Mon. Not. R. astr. Soc. (1977) **179**, 541–559

Cooling, dynamics and fragmentation of massive gas clouds: clues to the masses and radii of galaxies and clusters

M. J. Rees *Institute of Astronomy, Madingley Road, Cambridge CB3 0HA*
J. P. Ostriker *Department of Astronomy, Princeton University, Princeton, New Jersey 08540, USA*

Received 1976 November 5; in original form 1976 July 7

Summary. We investigate the extent to which the characteristic masses and sizes of galaxies (and clusters) are determined by processes occurring at the epoch when the pregalactic material has stopped expanding with the background Universe but has not yet fragmented into stars. Unless pregalactic clouds collapse in an exceedingly homogeneous fashion, their kinetic energy of infall will be thermalized via shocks before the contraction has proceeded by more than a factor ~ 2. What happens next depends on the relative value of the cooling and collapse timescales. Masses in the range 10^{10}–$10^{12} M_\odot$ cool so efficiently that they always collapse at the free-fall rate, and probably quickly fragment into stars. Larger masses, however, may experience a quasi-static contraction phase; and go into free fall (and fragment) only after their radii fall below a critical value r_{bc}. For masses $\gtrsim 10^{12} M_\odot$, r_{bc} has a mass-independent value ~ 75 kpc. We argue that this characteristic mass and radius may indeed be crucial determinants of the properties of galaxies, and discuss various complications that should be included in more refined calculations. Large masses ($\sim 10^{14} M_\odot$) which recollapse at recent epochs may be unable to cool at all; and there is no reason why galaxies of characteristic radii $\lesssim r_{\mathrm{bc}}$ should not still be forming at small redshifts. The typical giant galaxies which dominate the luminosity function must have formed at a relatively recent epoch $z \lesssim 10$. Opacity effects in collapsing protogalaxies are briefly discussed, the possibility that trapped Lyman α may inhibit fragmentation being the only important one. The effects of mass-dependent dissipation on clustering and the covariance function are outlined.

1 Introduction

The evolution of small-amplitude primordial perturbations – regarded as the hypothetical progenitors of galaxies and clusters – has been extensively investigated in the framework of 'hot big bang' cosmologies; and there have been several recent discussions of how galaxies (and their stellar content) evolve *after* they have developed into isolated self-gravitating units. We here aim to bridge the gap between these two approaches: we focus attention on the *intermediate* stages when protogalaxies and protoclusters enter the non-linear regime, contract, and begin to fragment before settling down into virial equilibrium.

Galaxies have typical masses $\sim 10^{45}$ g and radii $\sim 10^{23}$ cm which cannot be determined by purely gravitational processes; Einstein's equations contain no characteristic scales. But during this intermediate stage, when the material is presumably still mainly gaseous, the effects of shock heating, pressure gradients and radiation cooling are important; and the behaviour depends critically on the ratio of the cooling timescale and the gravitational free-fall timescale. A simple order-of-magnitude study of such effects yields suggestive and encouraging results: characteristic radii and masses can be derived which are relevant to the dimensions of galaxies and clusters, and which tend to reduce estimates of the redshift at which these entities formed.

We show that – provided that irregularities surviving from the big bang have sufficiently large amplitudes to have by now entered the non-linear regime – straightforward physical processes, occurring at epochs potentially accessible to observation, can imprint preferred scales on an initially featureless spectrum. As a bonus, these scales (which are insensitive to the cosmological parameters and to uncertain physical conditions at very early times) do have values which are comparable with those of large galaxies, or small groups.

In the present paper we first (Section 2) summarize the 'gravitational clustering' picture of galaxy formation, indicating how the masses, radii and velocity dispersions of eventual bound systems depend (in this picture) on the initial amplitude of perturbations superimposed on a background Friedmann cosmology. We then discuss (Section 3) the behaviour of gravitating gas clouds, emphasizing particularly the significance of two mass-dependent radii: (i) the radius r_{ab} at which the cooling time equals the Hubble time; and (ii) the radius r_{bc}, within which the cooling time is shorter than the collapse timescale, so no quasi-static equilibrium is in general possible. In Sections 4 and 5 we apply these results to particular models for the formation of bound systems. Finally (Section 6) we assess the extent to which earlier estimates of the epoch of galaxy formation, the covariance function, etc., might be affected by gas-dynamical dissipation processes. We conjecture that the most readily observable parameters characterizing the large-scale Universe – the masses and radii of galaxies and clusters – are determined by gas-dynamical processes occurring at $z \lesssim 10$ rather than being sensitive to the precise spectrum of irregularities that survives from the primordial fireball.

In order to keep our discussion general but relatively short, we do not attempt to fit specific models nor make detailed comparison with observation. Our discussion is therefore of an order-of-magnitude illustrative character: quoted density estimates are 'mean' values, which may depend by as much as a factor ~ 10 on the dynamical details, adopted density profiles, etc.; quoted masses are equally sensitive to model-dependent factors; and radii, redshifts and timescales are uncertain by a factor of 2 or 3.

2 'Non-dissipative' gravitational clustering

As our starting point, we recall the simplest possible model for galaxy formation: the evolution of a uniform, pressure-free, spherical density enhancement in a Friedmann universe. The expansion of such a region lags behind, and stops at a 'turnaround time' t_{turn}, corresponding to a redshift z_{turn}. The value of z_{turn} depends on the amplitude $(\delta\rho/\rho)_{rec}$ of the fluctuations on the appropriate mass-scale at the recombination epoch t_{rec}. The effects of pressure, etc., can probably be justifiably neglected for galactic (and larger) masses, from t_{rec} until the formation of the first bound systems capable of reheating the gas to temperatures $\gg 10^4$ K. The density at turnaround is $(3\pi/4)^2 \simeq 5.5$ times the background critical density $\rho_c = (6\pi G t_{turn})^{-2}$ at that time. The perturbed region then collapses, and – if its energy is conserved, and it is transformed promptly into point masses – undergoes some kind of violent relaxation and settles down into virial equilibrium ('virializes') after a further time $\sim 2t_{turn}$. Its mean harmonic radius would then be $\sim \frac{1}{2}$ that of a uniform sphere of radius r_{turn}.

One can straightforwardly calculate a relation between the mass M of such a perturbation, the value of z_{turn}, and the mean harmonic radius r of the eventual virialized object (*cf.* Gott & Rees 1975, and references therein). For an Einstein–de Sitter model, where the density parameter $\Omega = \frac{8}{3}\pi G\rho H^{-2}$ is unity,

$$r = (270/h^{2/3})\, M_{12}^{1/3}\, (t_{turn}/t_0)^{2/3}\ \text{kpc}, \tag{1}$$

where M is the mass measured in units of $10^{12} M_\odot$, h is the present Hubble constant in units of 100 km/(s Mpc),* t_0 is the present age of the Universe ($\frac{2}{3}H_0^{-1}$ if $\Omega = 1$), and t_{turn} is related to z_{turn} by

$$t = \tfrac{2}{3}H_0^{-1}(1+z)^{-3/2}. \tag{2}$$

For cosmological models with $\Omega < 1$, (1) is modified by an extra term $(\frac{3}{2}H_0 t_0)^{2/3}$ on the RHS, and the relationship between t_{turn} and z_{turn} is now given by

$$dt/dz = -H_0^{-1}(1+z)^{-2}(1+\Omega z)^{-1/2}, \tag{3}$$

which makes $t_{turn} \propto (1+z_{turn})^{-1}$ for $(1+z_{turn}) \lesssim \Omega^{-1}$ and $t_{turn} \propto (1+z_{turn})^{-3/2}$ for $(1+z_{turn}) \gtrsim \Omega^{-1}$. For most of our discussion, it will prove convenient to think in terms of t_{turn} directly. When $\Omega \ll 1$, a given *time* corresponds to a larger redshift z; but most of our discussion is insensitive to redshifts as such.

Relation (1) quantifies the obvious expectation that the mean density within bound systems, being (in this idealized case when subsequent dissipative and dynamical evolution is ignored) ~ 45 times the critical density at the turnaround time, increases with z_{turn}, i.e. low-density bound systems cannot have formed too early.

In practice, the masses of galaxies and clusters are of course inferred from their measured internal velocity dispersions along the line of sight. Expressions equivalent to (1) involving the *velocity dispersion* may therefore be useful

$$v_{disp} \simeq 47 r_{100} h (t/t_0)^{-1} (\tfrac{3}{2}H_0 t_0)^{-1}\ \text{km/s} \simeq 127 M_{12}^{1/3} (t_{turn}/h t_0)^{-1/3} \tag{4}$$

where r_{100} denotes the radius in units of 100 kpc.

Inserting appropriate values for r and v_{disp} readily shows that clusters of galaxies typically formed at $z_{turn} \lesssim 3$. As Oemler (1974) and others have emphasized, this means that the outer parts of clusters are still in the early stages of their dynamical evolution. When (4) is applied to individual galaxies, whose velocity dispersions are $\lesssim 200$ km/s, it is clear that the inferred t_{turn} is sensitive to the assumed radius: values of z_{turn} in the range 4–8 have been quoted by, for instance, Larson (1974a), who assumes infall from a radius of 50–100 kpc; and similar estimates have been given by Sunyaev (1971) and Sunyaev & Zeldovich (1972); but the extreme form of 'heavy halo' hypothesis (Einasto, Kaasik & Saar 1974; Ostriker, Peebles & Yahil 1974) entails increasing r still further, and this could decrease z_{turn} to $\lesssim 4$. Note also that a further time of $\sim 2t_{turn}$ may elapse before a galaxy settles down into equilibrium.

This means that not only clusters ($v_{disp} \simeq 1000$ km/s), but also individual galaxies

* Even though H_0 is probably closer to 50 than to 100 km/(s Mpc), we follow Peebles (1974b) in parameterizing it in terms of the latter 'rounder' figure.

(especially if they have extensive halos with $v_{\rm disp} \simeq 100\,{\rm km/s}$), may have formed at an epoch sufficiently recent to be potentially accessible to observations. A significant conclusion of Sections 3–5 of the present paper will be that dissipative effects modify (1) in such a way as to permit much *later* formation of systems with a given density. We note at this stage, however, that the foregoing argument is unrealistically simplified, even within the framework of non-dissipative models. For instance:

(i) A more realistic density perturbation would have a smooth profile. The centre (where $\delta\rho/\rho$ is largest) would turn around first, and would have virialized before the outer parts had completed their infall (*cf.* Gunn & Gott 1972; Gott 1973, 1975). The total kinetic energy is never zero, and so the characteristic radius need not contract by ~ 2 between turnaround and virialization.

(ii) A spherical perturbation is in any case an atypical special case. A somewhat more general perturbation is one whose density contours are ellipsoidal – the velocities along the minor axis turn around first, and collapse ensues in one direction while expansion continues in the others. The total kinetic energy is always non-zero, as for the inhomogeneous sphere; and it is possible for collapse to occur in one direction even if the ellipsoid has positive total energy and never becomes gravitationally bound. Generalization of (1)–(4) to ellipsoids where the maximum semi-axes are r_1, r_2 and r_3 is straightforward: in estimates of density, r is replaced by $(r_1 r_2 r_3)^{1/3}$; and in gravitational calculations by $\max(r_1, r_2, r_3)$. (If violent relaxation establishes virial equilibrium, the system may still be ellipsoidal even if the net angular momentum is zero: only the much slower process of two-body relaxation necessarily makes the velocities isotropic and the system spherical (Binney 1976).)

(iii) Dynamical processes occurring during virialization (and subsequently) would redistribute energy between different parts of the system, tending to enhance the central concentration, though this is unimportant unless there is a gradient in the velocity dispersion.

(iv) Even if dissipation processes can be ignored before and during the gravitational clustering process, one would expect a hierarchy of intermingled fluctuations, which would undergo complicated interactions even before 'turnaround' (Fall & Sauverne 1976; and references therein).*

In so far as the 'non-dissipative' approximation applies after $t_{\rm rec}$, the present mass–radius relation for bound systems – or, equivalently, the covariance function $\xi(r)$ – is determined by the spectrum of $(\delta\rho/\rho)$ at $t_{\rm rec}$. Detailed analysis of the spatial distribution of galaxies (Peebles 1974a, and references therein; Totsuji & Kihara 1969) reveal that the angular covariance function is quite smooth and approximately a power-law, exhibiting no preferred scale. This is consistent with the view that at $t_{\rm rec}$ the density irregularities themselves had a power-law spectrum.

$$(\delta\rho/\rho) \propto M^{-n}, \tag{5}$$

the inferred n being in the range $\tfrac{1}{3}$–$\tfrac{2}{3}$ (Peebles 1974a; Gott & Rees 1975). It also suggests that – at least on the scales where $\xi(r)$ is most reliably established – no mass-dependent dissipative process (nor the kind of dynamical evolution mentioned under (iv) above) has grossly affected the observed 'clumpiness' of matter. On the other hand it would perhaps be surprising if gas dynamical processes had *not* influenced the evolution of bound systems between $t_{\rm turn}$ and the present epoch; and any radiative cooling would have increased the binding energy. We consider such effects in the following Sections. We show how these processes operate selectively on certain mass-scales. Peebles (1974a) has conjectured that,

*Proper allowances for (i)–(iv) would probably affect quantitative estimates (e.g. Peebles 1974a; Gott & Rees 1975) of the relation between the present covariance function $\xi(r)$ and the spectrum $(\delta\rho/\rho)_{\rm rec}$ of density fluctuations at recombination.

despite the lack of a preferred scale in the covariance function, 'Galaxies are perceived as a distinct and separate class of entities because gas dynamics assures more or less complete destruction of subclustering on scales up to galaxies'. Our discussion – in which we focus particularly on the effects of radiative cooling – lends quantitative substance to Peebles' remark.

3 Equilibrium and dynamics of gravitating gas clouds

Gas pressure is important on scales $\lesssim$ the Jeans mass M_J, where

$$M_J \simeq (\pi k/G m_p \mu)^{3/2} T^{3/2} \rho^{-1/2} \simeq 10^8 T_4^{3/2} n^{-1/2} M_\odot. \tag{6}$$

The temperature and density are here written in the convenient units of 10^4 K and 1 particle cm^{-3} respectively, and μ is the molecular weight (~0.6 for an ionized 'primordial mixture' of H and He). Even if protogalaxies and protoclusters consist of gas rather than discrete subunits, pressure gradients are dynamically negligible if the masses are $M \gg M_J$. Equivalently, if we define a 'virial temperature' T_{virial} such that

$$T_{\text{virial}} \simeq (GMm_p/kr) \simeq 5 \times 10^5 r_{100}^{-1} M_{12}\ \text{K}, \tag{7}$$

then gas pressure is negligible if the gas temperature is $T \ll T_{\text{virial}}$. As we discuss in Section 4, however, anisotropic or inhomogeneous collapse of an initially cold cloud may generate temperatures $T \simeq T_{\text{virial}}$ before contraction has proceeded very far.

Pressure-supported clouds of mass M will lie close to a line of slope ⅓, with $M \simeq M_J$ (i.e. $T \simeq T_{\text{virial}}$), in the log T – log ρ diagram (Fig. 1). This type of diagram is familiar in studies

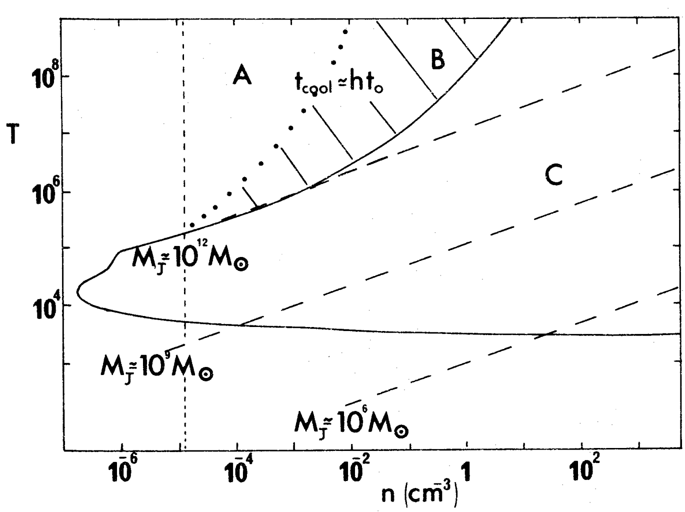

Figure 1. A gas of density n particles cm^{-3} and temperature T has a gravitational timescale $t_{\text{grav}} \propto n^{-1/2}$ and a cooling timescale $t_{\text{cool}} \propto T/n\Lambda(T)$. The solid contour denotes the points in the T–n plane for which $t_{\text{grav}}/t_{\text{cool}}$ is unity. Contours of points where this ratio has constant values different from unity would be the same shape, but shifted horizontally. The gas is assumed to consist of H and He only. Lines of constant Jeans mass M_J have slope ⅓ in the log T–log n plane. A cloud of mass M and density n particles cm^{-3} supported against gravity by gas pressure would need to have a temperature such that it lay close to the appropriate $M \simeq M_J$ line. The *dotted* line corresponds to $t_{\text{cool}} \simeq 10^{10}$ yr (h times the Hubble time, when the Hubble constant is expressed as $100h$ km/(s Mpc)). The vertical dashed line corresponds to 5.5 times the critical density for a cosmological model with $h = 1$: objects to the left of this line have such low densities that they would not have 'turned around' by the present time but would still be expanding with the Universe. A cloud in domain A cannot cool in the Hubble time. In domain B a cloud can be supported quasistatically by gas pressure. However, it will gradually radiate and contract, following a $T \propto n^{1/3}$ trajectory in the diagram until it encounters the critical contour. It then enters domain C, where no pressure support is possible, and goes into isothermal free fall with $T \simeq 10^4$ K. Note that no pressure-supported clouds can exist in the mass range 10^{10}–10^{12} $M_\odot$.

of star formation (e.g. Low & Lynden-Bell 1976). For a realistic inhomogeneous cloud, appropriately averaged values of T and ρ must be taken; and the precise location of the line in the diagram depends on the density profile, degree of rotational support, etc.

A gravitating cloud will lose energy via radiative cooling; and, as it does so, it will trace out a track in the T–ρ plane. The cooling rate per unit volume for a dilute collisionally ionized gas can be expressed in the form $n^2\Lambda(T)$. Detailed computations of $\Lambda(T)$ are reported by Tucker & Gould (1966) and Cox & Tucker (1969). For our present purposes we neglect the contribution from heavy elements. $\Lambda(T)$ then comprises a bremsstrahlung contribution $\sim 10^{-27} T^{1/2} \mathrm{erg\,s^{-1}\,cm^3}$, together with contributions due to H and He recombination, which dominate at $T \lesssim 10^6$ K. The cooling rate $\Lambda(T)$ cuts off sharply at temperatures below $\sim 10^4$ K, when the H would become predominantly neutral.

The cooling timescale can then be defined as

$$t_{\mathrm{cool}} \simeq (3kt/\Lambda(T))\, n^{-1}. \tag{8}$$

The influence of cooling on the fate of a gas cloud depends on the ratio of t_{cool} (evaluated at temperature T_{virial}) to the gravitational free-fall time

$$t_{\mathrm{grav}} \simeq (24\pi G\rho)^{-1/2} \simeq 2 \times 10^7 n^{-1/2}\,\mathrm{yr}. \tag{9}$$

The solid line in Fig. 1 is the contour for which $t_{\mathrm{cool}} \simeq t_{\mathrm{grav}}$: this is a curve such that $n \propto (T/\Lambda(T))^{1/2}$, and all curves for which $(t_{\mathrm{cool}}/t_{\mathrm{grav}})$ is constant are similar in shape but displaced along the n axis. Also shown is the contour for which t_{cool} is equal to the present 'Hubble time' $H_0^{-1} \simeq t_0$. There are thus three relevant domains in the T–ρ diagram:

A. $t_{\mathrm{cool}} \gtrsim t_0$,

B. $t_0 \gtrsim t_{\mathrm{cool}} \gtrsim t_{\mathrm{grav}}$.

and

C. $t_{\mathrm{grav}} \gtrsim t_{\mathrm{cool}}$.

Any pressure-supported cloud that comes into virial equilibrium within domain A would not have had time to radiate its gravitational binding energy, and would thus have remained essentially unchanged thereafter. But if its initial position lay in domain B, it *would* have had enough time to contract significantly. But, so long as this contraction occupies a timescale exceeding t_{grav}, the cloud can re-adjust quasi-statically, its temperature rising so that it remains on the appropriate $T \propto \rho^{1/3}$ 'constant Jeans mass' line. Eventually the trajectory of such a cloud will bring it to the 'critical line'. It then enters domain C: the temperature plunges to 10^4 K; and collapse proceeds on a free-fall timescale, unimpeded by gas pressure.

If the free-fall collapse proceeded in a smooth fashion, the only heat input into the gas would be the 'pdV' work. This is a fraction T/T_{virial} of the rate of conversion of binding into kinetic energy. The temperature would remain $\sim 10^4$ K, and (because the 'pdV' work done per unit mass varies as $\rho^{1/2}$) the level of ionization n_e/n *falls* as $\rho^{-1/4}$. In the more realistic case, the collapse will be irregular and each element of gas will be repeatedly shock-heated. Shocked gas would be able to cool. The radiation rate can then remain comparable with the rate of release of binding energy ($\propto \rho^{5/6}$ per unit mass). Even in this case of 'maximal' dissipation, n_e/n still falls as $\rho^{-1/12}$. In fact this shock heating would generate density inhomogeneities, thereby reducing the required mean ionization level further. Dense shocked regions of gas would be the preferential sites of star formation, and the fraction of material converted into stars cannot participate in any further dissipation. (The stars may *inject* heat into the gas, and we return to discuss this in Section 4.)

Note that only clouds whose masses exceed $\sim 10^{12} M_\odot$ are permitted a period of quasi-static collapse. Below this mass, any cloud which has turned around by the present epoch (i.e. with $\rho > \rho_c$) is in the free-fall domain (*cf.* Lynden-Bell 1966).

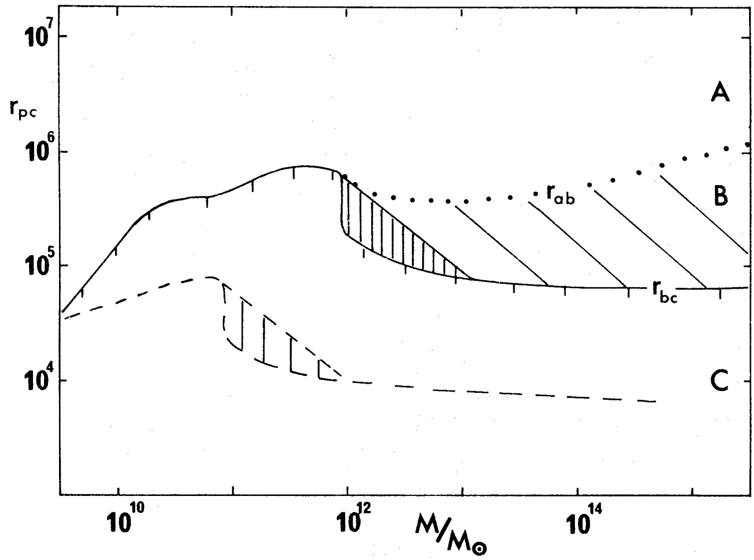

Figure 2. Information essentially similar to that in Fig. 1 is here plotted in terms of cloud mass and cloud radius. The mass-dependent radii demarcating domains A, B and C are denoted by r_{ab} and r_{bc}. Note that at large masses r_{bc} has a constant value $\mathscr{R} \simeq 75$ kpc. When a fraction x of the cloud has been converted into stars, the remaining gas can be quasistatically supported down to a smaller radius, obtained by shifting the previous curve downwards and to the left by a factor $(1-x)^{-1}$. The dashed curve corresponds to the case $x = 0.9$. When T_{virial} is in the range of temperatures where $\Lambda(T)$ decreases with T, a cloud can in fact radiate *more* energy at a *lower* temperature. This could increase r_{bc} for masses 10^{12}–$10^{13} M_{\odot}$ to an extent that is indicated by the region of vertical shading.

In Fig. 2 we plot essentially equivalent results in terms of a radius–mass relation, the 'radius' being naively defined as $(3M/4\pi\rho)^{1/3}$. As was first emphasized by Ostriker (1974), for large masses domains B and C ($t_{\text{grav}} \lesssim t_{\text{cool}}$) are demarcated by a line of *constant radius* $r_{bc} \simeq 75$ kpc. This result applies only to masses for which T_{virial} – equation (7) – is high enough for bremsstrahlung cooling ($\propto T^{1/2}\rho$ per unit mass) to dominate.*

In Section 4.5 we shall express these quantities in terms of 'fundamental' quantities. For lower masses, the critical radius r_{bc} between domains B and C *increases.* Indeed if $\Lambda(T) \propto T^{\alpha}$ then one readily shows that

$$r_{bc} \propto M^{-[(1-2\alpha)/(1+2\alpha)]}. \tag{10}$$

This indicates why the dependence of r_{bc} on M is so sensitive to the change from $\alpha \simeq \frac{1}{2}$ at high temperatures, to $\alpha \simeq -\frac{1}{2}$ in the range 10^5–10^6 K.

No quasi-static support is possible for clouds of 10^{10}–$10^{12} M_{\odot}$. Masses below $10^{10} M_{\odot}$ composed of atomic H and He can remain in equilibrium for a Hubble time if their radii are sufficiently large that T_{virial} is well below $\sim 10^4$ K. Of course, the situation is more complicated when heavy elements are present, or when H_2 can form and provide efficient cooling (*cf.* Yoneyama 1972; Low & Lynden-Bell 1976; Silk 1977). These processes undoubtedly play an important role in star formation. But our present concern is with the early pre-stellar era of galactic evolution, and with masses $\gtrsim 10^{10} M_{\odot}$; so we do not investigate them further here.

At large redshifts one must allow for an extra kind of cooling: Compton scattering of the thermal microwave background radiation, with temperature $\sim 2.7(1+z)$ K. This process gives

* The condition $t_{\text{grav}} = t_{\text{cool}}$ implies that a thermal electron, moving $\sim (m_p/m_e)^{1/2}$ faster than the free-fall speed, radiates its energy in a pathlength $\sim (m_p/m_e)^{1/2} r$. The mass-independence of r_{bc} is then a consequence of the result that the bremsstrahlung cross-section is $\propto v_{\text{electron}}^{-2}$, and $v_{\text{electron}}^2 \propto T_{\text{virial}} \propto M/r$.

a cooling rate (per unit mass) which is independent of ρ and proportional to the electron temperature. The Compton cooling timescale is

$$t_{\text{Comp}} \simeq 3 \times 10^{12}(1+z)^{-4} \text{ yr}. \tag{11}$$

At redshifts exceeding $\sim 10 (\Omega = 1)$ or $\sim 6 (\Omega \ll 1)$, t_{Comp} is less than t_0 – derived from (2) or (3). *All* clouds are then able to cool in Hubble time. If one were to draw versions of Figs 1 and 2 appropriate to these early epochs, then 'domain A' would entirely disappear.

The results displayed in Figs 1 and 2 show that there are elementary physical reasons for expecting mass-dependent dissipative effects during the formation of large gaseous gravitating bodies. Radiative cooling would enhance the binding energy and modify the simple relations outlined in Section 2 between radius, mass and 'turnaround time'. The interesting features of the diagrams relate to *masses* in the range 10^{12}–$10^{14} M_\odot$ characteristic of galaxies and clusters; and the characteristic *radii* of such systems also emerge in a simple way (Ostriker 1974; Silk 1977). This raises a deep suspicion that such physical considerations may indeed by genuinely relevant: they could, after all, have singled out *any* logarithmic interval of masses and radii, so even an order-of-magnitude concordance is suggestive. But before feeling too encouraged one should enquire whether plausible 'scenarios' exist in which the immediate precursors of galaxies and clusters could have resembled the contracting clouds whose behaviour we have been discussing.

4 Scenarios for collapsing protogalaxies

4.1 THE COSMOLOGICAL CONTEXT

At (re)combination matter and radiation become almost completely decoupled, both dynamically and thermally, and the Jeans mass drops to $\sim 10^6 \Omega^{-1/2} h^{-1} M_\odot$. If there were no subsequent heat input at all into the gas, T would fall almost at the adiabatic rate ($\propto \rho^{2/3}$) and M_J would fall with time (i.e. with decreasing z) as $M_J \propto (1+z)^{3/2}$. To maintain n_e/n of order unity at redshifts 10^2–10^3 would require an astrophysically unreasonable heat input, and would distort the microwave background spectrum to an extent incompatible with the data. However, there is no bar to maintaining the gas at a constant temperature of a few thousand degrees with an ionization level of (say) $n_e/n \simeq 10^{-2}$–10^{-3}. The requisite heat input could come from dissipating primordial motions, or from the condensation of a small fraction of the material into stellar or massive objects soon after decoupling. In this (isothermal) case, M_J would increase with time as $M_J \propto (1+z)^{-3/2}$.

The mass range of stable bound clouds which can be in virial equilibrium at $T \lesssim 10^4$ K (and therefore not be subject to cooling) is restricted to the triangular area of the z–M diagram (Fig. 3). This range extends up to only $\lesssim 10^{10} M_\odot$, even at small redshifts. The relative importance of pressure and gravitational effects is measured by the parameter

$$\delta = T/T_{\text{virial}} \tag{12}$$

suggesting that clouds of $\gg 10^{10} M_\odot$ will have $\delta \ll 1$ when they turn around. Will they then get heated to $T \simeq T_{\text{virial}}$ so that they resemble the initial state of the evolutionary tracks described in Section 3?

Suppose that such a cloud did indeed have $\delta \ll 1$ at the moment of turnaround. If it were completely homogeneous, and collapsing spherically symmetrically, it would heat adiabatically ($T \propto r^{-2}$). Unless cooling intervened, the cloud would 'bounce' when it had contracted to $\sim r_{\text{turn}} \delta$. If it entered domain C before this stage, no bounce or quasistatic phase would ever occur. But a realistic cloud will be both inhomogeneous and asymmetric, and

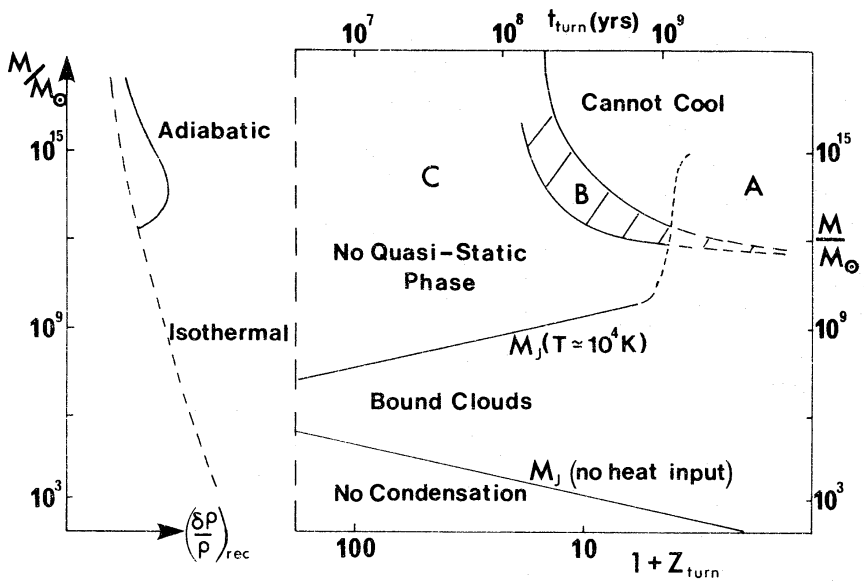

Figure 3. The expected mass spectrum of density perturbations surviving at t_{rec} depends on whether the initial fluctuations were adiabatic or isothermal. The two possibilities are sketched on the left-hand side of the figure. The first masses to turn around and recontract will be those on scales for which $(\delta\rho/\rho)_{rec}$ is largest. Note that in any case we expect t_{turn} to increase with mass on scales $\gg 10^{13} M_\odot$. The horizontal scales on the right-hand part of the figure plot 'turnaround redshift' and 'turnaround time' for a cosmology with $\Omega = 1$ and $h = 1$. If there were no heat input at all after t_{rec}, the Jeans mass would fall as $(1+z)^{3/2}$, so there can definitely be no bound systems formed below this line. If the gas is maintained at $\sim 10^4$ K, then M_J varies as $(1+z)^{-3/2}$. Clouds can form, and remain in virial equilibrium at temperatures below $\sim 10^4$ K, in the triangular area marked. (There is independent evidence that the intergalactic gas was heated to $\gg 10^4$ K at $z \simeq 3$, and the effect of this on M_J is drawn schematically.) The subsequent evolution of massive clouds depends on t_{turn}. (The density at turnaround is $\propto t_{turn}^{-2}$, and $t_{turn} \propto (\delta\rho/\rho)^{-3/2}$.) Clouds that recontract sufficiently early will come into virial equilibrium with $r \simeq \frac{1}{2} r_{turn}$, but will then find themselves in domain B or domain C (*cf.* Figs 1 and 2) and so subsequently fragment. Those in domain A (large masses with late 'turnaround') would remain at hot gas spheres with radii $\sim \frac{1}{2} r_{turn}$. In calculating the curves, Compton cooling has been included at $z \gtrsim 10$.

either of these properties leads to the development of shocks before adiabatic collapse has proceeded very far.

Gull & Northover (1975), Lea (1976) and Cowie (1977, in preparation) have calculated the collapse of a *centrally-condensed* spherical cloud. These authors were concerned primarily with clusters of galaxies, but their results apply to gas spheres in general. A shock forms after a time comparable with the collapse timescale for the central region. If cooling of the shocked gas is slow, the shock propagates out, heating the bulk of the cloud to T_{virial}, and stemming its infall, before it has had time to collapse by a factor ~ 2.

The collapse of an *oblate spheroid* has been treated by Doroshkevich, Sunyaev & Zeldovich (1974) and by Binney (1976), with specific relevance to protoclusters and protogalaxies respectively. As for the inhomogeneous sphere, most of the gas is heated to a temperature T_{virial} (where r in (7) is defined as the *maximum* linear dimension). If $t_{cool} > t_{grav}$, the shock propagates away from the plane of symmetry, leaving the gas almost in virial equilibrium. A small fraction of the material lying close to the plane of symmetry is never shocked to a high temperature, and remains in a dense thin sheet in pressure balance with the rest; but this detail is sensitive to the precise velocity profile during the infall. If t_{cool} is shorter than the dynamical timescale, the shock never propagates out and *all* the gas remains in a thin 'pancake'.

The general collapsing cloud would of course not even be axisymmetric. But we would still expect shock formation to thermalize the energy of infall before r has decreased by more than ~ 2. Thus, even if δ were very small at turnaround, the cloud would have been heated to T_{virial} at a radius of $\sim \frac{1}{2} r_{turn}$, and its subsequent fate will depend on whether it is then in domain A, B or C.*

The redshift at which perturbations of a given mass enter the non-linear regime and turn around depends on the values of $(\delta\rho/\rho)_{rec}$ on that mass scale. (Before turnaround, the behaviour on scales $\gtrsim 10^{10} M_\odot$ is probably adequately approximated by the pressure-free case discussed in Section 2.) There have been many discussions of the processes occurring before t_{rec} which influence $(\delta\rho/\rho)_{rec}$: see Jones (1976) for a comprehensive recent review. Broadly speaking, there are two possibilities: either (i) the perturbations are adiabatic and isentropic, in which case radiative viscosity damps out all scales up to a critical mass of 10^{12}–$10^{14} M_\odot$; or (ii) the initial irregularities involve (at least to some extent) entropy perturbations, and these are not significantly damped. The large-scale homogeneity of the Universe implies that the amplitudes fall off on the largest scales. If the initial spectrum is of power law form, then n – equation (5) – must be positive, and the first scales to turn around will be the smallest undamped ones that exceed the post-recombination Jeans mass: i.e. $\sim 10^{13} M_\odot$ in the purely isentropic case, but $\sim 10^6 M_\odot$ otherwise.

(i) *Isentropic fluctuations.* The minimum surviving mass, calculated first by Silk (1968), depends on Ω and h but is typically $\sim 10^{13} M_\odot$. Doroshkevich *et al.* (1974) suggest that in this case the first objects to collapse should be identified with clusters rather than individual galaxies, the latter being then a secondary phenomenon. (If the covariance function did indeed continue smoothly right down to individual galaxies, then this scheme would perhaps run into difficulties unless some compensating effect could be found (Peebles 1974).) Doroshkevich *et al.* propose that the primordial clusters collapse to form 'pancake' shocks the parameters being such that the bulk of the gas would rebound and end up in domain A or B. The resulting cloud would have a stratified density distribution and be in approximate pressure equilibrium. The parts that would later cool and fragment into galaxies would be those which were squeezed most, or shock-heated up to only a modest temperature. These protogalaxies would be able to cool only if their initial conditions placed them in Domain B, and they would presumably fragment only when they contracted to the critical radius r_{bc}.

(ii) *Isothermal or entropy fluctuations.* Clouds of mass $\sim 10^6 M_\odot$ probably could not turn efficiently into stars because, being loosely bound (escape speed $\lesssim 10$ km/s) a small fraction of material converted into stars could readily release enough energy to unbind the remainder (*cf.* Larson 1974b). Such clouds would therefore either remain gaseous, or most of their material would be expelled and then reconstitute itself into larger clouds. One might then envisage that clouds form into groups by gravitational interaction, and that cloud collisions build up larger masses (*cf.* Gott & Thuan 1976).

Figs 1 and 2 then show why this process should naturally terminate – or at least slow down – when masses of $\gtrsim 10^{12} M_\odot$ are attained. When the masses are lower than this, collisions are inelastic; the shocked gas cools, and can then fragment into stars. But collisions between clouds of $\gtrsim 10^{12} M_\odot$ heat the combined debris to temperatures that can land them in

*Two points of detail: (i) If a cloud is dissipating its binding energy via a succession of weak shocks, such that no gas element is ever heated to T_{virial} but each element is shocked repeatedly in a dynamical timescale, the radiation rate could in principle exceed that possible at T_{virial} if this latter temperature lay near the dip in $\Lambda(T)$ at $\sim 10^6$ K. This possibility would raise the value of r_{bc} for a range of masses around $10^{13} M_\odot$ (see Fig. 2 and caption). (ii) The ionization energy tends to make shocks very inelastic when the velocities are insufficient to generate temperatures exceeding $\sim 10^5$ K. The calculations by Binney (1976) include this effect.

the pressure-supported domain B. Analogous arguments show that a stellar (rather than gaseous) system of $\gtrsim 10^{12} M_\odot$ cannot readily gain mass by accreting gas: even if a gas mass comparable with its own original mass were to fall into its potential well, this gas could be pressure-supported ($t_{\text{cool}} > t_{\text{grav}}$) and need not immediately fragment into stars. Stellar systems whose dimensions placed them in domains A or B of the M–r diagram are likely to form only via coalescence of subunits that are already in stellar rather than gaseous form.

We therefore conclude that, for either of the two extreme hypotheses about $(\delta\rho/\rho)_{\text{rec}}$ corresponding to (i) and (ii) above, *stellar systems of mass M must have radii satisfying*

$$r \lesssim \min[\![[r_{\text{turn}}(M)]/2,\ [r_{\text{bc}}(M)]/2]\!] \tag{13}$$

unless they result from coalescence of pre-formed stellar systems. Approximate equality will prevail in (13) if star formation is reasonably efficient (*cf.* Section 4.2). We also have an inequality involving the 'turnaround' radius

$$r_{\text{turn}} \lesssim 2 r_{\text{ab}}. \tag{14}$$

If this were violated, the gas cloud would come into pressure equilibrium in domain A and would not have time to cool thereafter.

The mean density of an object at its 'turnaround time' scales as t_{turn}^{-2}, so (14) implies a lower limit on the redshift at which clouds must start to contract if they are ever to fragment into stars. The constraints are displayed quantitatively in Fig. 3 (see caption for fuller explanation). In the mass range 10^{13}–$10^{15} M_\odot$, the minimum value of z_{turn} required if a cloud is eventually to fragment increases with M; but if (as one believes) $(\delta\rho/\rho)_{\text{rec}}$ decreases as M increases, then so will z_{turn}. There is therefore an upper limit, depending somewhat on $(\delta\rho/\rho)_{\text{rec}}$, to the masses which can subsequently fragment into systems satisfying (13). Any larger masses which have turned around would still exist as pressure-supported gas clouds with radii $\sim \tfrac{1}{2} r_{\text{turn}}$. One should of course realize that it is an unrealistic oversimplification to envisage clouds of different masses as evolving independently: in reality the smaller masses will be embedded in larger ones.

When galaxies have formed, they provide a heat input into the remaining uncondensed material via supernova ejecta, galactic winds, etc. (*cf.* Schwarz, Ostriker & Yahil 1975). This will inhibit the formation of further large-scale condensations, as depicted schematically in Fig. 3. This reheating, and the consequent increase in the intergalactic pressure, may have a destabilizing effect on gravitationally bound H I clouds of mass $\lesssim 10^9 M_\odot$: if such a cloud were squeezed by external pressure it would behave adiabatically until its internal temperature attained $\sim 10^{40}$ K, and would thereafter be compressed isothermally. Although M_J rises during adiabatic compression it falls again during the isothermal stage, and there is a possibility that the external pressure would initiate gravitational collapse and fragmentation of such clouds. If the external pressure becomes Q times larger than that required to support an isolated cloud at $\sim 10^4$ K, this would trigger the collapse of all bound clouds with mass exceeding $Q^{-3/14}$ times the value M_J ($T = 10^4$) drawn in Fig. 3. Values of Q up to ~ 1000 are plausible. The relevance of this effect to 'young' low-mass galaxies will be discussed elsewhere (Ostriker & Rees 1977, in preparation).

We return to Section 6 to discuss the extent to which radiative dissipation during the pre-stellar phases can modify relations such as (1). Clearly there is some modification due to the decrease in radius during quasi-static contraction in domain B. But there may be significant further dissipation within domain C, which would lead to substantial inequality (rather than approximate equality) in relation (13). This issue is sensitive to the efficiency of star formation, a topic to which we next briefly turn.

4.2 FRAGMENTATION AND STAR FORMATION

A process whereby galaxies might fragment into stars was investigated by Hoyle (1953). He pointed out that a cloud of galactic mass and radius would be collapsing isothermally at $\sim 10^4$ K, essentially for the reasons explained in Section 3. The cloud would be unstable to fragmentation into subclouds of mass M_J. Since M_J decreases ($\propto \rho^{-1/2}$) for isothermal collapse, each subcloud would itself fragment further. Hoyle envisaged that this process of hierarchical fragmentation would terminate only when the smallest units in the hierarchy became optically thick. Isothermality would then break down. Hoyle calculated the masses of the 'non-fragmenting fragments' to be of stellar order. He assumed that the protogalaxy was composed of hydrogen, but one can show (Rees 1976) that this final mass is exceedingly insensitive to the details of opacity and cooling. In its simplest form, this process would be maximally efficient, in the sense that the only energy radiated would be that derived from 'pdV' work, which varies only logarithmically with the final density at which opacity sets in. There is no particular reason why any intermediate scale of the hierarchy would preserve its identity, and the destruction of the subunits would transfer their individual binding energies to the whole galaxy, thereby reducing its radius.

A slightly different scheme has been envisaged by some other authors (Larson 1974a, and references therein). A generic isothermal collapse should generate shocks throughout the clouds. If star formation (or at least the preliminary stages of fragmentation) is triggered by the passage of shocks, then it should proceed at a rate proportional to the collapse timescale:

$$1/\rho_{\rm gas}(d\rho_{\rm gas}/dt) = -C_s t_{\rm grav}^{-1} \tag{15}$$

(Larson 1974a). If C_s were indeed a constant, it would measure the fraction of gas converted into stars each time the collapsing cloud doubled its density. The final radius of the galaxy depends on how much shock dissipation (followed by radiative cooling) can occur after the onset of free-fall collapse. The amount of dissipation is therefore obviously sensitive to how quickly the gas is eaten up into stars: if C_s is $\gg 1$ most of the stars would already have formed by the time the galaxy had collapsed by a factor ~ 2, and we would expect close equality in (13). But it would not be worthwhile attempting actual calculations on the basis of (15). Even if it is valid initially, (15) would stop being so when the gas fraction had fallen well below unity, because the time available for instabilities to develop is $\sim t_{\rm grav} \propto \rho_{\rm total}^{-1/2}$ whereas the growth time for Jeans instability scales as $\rho_{\rm gas}^{-1/2}$: or, equivalently, the instability is inhibited by the tidal influence of the stellar components. Note also that for 'pancake' collapse, the collapse timescale varies as ρ^{-1} (i.e. it eventually becomes $\ll t_{\rm grav}$) and there is again no time for instabilities to develop.

Various kinds of negative feedback – e.g. reheating by supernovae, etc. – would tend to slow down star formation after the first generation (see Section 4.3). Larson (1974a) argues that C_s cannot be too much less than unity – otherwise his models of elliptical galaxies would predict too much central concentration. Other features of galaxies, e.g. the disc-to-halo ratio (Ostriker & Thuan 1975), are sensitive to C_s. But quantitative comparisons with observations are bedevilled by other uncertainties – for instance, the stellar mass function depends on the gaseous environment in a fashion that is not understood at all.

In some scenarios the protogalactic material may always have been at $\lesssim 10^4$ K; but in others (particularly when masses exceed $10^{12} M_\odot$) it may at one stage have been quasi-statically supported at a much higher temperature in domain B. We conjecture that the fragmentation proceeds rather differently in these two cases.

In the former case the process may resemble that proposed by Hoyle (1953), though shock waves may be the main cause of the density enhancements. There is no particular reason why any particular mass scale intermediate between stars and the whole galaxy should be singled out and preserved.

But in the second case, larger masses contracting from domain B would experience sudden cooling. In a realistic situation all parts would not cool simultaneously: if the cloud were isothermal, the denser central regions would cool first; but if the quasi-static cloud resembled an adiabatic sphere, it could be that the lower temperature outer regions (despite their reduced density) cool first and 'rain down' into the centre (Binney 1976). When cooling sets in, M_J drops suddenly by a factor of at least $(10^4/T_{\rm virial})^{3/2}$. Indeed, if pressure equilibrium is maintained during non-uniform cooling ($\rho T \simeq$ constant) the drop is by a factor $(10^4/T_{\rm virial})^2$. This means that, when an object of (say) $10^{13} M_\odot$ contracts through the critical radius $r_{\rm bc}$ and suddenly cools to 10^4 K, all scales in the fragmentation hierarchy between galaxies and globular clusters would be bypassed. Moreover, if pressure equilibrium is maintained, the dynamical timescale for the later stages of fragmentation is shorter by $(10^4/T_{\rm virial})^{1/2}$, implying that the process can go to completion in much less than a crossing time for the whole galaxy. This circumstance would favour the survival of globular-cluster-sized units as discrete entities (see Section 4.4).

4.3 FURTHER EFFECTS OF STAR FORMATION

Once a fraction x of the gas in a protogalaxy has been converted into stars, no simple law of the form (15) is expected to hold, and the subsequent evolution is affected in a number of further ways.

(i) The maximum cooling rate (per unit volume) of the uncondensed gas is reduced by a factor $(1-x)^2$. If the gas were subject to maximal dissipative heating by shocks, the heat input would fall by only one power of $(1-x)$. This permits the gas to remain pressure-supported down to a smaller critical radius (see the subsidiary lines in Fig. 2).

(ii) Contamination of the gas by ejected heavy elements would enhance the cooling rate. The evolution timescale for stars of $\gtrsim 10 M_\odot$ is $\lesssim 10^7$ yr (i.e. $\ll t_{\rm grav}$ at relevant times) implying that some elements are ejected 'instantaneously' once star formation begins.

(iii) The gas will become progressively more inhomogeneous. This enhances the cooling rate, and is likely to be more important than (ii) in so doing.

(iv) The heat input into uncondensed gas from supernova ejecta may exceed that derived from the protogalaxy's gravitational contraction. This would allow the gas to remain hot when $r < r_{\rm bc}$. Moreover the heat input from supernovae, unlike that provided by shocks generated in the irregular collapse, has no built-in thermostat which prevents $T_{\rm gas}$ exceeding $T_{\rm virial}$. Supernovae can therefore *eject* gas from the galaxy. Larson (1974b) has attempted to explore this process quantitatively. Its importance is plainly greater in galaxies of low binding energy (i.e. small $T_{\rm virial}$) but is uncertain because one knows neither what fraction of first-generation stars become supernovae nor the energy each explosion can inject into the gas. This process may well be important enough to reduce final galactic masses far below that of the contracting cloud from which they condensed; and this is yet another uncertainty that renders detailed comparison with observations premature.

4.4 CAN (STELLAR OR GASEOUS) SUBUNITS SURVIVE IN A BOUND SYSTEM?

Consider a virialized gravitationally bound object, radius R, composed of N subunits of individual mass m. It is interesting to consider how long the subunits will survive, in units of the crossing time

$$t_{\rm cross} \simeq (GNm/R^3)^{-1/2}. \tag{16}$$

Obviously

$$t_{\text{survival}}/t_{\text{cross}} \simeq (\pi R^2)/(N\sigma), \tag{17}$$

σ being the cross-section for collisions that result in coalescence or disruption.

Defining a 'generalized geometric cross-section' σ_g equal to πr^2, where r is the mean harmonic radius of each subunit, the answer is

$$t_{\text{survival}}/t_c \simeq N(\sigma_G/\sigma)\left[\frac{\text{binding energy per unit mass of subunits}}{\text{binding energy per unit mass of whole object}}\right]^2 \tag{18}$$

The value of (σ_G/σ) is always $\lesssim 1$ when the quantity in square brackets is $\gtrsim 1$; but when the latter quantity is $\ll 1$, (σ_G/σ) is ~ 1 if the subunits are gaseous but $\gg 1$ if they are themselves clusters of still smaller objects. Without discussing this question in detail we note that the following circumstances both tend to enhance the survival probability:

(i) *A large value of N.* When N is small, ordinary dynamical processes would in any case quickly affect the system, even if σ as defined here were very small.

(ii) *When the ratio in brackets is large.* This would tend to happen if the system had formed from primordial fluctuations where the spectrum was of form (5) with $n > 2/3$, or if radiative dissipation had been more important for the subunits than for the whole system.

4.5 NUMEROLOGICAL DIGRESSION

In Section 3 and the present Section, we have tried to emphasize the significance of a particular radius ~ 75 kpc and a particular mass $\sim 10^{12} M_\odot$. There are reasons why amorphous stellar systems much larger than $10^{12} M_\odot$ should all have this special radius. The significance of $\sim 10^{12} M_\odot$ lies in the fact that somewhat smaller masses cool so efficiently that they can never be supported quasistatically by gas pressure: it may therefore be easier for protogalaxies to grow to this mass than to get beyond it, and this perhaps accounts for galactic masses indeed being of this general order.

If these two quantities are indeed primary determinants of galactic structure, readers of numerological bent may like to see them expressed in terms of fundamental physical constants. Such expressions have in fact been given by Ostriker (1974) and Silk (1976). An expression for r_{bc} in the limit when T_{virial} is so high that bremsstrahlung cooling dominates is the constant radius

$$\mathscr{R} \simeq \left[\left(\frac{e^2}{Gm_p^2}\right)\left(\frac{e^2}{\hbar c}\right)\left(\frac{m_p}{m_e}\right)^{1/2}\right]\left(\frac{e^2}{m_e c^2}\right). \tag{19}$$

The critical mass (at which the r_{bc} curve in Fig. 2 goes vertical) is

$$\mathscr{M} \simeq \left[\left(\frac{Gm_p^2}{\hbar c}\right)^{-2}\left(\frac{e^2}{\hbar c}\right)^5\left(\frac{m_p}{m_e}\right)^{1/2}\right] m_p. \tag{20}$$

These quantities are of course independent of the 'classical' cosmological parameters H_0 and Ω.* They are also, obviously, independent of the parameter $\mathscr{S} = n_{\text{photon}}/n_{\text{baryon}}$, which measures the entropy per baryon of the 'hot big bang'.†

* Note that the mass $\mathscr{M}$ scales with G^{-2}. The characteristic masses of main sequence stars scale as $G^{-3/2}$, and the same is true (Rees 1976) of the minimum mass resulting from hierarchical fragmentation.

† This is true only up to a point. If $\mathscr{S}$ were $\gtrsim 10^3$ times larger than it actually is, Compton cooling would be important enough to raise r_{bc} even at the present epoch; and a large value of $\mathscr{S}$ would suppress gravitational instabilities if the Universe remained radiation-dominated until an epoch when $(1+z) \simeq \Omega^{-1}$.

This last point is pertinent because many attempts have been made to interpret galactic masses in terms of features impressed on the initial fluctuation spectrum by radiative viscosity, etc., before decoupling. The 'Silk mass', below which adiabatic oscillations would have been attenuated (Silk 1968; Weinberg 1971), is indeed of galactic order. However, it is sensitive to $\mathscr{S}$($\propto \mathscr{S}^{5/4}$ or $\propto \mathscr{S}^{1/2}$ according as the expansion is dominated by the mass-energy of matter or of radiation at $t_{\rm rec}$). The characteristic masses derived in the various cosmic turbulence theories, though somewhat different numerically, are all (for analogous reasons) sensitive to $\mathscr{S}$. Thus the present discussion differs from most other theories of galaxy sizes in assigning much less relevance to conditions at very early epochs.

5 Influence of radiation trapping and pressure

5.1 TRAPPING OF RADIATION

Our discussions of cloud collapse and fragmentation have so far ignored opacity and radiation pressure. Accurate inclusion of these effects would be very complicated and model-dependent, but it is fortunately not difficult to show that one is indeed generally justified in ignoring them.

Opacity is potentially significant whenever the optical depth $\tau \gtrsim 1$. The typical photon then takes a time $t_{\rm leak} \simeq \tau(r/c)$ to escape, the radiation energy density and pressure within the cloud being correspondingly enhanced. If $t_{\rm leak}$ exceeds the contraction timescale, which requires an optical depth such that $\tau \gtrsim c/v_{\rm contraction}$ (equivalent to $\tau > t_{\rm grav}\,(c/r)$ when the cloud is in free fall), then all radiation is effectively trapped.

For objects of protogalactic dimensions at $T \gtrsim 10^4$ K, only opacity due to *scattering* can ever be important: even if all the gravitational binding energy were turned into radiation and thermalized, the resulting blackbody temperature (i.e. the temperature of blackbody radiation with an equivalent energy density) would be only

$$T_{\rm equiv} \simeq 0.5\, T_{\rm virial}^{1/4} n^{1/4}\ {\rm K}, \qquad (21)$$

which is always $\ll 10^4$ K in cases of relevance. This means that the presence of trapped radiation can never inhibit cooling of the gas: even the centre of an opaque protogalaxy is a perfect sink for radiation emitted by gas at $\gtrsim 10^4$ K.

5.2 UNLIKELIHOOD OF RADIATION PRESSURE CAUSING 'BOUNCE'

Trapping of the radiation, though it would increase the radiation pressure, would not enhance $p_{\rm gas}$. In the limit of complete trapping and maximal dissipation, then

$$[1 + (p_{\rm rad}/p_{\rm gas})] \simeq T_{\rm virial}/10^4 \qquad (22)$$

and so, when $T_{\rm virial} \gg 10^4$ K, the pressure is mainly exerted by radiation. This means that even though radiation pressure could reverse a 'pancake' collapse, a spherical cloud could never 'bounce' in domain C, even when $t_{\rm leak} \gtrsim t_{\rm grav}$ (*cf.* Fish 1964). This is because the effective γ is then close to $4/3$; or, equivalently, because a given amount of energy exerts only half as much pressure in the form of radiation as in the form of non-relativistic thermal particle motions. (This argument would not prevent a bounce occurring – given a sufficient τ – if extra radiation were supplied by some agent (e.g. supernovae) which augmented that derived from the gravitational contraction of the whole protogalactic cloud.)

5.3 NATURE OF POSSIBLE OPACITY

The two main relevant types of opacity are (a) Thomson scattering by free electrons, and (b) scattering of trapped Lyman α.

The Thomson scattering optical depth is only

$$\tau_{es} \simeq 10^{-3}(n_e/n) M_{12} r_{100}^{-2}. \tag{23}$$

Even for complete ionization, τ_{es} is therefore usually $\lesssim 1$ (and certainly is unlikely to be so large that $t_{trap} \gtrsim t_{grav}$). Moreover, once the collapse enters domain C, the ionized fraction n_e/n will fall at least as fast as $r^{1/4}$.

When the gas does cool to $\sim 10^4$ K and start to recombine, the optical depth in Lyman α becomes enormous. On the assumption that bulk motions with the free-fall speed determine the line width, the optical depth in the line centre is

$$\tau_{Ly\alpha} \simeq 10^7(n_H/n) M_{12}^{1/2} r_{100}^{-3/2}. \tag{24}$$

Furthermore, most of the radiation will then be emitted as (or quickly converted into) Lyman α: the Lyman continuum would be opaque, so all hydrogen recombinations, even those directly to the ground state, would generate a Lyman α photon; and so indeed would any helium recombinations. Lyman α also results from collisional excitation, a process that contributes importantly to the cooling at $\sim 10^4$ K.

5.4 DOES RADIATION TRAPPING AFFECT FRAGMENTATION?

We have seen that, even when trapping is efficient, the overall collapse proceeds unimpeded. One's first thought would be that the modified fragmentation criterion could be simply obtained by multiplying M_J – equation (6) – by $[1 + (p_{rad}/p_{gas})]^{3/2}$. But this would be correct only if the photons are so well trapped that they cannot diffuse a Jeans length in a time t_{grav}; which is a more stringent condition than the requirement that the photons should be trapped for $\gtrsim t_{grav}$ within the whole cloud.

To describe the case when radiation is not trapped to this complete extent, let us define $t_{leak}(M)$ to be the time a typical photon takes to diffuse a distance $\sim (M/\rho)^{1/3}$. When $t_{leak} \ll t_{grav}$, the radiation pressure gradient that builds up to oppose gravitational instability on a mass-scale M would be only $q(M) \simeq [t_{leak}(M)]/t_{grav}$ times that expected for perfect trapping. The corrected Jeans mass is then given by

$$M_J^* = M_J[1 + q(M_J^*)(p_{rad}/p_{gas})]^{3/2} \tag{25}$$

M_J being the mass given by (6). Note that M_J appears implicitly on the right-hand side. For electron scattering opacity, $q \propto M^{2/3}$, which means that the relative importance of radiation pressure gradients and gravitational forces is independent of M, on all scales M such that $t_{leak}(M) < t_{grav}$. Therefore, radiation trapped within a cloud by electron scattering has an 'all or nothing' effect on the Jeans mass: if it is unable to impede fragmentation into masses such that $t_{leak}(M) \simeq t_{grav}$, it has no effect on any smaller scales either.*

*If a protogalaxy collapses at a large redshift, drag and pressure gradients due to trapped primordial background radiation may be more important than effects due to the cloud's own emission. When $\tau \gg 1$, the condition for the former to dominate is that the background temperature should exceed T_{equiv} – equation (21). Even when $\tau \lesssim 1$, the primordial radiation provides a drag force which tries to keep the matter and radiation comoving, thereby impeding the growth of $(\delta\rho/\rho)$. This drag operates on a timescale $t_{drag} \simeq (m_p/m_e)(n_e/n)^{-1} t_{Comp}$, where t_{Comp} is given by (11). The effect could be important at $z \gtrsim 100$ (Peebles 1967).

In the more important case when radiation is trapped in the cloud by Lyman line opacity, the value of $t_{leak}(M)$ that one would compute by envisaging a random walk in space would be $\gg t_{grav}$ on all scales M_J, suggesting complete trapping. However, a particular photon diffuses in energy as well as space, and the trapping problem is not straightforward (Hummer & Rybicki 1972, and references therein; Adams 1972). If there are no bulk motions, t_{leak} is 10–100 crossing times for all optical depths 10^4–10^8. Any large-scale velocity gradients would reduce t_{leak}: in particular, if the cloud is systematically contracting, each scattering will tend to blueshift the photons, leading to an upper limit $\sim t_{grav}(\Delta\nu/\nu)$ on the trapping time. In this expression $(\Delta\nu)$ should be taken as the overall line widths, which may be up to ~ 50 Doppler widths (Adams 1972). implying line widths of ~ 100 Å in protogalaxies.

Because of the effects of the frequency shifts, the trapping efficiency factor q is likely to depend only very slightly on M – it certainly will not vary as $M^{2/3}$, as does the corresponding factor for electron scattering. It is therefore possible that, even though Lyman α can never brake the overall collapse, it may inhibit fragmentation, raising M_J by as much as $\sim q^{3/2}(T_{virial}/10^4)^{3/2}$. If this were so, higher densities would be needed before Hoyle's (1953) hierarchical fragmentation could proceed down to a given mass.

6 Concluding comments

The main aim of this paper has been to discuss basic physical processes which play a dominant role in determining the masses and radii of galaxies, and the epoch at which they form. Our arguments have been of a fairly general character, so it might be useful to conclude by emphasizing their relevance to four specific aspects of galaxies and clusters.

6.1 THE EPOCH OF GALAXY FORMATION

The possibility that massive clouds can undergo a prolonged phase of quasistatic contraction before free fall (and fragmentation into stars) ensues obviously modifies the relation expressed in (1) between *present* radius r and turnaround time. For masses $10^{12} M_\odot$ we have an upper limit to t_{turn} because of the inequality (14): if 'turnaround' occurs too late, the resulting gas cloud is in domain A and never cools (see Fig. 3). However, if we mean by 'galaxy formation' the stage when fragmentation occurs, then we have no lower limit at all to the redshift at which it happens. If the protogalaxy starts off with a radius just below r_{ab}, it will take 10^{10}yr to contract to r_{bc}. Only then can it turn into stars, and the final galactic radius would be r_{bc} (relation (13) even if its radius at 'turnaround' were much larger. It is of course gratifying that r_{bc} is indeed comparable with the radii of massive galaxies.

This circumstance adds interest and optimism to searches for young galaxies. The likely properties of such objects have recently been discussed by Meier (1976) and Kaufman (1976). A collapsing protogalaxy would contain much more uncondensed gas than present-day galaxies. Shock waves generated during the collapse would have formed the gas into sheets or filaments; and such objects, lying along the line of sight to distant quasars, could cause systems of absorption lines of the type seen in quasar spectra.

6.2 TIDAL TORQUES AND GALACTIC ROTATION

If protogalaxies have spent a long time collapsing quasistatically from an initial radius r_{turn}, substantially larger than their present radius, tidal torques between near neighbours could clearly operate with enhanced effectiveness.

6.3 THE COVARIANCE FUNCTION, ETC.

If dissipative processes had been unimportant on any scales, the present clustering properties of matter would be a straightforward 'mapping' of the initial fluctuation spectrum $(\delta\rho/\rho)_{rec}$. But material that has at some stage constituted a massive gas cloud *will* have experienced radiative cooling; and moreover this effect operates preferentially on certain scales. An illustrative and perhaps realistic example is displayed in Fig. 4 (see caption for further explanation). Radiative effects enhance the binding energy by a factor $\beta \simeq r_{turn}/r_{bc}$ (and the factor could be still larger if star formation were inefficient and the final radii were much less than $\sim \frac{1}{2} r_{bc}$). This energy would manifest itself as a contribution to the X-ray background. In the usual covariance function jargon, the 'non-dissipative' $\xi(r)$ is modified so that $\xi'(r) = \beta^3 \xi(\beta r)$. As is clear from Fig. 4, this effect depends quantitatively on the adopted form of $(\delta\rho/\rho)_{rec}$, so we cannot make specific predictions (for similar reasons we have not ventured any discussion of the galactic luminosity function, etc.). There would, however, be a general tendency for the dissipation to suddenly cut off as the mass scale increased above some value $\sim 10^{14} M_\odot$. This feature would enhance the probability that sub-units could survive within clusters of total mass $\gtrsim 10^{14} M_\odot$ (*cf.* Section 4.4).

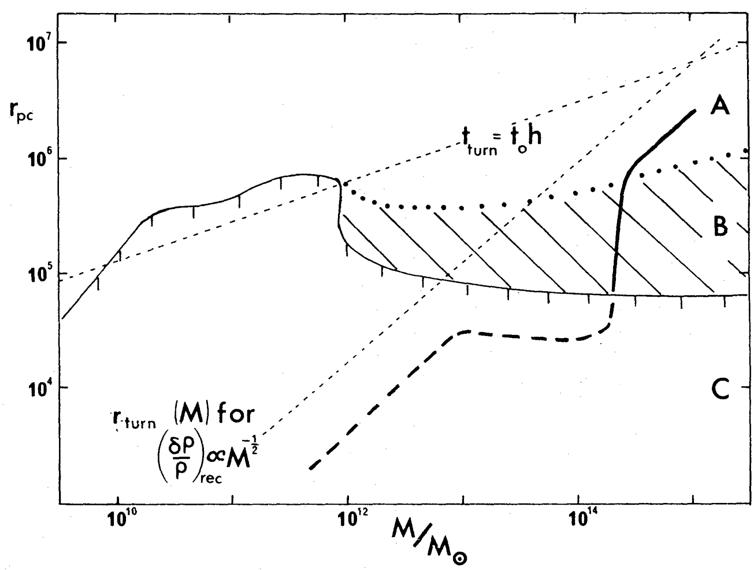

Figure 4. The continuous curve is $r_{bc}(M)$, as in Fig. 2. The upper dashed line, slope $\frac{1}{3}$, denotes the mass–radius relation for clouds with 5.5 times the critical density. Such clouds would be turning round at the present epoch. If the spectrum $(\delta\rho/\rho)_{rec}$ had larger amplitude on smaller scales, then smaller masses would have turned around at earlier times. The lower dashed line shows $r_{turn}(M)$ for the case $(\delta\rho/\rho)_{rec} \propto M^{-n}$ with $n = \frac{1}{2}$. This is an arbitrary choice (though n is thought to lie in the range $\frac{1}{3}$–$\frac{2}{3}$), but the height is normalized so as to agree with covariance function data on scales 10^{14}–$10^{15} M_\odot$. All mass scales would come into virial equilibrium with $r_{virial} \simeq \frac{1}{2} r_{turn}$. But clouds which were then located in domains B or C would cool radiatively, contract, and fragment, the final radii obeying (13). The final mass–radius relation would resemble the heavy line (scales that would have fragmented are indicated by the dashed portion). Note the extent to which dissipation has increased the binding energy on scales 10^{13}–$10^{14} M_\odot$, thereby modifying the quantitative details of the 'non-dispersive' model discussed in Section 2.

6.4 CLUSTERS OF GALAXIES

It is apparent from Figs 3 and 4 that clouds of mass $10^{14} M_\odot$ which turned around at redshifts $\lesssim 3$ could have remained essentially unchanged for 10^{10} yr. If such objects exist, they would be detectable only as thermal X-ray sources. If such a cloud *did* have time to cool, as would be the case if z_{turn} were somewhat larger (see Fig. 3), then at least part of its mass must have cooled and fragmented. It is perhaps suggestive that the cooling time of the hot gas in (e.g.) the Coma Cluster is $(1-2)\times10^{10}$ yr. Perhaps the protocluster had a shorter cooling time (i.e. lay in domain B) and fragmentation into galaxies proceeded until the gas was so depleted that the remainder could remain hot for $\gtrsim 10^{10}$ yr.

Acknowledgments

We have derived pleasure and profit from discussing the ideas in this paper with many colleagues. We are particularly grateful to James Binney, Bernard Jones, Donald Lynden-Bell, Jim Peebles, Joe Silk, Beatrice Tinsley and Scott Tremaine for help and criticism.

References

Adams, T. F., 1971. *Astr. Astrophys.*, **12**, 280.
Binney, J. J., 1976. *D Phil thesis,* Oxford (and work in preparation).
Cox, D. P. & Tucker, W. H., 1969. *Astrophys. J.*, **157**, 1157.
Doroshkevich, A. G., Sunyaev, R. A. & Zeldovich, Y. B., 1974. *Confrontation of cosmological theories and observational data,* p. 213, ed. M. S. Longair, Reidel, Holland.
Einasto, J., Kaasik, A. & Saar, E., 1974. *Nature,* **250**, 309.
Fall, S. M. & Sauverne, G., 1976. *Mon. Not. R. astr. Soc.*, **174**, 241.
Fish, R. A., 1964. *Astrophys. J.*, **139**, 284.
Gott, J. R., 1973. *Astrophys. J.*, **186**, 481.
Gott, J. R., 1975. *Astrophys. J.*, **201**, 296.
Gott, G. R. & Rees, M. J., 1975. *Astr. Astrophys.*, **45**, 365.
Gott, J. R. & Thuan, T. X., 1976. *Astrophys. J.*, **204**, 649.
Gull, S. L. & Northover, K. V. E., 1975. *Mon. Not. R. astr. Soc.*, **173**, 585.
Gunn, J. E. & Gott, J. R., 1972. *Astrophys. J.*, **176, 1.**
Hoyle, F., 1953. *Astrophys. J.*, **118**, 513.
Hummer, D. G. & Rybicki, G., 1971. *A. Rev. Astr. Astrophys.*, **9**, 237.
Jones, B. J. T., 1976. *Rev. mod. Phys.*, **48**, 107.
Kaufman, M., 1976. *Astrophys. Sp. Sci.*, **40**, 369.
Larson, R. B., 1974a. *Mon. Not. R. astr. Soc.*, **166**, 585.
Larson, R. B., 1974b. *Mon. Not. R. astr. Soc.*, **169**, 229.
Lea, S. M., 1976. *Astrophys. J.*, **203**, 569.
Low, C. & Lynden-Bell, D., 1976. *Mon. Not. R. astr. Soc.*, **176**, 367.
Lynden-Bell, D., 1966. *Radio astronomy of the Galaxy,* ed. H. von Woerden, Reidel, Holland.
Meier, D. L., 1976. *Astrophys. J. Lett.*, **203**, L103.
Oemler, A., 1974. *Astrophys. J.*, **194**, 1.
Ostriker, J. P., 1974. *Talk at 7th Texas Conference.*
Ostriker, J. P., Peebles, P. J. E. & Yahil, A., 1974. *Astrophys. J. Lett.*, **193**, L1.
Ostriker, J. P. & Thuan, T. X., 1975. *Astrophys. J.*, **202**, 353.
Peebles, P. J. E., 1967. *Astrophys. J.*, **147**, 859.
Peebles, P. J. E., 1974a. *Astrophys. J. Lett.*, **189**, L51.
Peebles, P. J. E., 1974b. *Origin and evolution of galaxies,* ed. J. R. Shakeshaft, Reidel, Holland.
Rees, M. J., 1976. *Mon. Not. R. astr. Soc.*, **176**, 483.
Schwarz, J., Ostriker, J. P. & Yahil, A., 1975. *Astrophys. J.*, **202**, 1.
Silk, J. I., 1968. *Astrophys. J.*, **151**, 459.
Silk, J. I., 1977. *Astrophys. J.*, **211**, 638.
Sunyaev, R. A., 1971. *Astr. Astrophys.*, **12**, 190.
Sunyaev, R. A. & Zeldovich, Y. B., 1972. *Astr. Astrophys.*, **20**, 189.
Totsuji, M. & Kihara, T., 1969. *Publ. astr. Soc. Japan,* **21**, 211.
Tucker, W. H. & Gould, R. J., 1966. *Astrophys. J.*, **144**, 244.
Weinberg, S., 1971. *Astrophys. J.*, **168**, 175.
Yoneyama, T., 1972. *Publ. astr. Soc. Japan,* **24**, 87.