Galaxy formation

P. J. E. Peebles†

Joseph Henry Laboratories, Princeton University, Princeton, NJ 08544

ABSTRACT It is argued that within the standard Big Bang cosmological model the bulk of the mass of the luminous parts of the large galaxies likely had been assembled by redshift \( z \sim 10 \). Galaxy assembly this early would be difficult to fit in the widely discussed adiabatic cold dark matter model for structure formation, but it could agree with an isocurvature version in which the cold dark matter is the remnant of a massive scalar field frozen (or squeezed) from quantum fluctuations during inflation. The squeezed field fluctuations would be Gaussian with zero mean, and the distribution of the field mass therefore would be the square of a random Gaussian process. This offers a possibly interesting new direction for the numerical exploration of models for cosmic structure formation.

Galaxies are not mentioned much in modern studies of structure formation; the emphasis is on phenomena on the smaller mass scales of the Lyman \( \alpha \) forest clouds and on the larger scales of the extended dark matter concentrations that might be sites for formation of galaxies and clusters of galaxies. That is, the focus is on simpler problems, involving systems that are not strongly nonlinear and not likely to have been greatly disturbed by star formation. The path is sensible but not a very direct route to the lessons to be drawn from the galaxies.

An important lesson may be found in the high mass densities within galaxies. The mean density within our position in the Milky Way is equivalent to about one proton \( \text{cm}^{-3} \), six orders of magnitude greater than the large-scale average. Why is the contrast so big? In the commonly discussed CDM model, protogalaxies form at relatively low redshift and reach high contrast because they collapse by a large factor. Another possibility is that the mass was assembled at high redshift, when the mean density was larger and the collapse factor thus more modest. In the next section I argue that the latter agrees with the interpretation of rich clusters of galaxies as scaled versions of protogalaxies, and that late assembly is not easy to reconcile with constraints from the physics and astrophysics.

We must take care to distinguish the epochs of formation of star populations from the epoch of assembly of the bulk of the mass \( (1) \). If, as the evidence reviewed here suggests, the latter happened early, we likely will have to find an alternative to the adiabatic CDM model for structure formation. The success of the standard picture for the light elements [as discussed by Schramm in this issue of the Proceedings (2)] leads me to stay with the idea that the dominant mass in clusters of galaxies and the outer parts of individual galaxies is nonbaryonic, the commonly discussed CDM. A picture that seems worth exploring is that this CDM is the remnant of quantum fluctuations frozen into a massive scalar field during inflation.

Clusters of Galaxies as Scaled Protogalaxies

We have a remarkably tight set of checks on the mass distribution within a cluster. The gravitational potential well needed to contain the motions of the galaxies agrees with what is needed to contain the intracluster plasma detected by its x-ray emission and by its effect on the thermal background radiation. The gravitational potential also agrees with the observations of gravitational lensing. A useful approximation to the mass distribution is a limiting isothermal sphere with mass \( M(< r) = 2\sigma r^2/G, \) where \( \sigma \) is the line-of-sight velocity dispersion. Characteristic numbers for clusters are

\[
r_{cl} = 1.5 h^{-1} \text{Mpc}, \quad \sigma = 750 \text{ km s}^{-1},
\]

where \( r_{cl} \) is the Abell radius, \( \sigma \) is an rms mean for Abell clusters, Mpc is megaparsecs, and Hubble’s constant is \( H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1} \). The isothermal model also gives a useful first approximation to the mass distribution within the luminous parts of L* spiral galaxies (objects like the Milky Way that contribute the bulk of the luminosity density); characteristic numbers here are

\[
r_{g} = 10 h^{-1} \text{kpc}, \quad v_c = 200 \text{ km s}^{-1},
\]

where \( v_c \) is the circular velocity and kpc is kiloparsecs. I have chosen the radius \( r_g \) to be large enough to include most of the luminous and enough of the dark halo that, in the standard model for an L* spiral, the contained mass typically is dominated by CDM.

The mass density contrasts within the radii in Eqs. 1 and 2 are:

\[
\frac{3M_{cl}}{4\pi r_{cl}^3 \rho_b} \approx \frac{100}{\Omega}, \quad \frac{3M_g}{4\pi r_{g}^3 \rho_b} \approx \frac{10^5}{\Omega(1 + z_g)^3}.
\]

The cosmic mean mass density is \( \rho_b \), and the density parameter \( \Omega \) is the ratio of \( \rho_b \) to the critical Einstein–de Sitter value. The contrast for clusters is evaluated at the present epoch. The contrast for galaxies could be evaluated at redshift \( z_g \) if galaxies existed then with about the same characteristic mass within the physical radius \( r_g \). The fractions of the net mass of the universe within radii \( r_{cl} \) and \( r_g \) in clusters and in galaxies are:

\[
\frac{n_{cl}M_{cl}(< r)}{\rho_b} \sim \frac{0.006}{\Omega}, \quad \frac{n_gM_g(< r)}{\rho_b} \sim \frac{0.003}{\Omega}.
\]

where \( n_{cl} \) and \( n_g \) are the characteristic number densities of clusters and galaxies. The ratio of the object size to the separation defined by the number density is

Abbreviations: CDM, cold dark matter; CBR, cosmic background radiation.

†To whom reprint requests should be addressed. e-mail: pjep@pupgprinceton.edu.

© 1998 by The National Academy of Sciences 0027-8424/98/9567-582.00/0
PNAS is available online at http://www.pnas.org.
As in Eq. 3 this ratio could be evaluated at redshift \( z_g \) if the galaxies were present then. Gravity is scale-invariant at distances small compared with the Hubble length, and the expansion of the universe is scale-invariant or close to it. If the initial conditions were scale-invariant and we could ignore nongravitational interactions then structure formation would follow a scaling law where the length scale as a function of redshift is set by the spectrum of primordial density fluctuations. Eqs. 3–5 show that galaxies at redshift

\[ z_g \sim 10 \]  

resemble scaled versions of present-day clusters of galaxies.

We might pause to consider the slope of the power spectrum of density fluctuations that would scale protogalaxies to clusters. The ratio of comoving wavenumbers that enclose the mass defined by Eqs. 1 and 2 is

\[ \frac{k_g}{k_{cl}} = \left( \frac{M_{cl}(<r_g)}{M_g(<r_g)} \right)^{1/3} \sim 15. \]  

The ratio of the mean square primordial density fluctuations on these two scales is

\[ \frac{k_g^2P(k_g)}{k_{cl}^2P(k_{cl})} = \left( \frac{k_g}{k_{cl}} \right)^{3+n} = (1 + z_g)^2D(\Omega)^2. \]  

The power spectrum is assumed to vary as \( P \propto k^n \). The last expression is the ratio of the gravitational growth factor up to the present for clusters and up to redshift \( z_g \) for protogalaxies, where \( D(\Omega) \) is the departure from the Einstein–de Sitter growth factor. In a cosmologically flat model \( D(\Omega) \) is close to unity, and Eqs. 7 and 8 with \( z_g \sim 10 \) require \( n \sim -1.3 \). This is not far from the measured slope of the power spectrum on larger scales where the present galaxy distribution is thought to be linearly related to the primordial mass distribution. That is, the scaling picture seems to be consistent with what is known about the present mass distribution.

Substructure shows that there is considerable ongoing merging and relaxation in present-day clusters of galaxies. Within this scaling picture one similarly would expect substantial merging and relaxation at \( z \sim 10 \) in young galaxies that are considerably more irregular than they are now. The main point is that because nature manifestly is capable of producing bound mass concentrations on the scale of clusters now it is capable of doing about the same on the scale of galaxies at \( z = 10 \).

The value for \( z_g \) from this argument depends on the choice of \( r_g \) in Eq. 2. A larger radius would yield a smaller value for \( z_g \) which in this picture would be interpreted as a characteristic epoch for the assembly of the extended dark halos of spiral galaxies. A substantially smaller value for \( r_g \) would probe the baryon-dominated parts of galaxies, which certainly can behave differently from the CDM that is assumed to dominate clusters and the outer parts of individual spirals.

The scaling picture would fail if the mass concentration within the Abell radius of a cluster typically were evolving on the scale of the Hubble time, but this is inconsistent with the evidence that clusters are close to dynamical equilibrium at the Abell radius, and that many clusters at \( z = 0.5 \) have central masses comparable to present-day systems [as discussed in this issue of the Proceedings by Mushotzky (3)]. The picture would fail if the spectrum of primordial density fluctuations were cut off at about the mass of a galaxy: pancake collapse at the cutoff might allow a larger collapse factor than is seen in clusters of galaxies. The point is worth exploring but not relevant to the adiabatic CDM model or the proposed isocurvature variant. I turn now to an issue that could be very relevant, the breaking of scaling by dissipation by gaseous baryons.

**Dissipative Galaxy Assembly at Redshift \( z \leq 3 \)**

If protogalaxies were assembled well after redshift \( z = 10 \) they would have to have collapsed by a considerable factor to reach the densities characteristic of the bright parts of galaxies. Gaseous baryons can be strongly dissipative, and dissipation can result in strong settling. This is a commonly held basis for the assumption of relatively late galaxy assembly, but some issues should be considered.

Fig. 1 shows models for galaxy assembly at relatively early and late epochs. The curves contain baryonic and total masses

\[ M_B = 1 \times 10^{11}M_{\odot}, \quad M = 6 \times 10^{11}M_{\odot}. \]  

The former is typical of an \( L^* \) spiral like the Milky Way. The ratio of baryonic to CDM mass is in the range now under discussion (4, 5); the uncertainty in the ratio does not affect these considerations much. The background cosmological model has density and Hubble parameters

\[ \Omega = 0.3, \quad h = 0.7. \]  

If the mass distribution may be modeled as spherical and homogeneous (to avoid orbit crossing) then the relation between the nominal redshift \( z_c \) at collapse to zero radius and the radius at maximum expansion may be approximated as

\[ R_{max} = \frac{2(G]\psi)^{1/3}}{(3\pi H_o)^{2/3}V^{1/3}} \frac{1}{1+z_c} = \frac{500}{1+z_c} \text{ kiloparsecs.} \]  

The second expression uses the numbers in Eqs. 9 and 10. The curves in Fig. 1 follow the spherical model through maximum expansion and nearly free collapse to a radius typical of a large spiral galaxy. In a spherical and homogeneous mass distribution the kinetic energy vanishes at maximum expansion, and collapse by a factor of two in radius produces enough kinetic energy for virial equilibrium at density contrast \( \sim 100 \). A real protogalaxy or protocluster likely has appreciable kinetic energy at nominal maximum expansion, which would tend to reduce the collapse factor. Thus it is not surprising that clusters are close to equilibrium at least as far out as the Abell radius.
at the density contrast in Eq. 3. In the model for late galaxy assembly in Fig. 1, the radius at maximum expansion is 130 kiloparsecs, and the protogalaxy must dissipatively collapse by more than an order of magnitude in radius. Is this reasonable? I offer two points.

First, in models like adiabatic CDM and the isocurvature variant, structure formation is hierarchical: protogalaxies are assembled out of subconcentrations that were assembled out of sub-subconcentrations. If nongravitational dissipation of nonradial motions could cause the baryons in a protogalaxy to collapse nearly freely by a factor of 10 in radius, why would the same effect not have caused strong collapse earlier in major subconcentrations? In the late formation scenario in Fig. 1 that would have produced dwarf galaxies in a group with diameter ~100 kiloparsecs, which would take a long time to merge into something resembling the spheroid component of a large galaxy. The problem is avoided if nonradial motion prevented catastrophic collapse of the major subconcentrations, but the same effect of nonradial motion would have slowed collapse of the protospheroids, contrary to the late formation model in Fig. 1.

Second, the evidence is that in large spirals the rotation curve is close to flat from the inner parts dominated by baryons to the outer parts that are presumed to be dominated by dissipationless CDM. In the late assembly picture, dissipative collapse increases the baryon density by a large factor. The growing potential would drag CDM with it, but the increase in density has to be much less than for the baryons because the CDM is not allowed to dissipate energy. Why would the large difference of collapse factors for the baryons and the CDM not have produced a Keplerian feature in the rotation curves of spiral galaxies?

Galaxies at High Redshift

The remarkable advances in observations of galaxies at z > 1, described in this issue of the Proceedings by Steidel (6), offer valuable clues to the epoch of galaxy assembly. We have to consider the possible difference between the epochs of formation of star populations and of the assembly of the mass, however (1). The peak in the comoving number density of quasars at z ~ 2.5 may be evidence for late galaxy assembly, or instead it could be a measure of the time required for the development of an active galactic nucleus central engine after assembly of the protogalaxy. A peak in the star formation rate is at z ~ 1 (7), but a major contributor is a population of relatively low mass galaxies that since have faded or merged (8); the evidence is that the large spirals and ellipticals are in place at z ~ 1. The small scatter in the color-magnitude relation suggests the stars in ellipticals and the spheroid components of the spirals formed at high redshift. Kauffmann (9) shows that this does not contradict the adiabatic CDM model; the stars can have formed before the ellipticals were assembled. On the other hand, Prochaska and Wolfe (10) show that the damped Lyman alpha systems at z ~ 3 have the properties expected of rotationally supported gaseous discs; they are reasonable candidates for protodiscs draped around the spheroid bulges in the dark matter halos of protospirals, an example of star formation well after assembly of the mass.

A CDM Model for Early Galaxy Assembly

I turn now to a model for early galaxy assembly within the inflation scenario for the very early universe. These qualitative comments are meant to show that a more detailed assessment by numerical simulations could be interesting.

In early discussions of the inflation scenario people considered two ways homogeneity could be broken. In the first model, quantum fluctuations δφ(ξ) frozen into the field φ whose energy is driving inflation produce adiabatic fluctuations with mass density contrast

$$\delta(\xi) = \delta\rho/\rho \propto \delta\phi(\xi). \quad [12]$$

The perturbation to the density is proportional to the perturbation to the field, because |δφ| ≪ |φ|. In simple models for inflation this makes the primeval mass density contrast δ(ξ) a random Gaussian process characterized by a power spectrum (the Fourier transform of the mass autocorrelation function) that is close to a power law,

$$P_\rho(k) \propto P_\delta(k) \propto k^n. \quad [13]$$

The case n = 1 first attracted attention because the perturbation to space-time curvature diverges only as the logarithm of k, so the power law can apply to a broad range of scales. In the second model a single-component scalar field ψ that is present during inflation ends up as (or decays into) the present-day CDM. The field is supposed to acquire mass m_ψ, so on scales large compared with the Compton wavelength m_ψ/c, the CDM mass distribution well after inflation is

$$\rho_\psi(\xi) \sim m_\psi^2 \delta(\xi)^2. \quad [14]$$

Here the mass fluctuations are non-Gaussian.

If inflation is cool, adiabatic departures from homogeneity may be subdominant. Under this isocurvature condition, the net mass distribution is homogeneous at high redshift, and the composition is a function of position. In the model in Eq. 14 the inhomogeneity is the local ratio of mass in the CDM to the mass in all the other components. If ψ(ξ) is a random Gaussian process with zero mean the autocorrelation function of the mass distribution in the field is

$$\xi_\rho(x, z) = \frac{\langle p_\rho^2 \rangle}{\langle p^2 \rangle} - 1 = 2 \frac{\langle p_\psi^2 \rangle}{\langle p^2 \rangle} - 1. \quad [15]$$

If the power spectrum of ψ(ξ) is close to a power law,

$$P_\psi(k) \propto k^m, \quad [16]$$

with −3 < m_ψ < 0, then the autocorrelation function of ψ(ξ) is

$$\xi_\rho(x) = \frac{\langle p_\psi^2 \rangle}{\langle p^2 \rangle} = \frac{x^3}{x^2}, \quad [17]$$

and the mass fluctuation power spectrum from Eq. 15 is

$$P_\rho(k) \propto k^m, \quad m = 3 + 2m_\psi. \quad [18]$$

for m_ψ < −1.5. Continuity requires δ_ψ(ξ) → 1 at x → 0, so P_ψ must break from the power law form in Eqs. 16 and 17 at a comoving length x_1 shortward of which it might approach the scale-invariant form P_ψ ∝ k^{-3}. At x → 0 the primeval CDM mass density fluctuations in this model are mildly nonlinear, δρ/ρ → 2^{1/2} (Eq. 15).

In simple inflation scenarios the power law indices are n = 1 in the adiabatic model in Eq. 13 and m_ψ = −3 in the isocurvature model in Eq. 18. The demonstration by Efstathiou and Bond (11) that the latter case violates the bounds on the large-scale anisotropies of the thermal cosmic background radiation (CBR) was a significant factor in focusing the attention of the community on the adiabatic case with n = 1. The Cosmic Background Explorer measurement of the large-scale CBR anisotropy showed that if Ω = 1 the adiabatic power law index has to be tilt to n < 1 to avoid unacceptable small-scale mass fluctuations. Tilt also is possible and surely equally socially acceptable in the isocurvature case; a value m_ψ = −2.5 gives a reasonable fit to the CBR angular fluctuations and the large-scale mass distribution. The details are presented
Fig. 2. Mass distributions with power spectrum $P \propto k^{-1.8}$, smoothed through a spherical window with radius equal to 1.2% of the box width in the figure. The lowest contour is at the mean and the contour interval is one standard deviation. The left side is a realization of a Gaussian process. The right side shows a realization of the CDM distribution in the model in Eq. 14.

elsewhere (12); here I comment on the possible relation of a tilted isocurvature model to the argument for early galaxy assembly.

Fig. 2 compares realizations of the primeval mass distributions in the Gaussian adiabatic and non-Gaussian isocurvature models. The power spectra are the same and give about the same fit to the second moments of fluctuations in the CBR and the large-scale galaxy distribution. In the isocurvature case the figure applies on scales larger than the cutoff $x_c$ (Eq. 17) and smaller than about 300 megaparsecs. (On larger scales the radiation pressure gradient force is not important and the net mass distribution ends up as it began, close to homogeneous, the primeval density fluctuations in the now dominant CDM being replaced with large-scale fluctuations in the distribution of the baryons and radiation. On scales <300 megaparsecs radiation pressure forces the radiation and baryons to remain nearly homogeneous up to decoupling, leaving the CDM with close to its primeval distribution.) The mass distributions in this figure are constructed from Gaussian realizations with zero means in a cube of $256^3$ cells, and the figure shows a cut through the cube. For the isocurvature case the Gaussian realization is squared to get the primeval CDM mass distribution, and the mean subtracted. Finally, the distributions are smoothed through a window of radius 3.1 cell widths. The fit to the second moments of the CBR and galaxy distributions requires similar mass fluctuation power spectra extrapolated to the present in linear perturbation theory, but one sees from Fig. 2 that there would be quite different galaxy formation histories in the adiabatic and isocurvature CDM models.

The adiabatic model gives a successful account of the formation and dissipation of the Lyman-α forest clouds at $z \sim 3$ (13, 14). In a Gaussian model it is reasonable to think that the galaxies, which are more massive than the clouds, form no earlier, that is, at redshifts <3, which I have argued does not seem reasonable. Three points suggest to me that the isocurvature version looks promising.

First, one sees in the figure that outside the concentrations of prominent upward density fluctuations conditions in the isocurvature case are not greatly different from the general mass distribution in the adiabatic CDM case. For example, in the probability distribution of the mass smoothed through a spherical window, about 90% of the mass is outside windows centered on peaks at $\delta \rho/\rho > 2\sigma$ in the Gaussian and the non-Gaussian cases with the same second moment. The adiabatic and isocurvature models thus might give similar accounts of the Lyman-α forest at $z \sim 3$. Second, at the same normalization of the power spectrum the prominent upward density fluctuations are a factor of three larger in the isocurvature case, meaning the first generation of galaxies would be assembled at $1 + z$ larger by a factor of about three. This could be consistent with the lower curve in Fig. 1, which we have seen agrees with the scaling relation of rich clusters and protogalaxies. Third, the fact that the prominent peaks in the isocurvature case tend to appear in concentrations qualitatively agrees with the clustering of young galaxies (15). At $\delta \rho/\rho > 6\sigma$ the correlation function of peak positions is $\langle \delta \rho \rangle = 3x^{-1} p_{\delta x} \rho = 1$ at $x$ equal to half the mean interpeak separation $x_p$, and $\langle \delta \rho \rangle \sim 0.2$ at $x = x_p$. Because the model assumes mass clusters with the galaxies we have to assume that the gravitational growth of clustering has emptied the voids and swept the dwarf galaxies into the concentrations of giants. Whether this could happen without causing the galaxies to cluster too strongly remains to be investigated. Other issues are discussed in ref. 11. It will be interesting to see tests of the issues by the now well developed methods of numerical N-body simulations of structure formation.

Concluding Remarks

If the evidence for early galaxy assembly really were as compelling as I have argued why would so many capable people have devoted so much time to the analysis of the adiabatic CDM model? The commonly held opinion is that the evidence for early assembly is not at all compelling, but rather that the recent galaxy assembly predicted by the CDM model is quite reasonable. Numerical N-body simulations of pure gravitational dynamics show the formation of regions with large density contrast, and dissipation by diffuse baryons naturally enhances the contrast. But the conclusions by several groups that the adiabatic CDM model successfully accounts for the properties of the great clusters of galaxies indicates to me that the collapse factor in pure gravitational simulations need not be so large. Dissipation in the baryon component certainly can greatly enhance collapse of a protogalaxy that collects around a single seed mass, but in models for hierarchical assembly of protogalaxies I cannot see how the proposed strong dissipative collapse would not instead produce unacceptably large numbers.
of compact groups of dwarf galaxies. It is only fair to admit that my colleagues by and large do not seem to be impressed, however.

The question is being put to observational tests. Key to the argument for early assembly is the idea that present-day clusters of galaxies yield a consistent scaled picture for young galaxies as irregular mass concentrations at redshift $z \sim 10$. Ongoing advances in observations of the evolution of the clusters back to $z = 1$ will help test this picture and may offer lessons on how protogalaxies evolved. The exploration of the universe at redshifts greater than unity already has given us a remarkably detailed picture of the history of the galaxies—a complex mix of evolution and stability—the meaning of which is under intense discussion. Further advances in observations of the state of the universe at high redshift will be followed with interest by all sides of the debate on how it all began, what it all might mean, and how it all will end.

This work was supported in part by the National Science Foundation.


