipationless dark matter, the collapse would proceed in such a way that the total angular momentum $\boldsymbol{J}$ and the total mass m would be conserved, but the energy would scale as $1/R$ where R is the collapse factor, so that $\lambda \propto JE^{1/2}M^{-5/2} \propto 1/\sqrt{R}$. The disk would then need to collapse by a factor of $(0.5/0.05)^2 \sim 100$ to obtain its observed dimensionless angular momenta, and this would take longer than the age of the universe for a 10-kpc disk! However, if the gas cloud collapses inside a dark matter halo, for which it represents only a fraction f of the mass, then the angular momentum $\boldsymbol{J}$ and mass $\boldsymbol{M}$ would scale by a factor f and the energy $\boldsymbol{E}$ would scale by a factor f^2, so that $\lambda \propto JE^{1/2}M^{-5/2} \propto 1/(f^{1/2}R^{1/2})$. For a typical estimate of the baryon fraction, $f \sim 0.1$, the gas cloud would then only need to collapse by a factor of 10, easily accommodated in current theories.

Despite the simplicity of this picture, a significant portion of the available gas cools to form GALAXIES AT HIGH REDSHIFT. Detailed simulations which follow the evolution and merging of these galaxies into larger and larger systems produce disks whose sizes are much smaller than those observed (Steinmetz and Navarro 1999) because of substantial angular momentum transfer from the baryons to their dissipationless halos.

Feedback

A nonnegligible fraction of stars end their lives as SUPERNOVAE, injecting much of this energy into the ambient gaseous medium. This energy serves to heat the gas, either expelling it from the star-forming environment or making it too hot to be conducive to star formation. Hence, the formation of stars serves to suppress further star formation and hereby regulates itself. This process is quite logically called feedback. The presence of feedback, particularly in disk galaxies, explains why the conversion of gas into stars frequently requires ten to hundreds of dynamical time scales ($\sim 10^{10}$ yr) instead of just several dynamical time scales ($\sim 10^8$ yr).

Feedback also provides the preferred explanation for the flattening of the luminosity function relative to the mass function at low masses (see the section above on galaxy luminosity function). DWARF GALAXY potential wells are shallow, and interstellar gas is readily energized above the escape velocity and therefore blown out in a galactic wind. Evidence for galactic winds is commonly found for STARBURST GALAXIES, often of relatively low mass.

Elliptical galaxy formation

Ellipticals make up the other principal component of the local galaxy census. Ellipticals possess elliptical isophotes with projected ellipticities $\epsilon = a/b$ (a being the major axis and b the minor) ranging from 0 to 0.7, the former being denoted an E0 and the latter an E7. Low redshift ellipticals possess an abundance of low-mass stars and are therefore very red. The lack of short-lived blue stars is generally taken as an indication that these galaxies are very old and have not formed stars for at least 5–10 Gyr. Like spirals, the luminosity function for ellipticals can also be described by a Schechter function, but with a much shallower faint-end slope (Bromley *et al* 1998, Folkes *et al* 1999). Unlike spirals, ellipticals are predominantly found in dense regions, i.e., galaxy clusters (Dressler 1980).

Ellipticals are known to have approximately de Vaucouleurs surface brightness profiles:

$$I(r) \propto \exp(-7.67(r/r_e)^{1/4}) \tag{104}$$

where r is the radius and r_e is the half-light radius. To higher order, the surface brightness profiles of ellipticals show an important dichotomy. Some ellipticals, known as disky ellipticals, appear to have power-law profiles all the way into their center, and other ellipticals, known as boxy ellipticals, exhibit a sharp break from this power-law at some radius near the center.

Like spirals, the global structural properties of ellipticals are known to populate a two-dimensional manifold, commonly known as the fundamental plane. These are known according to various names: the Faber–Jackson (Faber and Jackson 1976) relationship ($L \propto \sigma^4$), the Kormendy luminosity–radius (Kormendy 1977) relationship, and the D_n-σ (Dressler *et al* 1987) relationship. It has largely been agreed that the fundamental plane is essentially a consequence of the virial theorem and a relatively homologous formation scenario where the mass-to-light ratio varies as a small power of the mass ($M/L \propto M^{1/6}$).

There are two prevailing scenarios for the formation of elliptical galaxies: one in which ellipticals formed as the result of mergers from spiral galaxies and one in which ellipticals formed at high redshift from monolithic collapse. We begin by presenting the monolithic collapse scenario.

Monolithic collapse

One possible mechanism for the formation of elliptical galaxies is the early formation of stars from the gas collapsing onto the center of a dark halo. Early collapse and fragmentation into stars prior to the collapse of the halo can constitute the core of the elliptical while

stars formed from the secondary gas infall can constitute the shallower wings. An examination of the velocity-dispersion rotational-rate relationship demonstrates that ellipticals are essentially pressure-supported and that rotational flattening is not important in imparting ellipticity to these galaxies. In fact, detailed comparisons show that the dimensionless angular momenta of slow-rotating ellipticals seems to be no larger 0.05. In order to obtain the typical mass M ($\sim 10^{11} M_{\odot}$), radius R (~ 10 kpc), and angular momentum without recourse to dissipation, it would be necessary for the halo to collapse at redshifts beyond 10. On the other hand, with dissipation, one could easily obtain galaxies of the desired mass and radius, but the dimensional angular momentum would be too large (unless the initial angular momentum for the halo just happened to be particularly small), and the galaxy would resemble a disk.

Merger-based origin

Another mechanism for the formation of ellipticals is through the mergers of spiral galaxies. This mechanism provides a natural way of resolving the angular momentum problem, the crucial point being that since the spin angular momenta are randomly oriented with respect to each other, the resultant spin angular momentum for the formed elliptical can be considerably smaller than the spin angular momentum of the colliding disks. There are a number of other attractive features to this scenario. First, there are numerous examples of disk galaxies merging to form objects with de Vaucouleurs profiles in the local universe (Schweizer 1982, 1986), and it is quite conceivable that mergers were more frequent in the past. Secondly, nearly half of elliptical galaxies (Malin and Carter 1983, Schweizer and Ford 1984) possess features, such as shells or other sharp features, indicative of mergers or an otherwise violent formation. Thirdly, detailed N-body simulations of collisions between disk galaxies embedded in dark halos produce galaxies with de Vaucouleur profiles similar to those found in nearby ellipticals. Fourth, the GLOBULAR CLUSTER populations around ellipticals have bimodal metallicity distributions, indicative of a multi-stage formation scenario (Ashman and Zepf 1992, Zepf and Ashman 1993). All these features point toward the conclusion that at least *some* ellipticals formed by merging.

Before discussing the relative merits of the two formation scenarios for ellipticals, it is interesting to look at several of the difficulties which only arise in the merging scenario because of the close relationship between ellipticals and their progenitors (spirals). First, the energy per particle and phase space density are higher at the centers of ellipticals than any observed spiral, and therefore the merging process must be accompanied by a great deal of gas dissipation and cooling both to form a much deeper central potential and to obtain the high phase space density observed there if we presume this scenario is correct. In fact, nuclear starbursts are frequently observed to accompany such mergers (Schweizer 1990). Second, the number of globular clusters ($\sim 10^4$–$10^5 M_{\odot}$ compact star clusters) per unit luminosity for ellipticals is typically 4–10 times larger than that for spirals (van den Bergh 1990), so disk-disk mergers must result in the formation of a large number of globular clusters if we presume this scenario is correct. Finally, while ellipticals might be expected to show relative alpha-to-iron abundances typical of spirals, ellipticals contain significantly larger abundances of alpha elements than iron elements, so a substantial fraction of the stars present in ellipticals must have formed in the merger events between two spiral galaxies.

A comparative evaluation

The principal observational differences between the monolithic collapse and merger scenarios for elliptical formation concern their predictions for the formation history of ellipticals. Monolithic scenarios tend to form elliptical galaxies at very high redshifts ($z > 3$) while the elliptical population builds up more gradually in hierarchical scenarios.

Consequently, the merger scenarios, with their more diverse and contemporary formation histories, show more scatter in both the colour-magnitude relationship and the fundamental plane than monolithic collapse scenarios. Observationally speaking, ellipticals show a high degree of uniformity both in their small color-magnitude scatter, i.e. $\sigma(U - V) = 0.15$ (Bower *et al* 1992) and their tightness around the fundamental plane (Renzini and Ciotti 1993). This observed tightness about the fundamental plane extends to $z \sim 1$ (Aragon-Salamanca *et al* 1993, Stanford *et al* 1998). The observed tightness supports a monolithic collapse scenario where ellipticals form early and somewhat coevally. Of course, in hierarchical scenarios, most galaxies assemble quite early ($z \sim 2$) in the rich clusters, where the most compelling examples of tight fundamental planes are observed, so apparent difficulties with this scenario are not as strong as they first might seem (Kauffmann and Charlot 1998a).

Due to the different formation times for ellipticals, these scenarios also yield remarkably different predictions for the evolution in the number density of early-type galaxies as a function of redshift. While there has been an increasing number of studies reporting a devolution in the number and luminosity of ellipticals at high redshift relative to that found in the local universe (Kauffmann and Charlot 1998b, Kauffmann *et al* 1996, Zepf 1997, Barger *et al* 1999, Menanteau *et al* 1999) as would be expected in a hierarchical scenario where their formation is more gradual, these results remain somewhat controversial (Broadhurst and Bouwens 1999, de Propris *et al* 2000).

Another important difference between these scenarios is the star formation rates they predict at high redshift. In the hierarchical scenario, galaxies start out small and slowly build up to the massive entities we observe in the universe today. Clearly, we do not expect large star formation rates here at early times except possibly when two galaxies merge. On the other hand, in the monolithic scenario, ellipticals need to undergo huge star formation

rates ($\sim 100\,M_\odot$/yr) to form the typical $10^{11}\,M_\odot$ stars observed in nearby giant elliptical galaxies at high redshift since there is only a period of $\sim 10^9$ yr available. In fact, very few galaxies with these huge star formation rates ($\sim 100\,M_\odot$/yr) have been found in either emission-line searches or Lyman-dropout searches at moderate redshifts ($1 < z < 5$), pointing to either a high redshift of formation or dust-enshrouding. Recently, however, SCUBA results and subsequent follow-up work have revealed a population of ultraluminous infrared galaxies at moderate redshifts with high enough star formation rates ($\sim 100\,M_\odot$/yr) to match those needed in a monolithic collapse scenario. Nevertheless, the exact nature of this population, its number density, and its relevance remain unclear.

Bulges

Many spiral galaxies feature a bulge, or a spheroid, at their centers. Spheroids resemble elliptical galaxies in many important respects including their overall appearance and placement in the fundamental plane. This suggests that bulges are nothing but elliptical galaxies upon which gas has later accreted. Note, however, that somewhat contrary to ellipticals, in particular ellipticals with boxy isophotes, is the presence of considerable rotational flattening in many bulges (Davies *et al* 1983, Davies 1987). This is in agreement with what one might expect from dissipational collapse and, in particular, from the formation of bulges via disk instabilities (van den Bosch 1998).

Summary

While there are many things we do not understand about galaxy formation, many pieces of the picture now seem to be clear. Galaxies seem to form in a homogeneous, isotropic universe that is expanding according to Friedmann's equations. Inflation, though not unique, appears to be a relatively successful way of producing the scale-free spectrum of density fluctuations out of which galaxies have formed. Growth of the fluctuations can be followed initially with linear growth theory and later using a spherical collapse model. Press–Schechter theory provides a relatively successful way of putting these ingredients together to predict the mass spectrum of collapsed objects. The relative magnitudes of the cooling and dynamical time scales are important for determining the mass range of galaxies, galaxies forming when the cooling time is smaller than the dynamical time. Disk galaxies form from the cooling of gas onto the centers of collapsed halos, the gas settling into a disk supported by its angular momentum. Elliptical galaxies, on the other hand, seem to form by disk–disk mergers or by gas cooling within a halo of low intrinsic angular momentum (monolithic collapse).

Many important questions remain. For example, what is the relative importance of different mechanisms for the formation of both ellipticals and bulges? How do the sizes, luminosities, star formation rates, number densities, and metallicities of various galaxy types evolve over the history of the universe? What mechanisms are responsible for the tight correlation between the global properties of ellipticals and spirals? While theoretical simulations are becoming increasingly sophisticated, the inherent nonlinearity of galaxy formation processes make the role of new observations tantamount. To give the reader a taste of the improvements we will see in the next ten years in probing galaxy formation in the most remote regions of the universe, in figure 2 we have included some simulations for a hierarchical merging model using two current generation instruments (WFPC2 and NICMOS) and two future generation instruments (ACS and NGST). The obvious increase in depth will clearly bring our already moderately mature understanding of galaxy formation further into focus.

Bibliography

Aragon-Salamanca A, Ellis R S, Couch W J and Carter D 1993 *Mon. Not. R. Astron. Soc.* **262** 764
Ashman K M and Zepf S E 1992 *Astrophys. J.* **384** 50
Bardeen J M, Bond J R, Kaiser N and Szalay A S 1986 *Astrophys. J.* **304** 15
Barger A J, Cowie L L, Trentham N, Fulton E, Hu E, Songaila A and Hall D 1999 *Astron. J.* **117** 102
Binney J 1977 *Astrophys. J.* **215** 483
Bower R G, Lucey J R and Ellis R S 1992 *Mon. Not. R. Astron. Soc.* **254** 601+
Broadhurst T and Bouwens R 2000 *Astrophys. J.* **530** 53
Bromley B C, Press W H, Lin H and Kirshner R P 1998 *Astrophys. J.* **505** 25
Catelan P and Theuns T 1996 *Mon. Not. R. Astron. Soc.* **282** 455
Cole S and Lacey C 1996 *Mon. Not. R. Astron. Soc.* **281** 716+
Dalcanton J, Spergel D N, Gunn J E, Schmidt M and Schneider D P 1997 *Astron. J.* **114** 2178+
Davies R 1987 *Structure and Dynamics of Elliptical Galaxies* IAU Symposium vol 127, ed T de Zeeuw (Dordrecht: Reidel)
Davies R L, Efstathiou G, Fall S M, Illingworth F and Schechter P L 1983 *Astrophys. J.* **266** 41
de Bernadis P *et al* 2000 *Nature* **404** 955
de Jong R and Lacey C 1999 *Astrophys. Sp. Sci.* **269** 569
de Jong R S 1996 *Astron. Astrophys.* **313** 45
de Propris R, Sanford A, Eisenhardt P, Dickenson M and Elston R 1999 *Astron. J.* **118** 719
Disney M J 1976 *Nature* **263** 573
Dressler A 1980 *Astrophys. J.* **236** 351
Dressler A, Lynden-Bell D, Burstein D, Davies R L, Faber S M, Terlevich R and Wegner G 1987 *Astrophys. J.* **313** 42
Faber S M and Jackson R E 1976 *Astrophys. J.* **204** 668
Folkes S, Ronen S, Price I, Lahav O, Colless M, Maddox S, Deeley K, Glazebrook K, Bland-Hawthorn J, Cannon R, Cole S, Collins C, Couch W, Driver S, Dalton G, Efstathiou G, Ellis R, Frenk C, Kaiser N, Lewis I, Lumsden S, Peacock J, Peterson B, Sutherland W and Taylor K 1999 *preprint* astro-ph/9903456
Freeman K C 1970 *Astrophys. J.* **160** 811+
Hanany S *et al* 2000 *preprint* astro-ph/0005123
Jing Y P and Suto Y 2000 *Astrophys. J* **529** L69

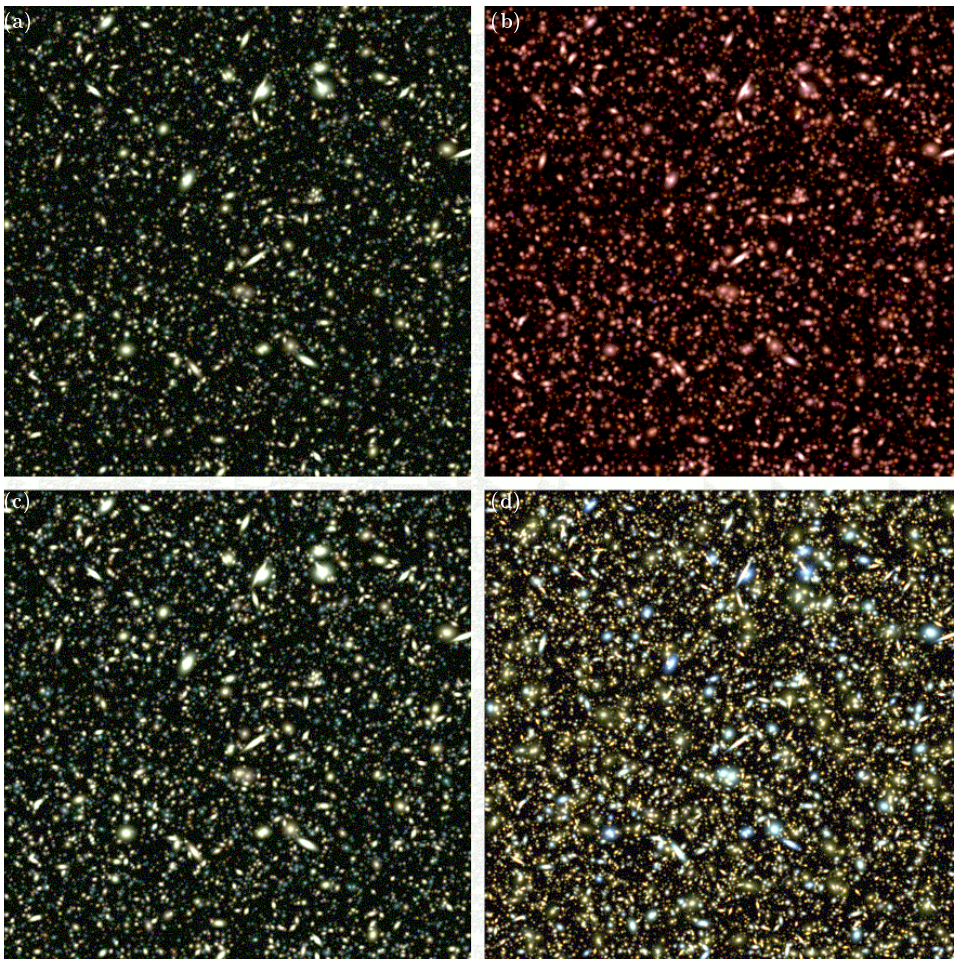

Figure 2. A simulation of HDF-depth ($\sim$ 30 orbits) exposure over a $7' \times 7'$ field of view for WFPC2 (a), NICMOS (b), ACS (c) and NGST (d). WFPC2 and NICMOS are instruments currently on the Hubble Space Telescope. ACS is an instrument that will be placed on the Hubble Space Telescope in 2001, and NGST is a completely new space telescope, which will be launched in $\sim$ 2007. A hierarchical merging model was used as the model inputs for the simulation (see Bouwens and Silk 1999 for details). 5 μm, 3 μm and 1 μm wavelengths are assumed for the RGB channels in the false-colour NGST image while for the NICMOS image, the K (2.2 μm), H (1.6 μm) and J (1.2 μm) bands are assumed. **This figure is reproduced as Color Plate 18.**

Kauffmann G and Charlot S 1998a *Mon. Not. R. Astron. Soc.* **294** 705+
Kauffmann G and Charlot S 1998b *Mon. Not. R. Astron. Soc.* **297** L23
Kauffmann G, Charlot S and White S D M 1996 *Mon. Not. R. Astron. Soc.* **283** L117
Kormendy J 1977 *Astrophys. J.* **218** 333
Malin D F and Carter D 1983 *Astrophys. J.* **274** 534
McGaugh S S 1996 *Mon. Not. R. Astron. Soc.* **280** 337
Menanteau F, Ellis R, Abraham R, Barger A and Cowie L 1999 *Mon. Not. R. Astron. Soc.* **309** 208
Mo H J, Mao S and White S D M 1998 *Mon. Not. R. Astron. Soc.* **295** 319
Moore B, Quinn T, Governato F, Stadel J and Lake G 1999 *Mon. Not. R. Astron. Soc.* **310** 1147
Navarro J F, Frenk C S and White S D M 1997 *Astrophys. J.* **490** 493+
Peacock J A 1997 *Mon. Not. R. Astron. Soc.* **284** 885
Peebles P 1980 *The Large-Scale Structure of the Universe* (Princeton, NJ: Princeton University Press)
Press W H and Schechter P 1974 *Astrophys. J.* **187** 425
Rees M J and Ostriker J P 1977 *Mon. Not. R. Astron. Soc.* **179** 541
Renzini A and Ciotti L 1993 *Astrophys. J.* **416** L49
Schweizer F 1982 *Astrophys. J.* **252** 455
Schweizer F 1986 *Science* **231** 227
Schweizer F 1990 *Dynamics and Interactions of Galaxies* ed R Wielen (Berlin: Springer)
Schweizer F and Ford W K 1984 *Bull. Astron. Astrophys. Soc.* **16** 889+
Silk J 1977 *Astrophys. J.* **211** 638
Silk J 1997 *Astrophys. J.* **481** 703+
Sprayberry D, Impey C D, Irwin M J and Bothun G D 1997 *Astrophys. J.* **482** 104+
Stanford S A, Eisenhardt P R and Dickinson M 1998 *Astrophys. J.* **492** 461+

Steinmetz M and Bartelmann M 1995 *Mon. Not. R. Astron. Soc.* **272** 570
Steinmetz M and Navarro J F 1999 *Astrophys. J.* **513** 555
van den Bergh S 1990 *Dynamics and Interactions of Galaxies* ed R Wielen (Berlin: Springer)
van den Bosch F C 1998 *Astrophys. J.* **507** 601
Warren M S, Quinn P J, Salmon J K and Zurek W H 1992 *Astrophys. J.* **399** 405
Zel'dovich Y B 1970 *Astron. Astrophys.* **5** 84+
Zepf S E 1997 *Nature* **390** 377+
Zepf S E and Ashman K M 1993 *Mon. Not. R. Astron. Soc.* **264** 611+

Joseph Silk and Rychard Bouwens