Mon. Not. R. astr. Soc. (1978) **183**, 341–358

Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering

S. D. M. White and M. J. Rees *Institute of Astronomy, Madingley Road, Cambridge*

Received 1977 September 26

Summary. We suggest that most of the material in the Universe condensed at an early epoch into small 'dark' objects. Irrespective of their nature, these objects must subsequently have undergone hierarchical clustering, whose present scale we infer from the large-scale distribution of galaxies. As each stage of the hierarchy forms and collapses, relaxation effects wipe out its substructure, leading to a self-similar distribution of bound masses of the type discussed by Press & Schechter. The entire luminous content of galaxies, however, results from the cooling and fragmentation of residual gas within the transient potential wells provided by the dark matter. Every galaxy thus forms as a concentrated luminous core embedded in an extensive dark halo. The observed sizes of galaxies and their survival through later stages of the hierarchy seem inexplicable without invoking substantial dissipation; this dissipation allows the galaxies to become sufficiently concentrated to survive the disruption of their halos in groups and clusters of galaxies. We propose a specific model in which $\Omega \simeq 0.2$, the dark matter makes up 80 per cent of the total mass, and half the residual gas has been converted into luminous galaxies by the present time. This model is consistent with the inferred proportions of dark matter, luminous matter and gas in rich clusters, with the observed luminosity density of the Universe and with the observed radii of galaxies; further, it predicts the characteristic luminosities of bright galaxies and can give a luminosity function of the observed shape.

1 Introduction

A central issue in theories of galaxy formation is the relative importance of purely gravitational processes (*N*-body effects, clustering, etc.) and of gas-dynamical effects involving dissipation and radiative cooling. The large-scale distribution of galaxies is consistent with a smooth 'hierarchical clustering' picture lacking any preferred scale, and with the results of *N*-body simulations where no non-gravitational effects are included. On the other hand, dissipation must have played a role in the formation of disc galaxies (and perhaps of the luminous central parts of elliptical galaxies as well); uncondensed gas exists in clusters of galaxies and in individual galaxies; and the characteristic mass and size of galaxies finds no natural interpretation in a purely gravitational picture. We shall argue, in fact, that a purely dissipationless clustering process cannot be responsible for forming both individual galaxies and galaxy clusters.

We here develop a model which incorporates aspects of both these schemes: the distribution of the dominant mass component on all scales arises from purely gravitational clustering (Press & Schechter 1974); but the observed sizes and luminosity functions of galaxies are determined by gas-dynamical dissipative processes, for reasons which are a straightforward extension of arguments given by Binney (1977), Rees & Ostriker (1977) and Silk (1977).

A satisfactory theory of galaxy formation must account for the large amount of non-gaseous 'dark matter' which apparently provides $\gtrsim 80$ per cent of the virial mass in clusters like Coma and which may constitute massive halos around large galaxies. Indeed, the evidence is consistent with the view that all systems $\gtrsim 100$ kpc in size contain dark matter and luminous matter (i.e. visible stars and gas) in a universal ratio of (5–10):1. We shall not attempt here to assess the various arguments bearing on this question, but we do take the point of view that the discrepancy between the well-determined mass to light ratios found for rich galaxy clusters and for the main body of galaxies suggests that $\gtrsim 80$ per cent of the gravitating matter in the Universe may well be in some dark form that is still undetected except indirectly through its gravitational effects.

This segregation of high- and low-luminosity material seems incompatible with any theory which tries to build up galaxies and clusters from smaller units in an entirely dissipationless way, since one expects efficient mixing to occur during this process. Further, the luminous contents of galaxies appears too concentrated to be on a continuous hierarchy embracing clusters of galaxies (*cf.* Section 4). Finally the survival of galaxies as discrete subcondensations within clusters of short crossing time is inconsistent with dissipationless clustering, in which substructure is destroyed during the collapse of any bound unit. Numerical simulations of this process show neither stable clusters of clusters, nor structures with many subcondensations (*cf.* Section 2 and Appendix).

The need to invoke some dissipative process in galaxy formation is also suggested by the fact that the characteristic mass of a large galaxy is not the same as the 'turn-around' mass scale in a gravitational clustering model. Press & Schechter (1974) appreciated this problem, and postulated – *ad hoc* – that galaxies all formed at some well-defined past epoch by an unspecified process, the galactic luminosity function then being a scaled-down version of the cluster luminosity function. There is, of course, obvious evidence for gaseous dissipation in disc systems.

What does the dark mass consist of? Of the many possibilities, the most plausible candidates are low-mass stars, burnt-out remnants of high-mass stars, or the remnants of supermassive stars, any of which might have formed soon after the primordial plasma recombined (i.e. $z \simeq 1000$, $t \simeq 10^{13}\Omega^{-1/2}h^{-1}$ s the Hubble constant being taken as $100h$ km/(s Mpc)). Indeed, the conditions at this epoch (density $\sim 10^4 \Omega h^2$ particles cm^{-3}, Jeans mass $\sim 10^6 \Omega^{-1/2} h^{-1} M_\odot$) would seem much *more* propitious for gravitational instability and fragmentation than the present-day environments where star formation is usually assumed to occur. A plausible extrapolation to smaller mass scales of the fluctuation spectrum needed (in any theory) to explain galaxies and clusters suggests that the inhomogeneities on scales $\sim 10^6 M_\odot$ may already be of order unity, so that massive clouds would collapse and fragment immediately after decoupling from the background radiation field. Only in the special case of purely adiabatic fluctuations, which are attenuated by radiative viscosity on all scales $\lesssim 10^{13} M_\odot$, could this early fragmentation be suppressed. There are some processes of negative feedback which could limit the rate at which this fragmentation proceeded (e.g.

Compton drag following reheating); but these processes cannot readily have prevented most of the gas from condensing into bound objects by $z \simeq 100$. One cannot confidently say what mass range these objects would lie in (Rees 1977), but for the present discussion we need merely postulate that the dark mass behaves as an assemblage of point particles of individual mass $\lesssim 10^6 M_\odot$. In fact nothing in our discussion would change if the dark mass consisted of, for instance, massive neutrinos, or black holes which formed before recombination.

At redshifts $z \gtrsim 100$, the dark mass could not have been clustered on mass scales as large as galaxies. However, we shall take the point of view that it subsequently underwent hierarchical gravitational clustering (in the manner that can be simulated by N-body computations) and formed progressively larger systems; its present distribution can be inferred from the observed galaxy distribution and it constitutes the dark mass in clusters of galaxies and in galactic halos. We suggest, however, that the entire observed stellar content of galaxies condensed from residual gas whose distribution followed that of the dark material until cooling allowed it to settle within the gravitational potential wells of pre-existing 'halos'. A prerequisite for the occurrence of 'secondary' star formation is that the gas should have time to cool radiatively and fragment. This requirement sets a characteristic upper limit to galactic masses and sizes. When the clustering properties of the dark matter are matched to the observed galaxy covariance function, the inferred dimensions accord gratifyingly with the observations. We can, furthermore, make a preliminary attempt to derive a galactic luminosity function, on the assumption that low-mass galaxies condensed at early times when the hierarchical clustering of the dark matter did not extend to large masses and then retained their identity throughout the subsequent growth of the hierarchy. A consistent model can be obtained even on the restrictive assumption that all secondary star formation occurs with the same initial mass function (IMF). This contrasts with the situation in alternative theories: as noted above, dissipationless theories cannot explain why M/L is systematically lower in high-density regions; a purely gas-dynamical model would require an IMF dependency on density to account for this observation. The time-evolution of our model is illustrated in Fig. 1.

Within our picture, there is no process that can segregate the three components (dark material, luminous material and gas) on scales as large as a rich cluster. For consistency, we therefore require that the universal M/L be the same as that measured in clusters such as Coma ($M/L \simeq 400\,h$). This implies that $\Omega \simeq 0.2$. Assuming that the luminous material has $M/L \simeq 40\,h$, a typical value for the main body of elliptical galaxies, ~ 80 per cent of cosmic matter must have condensed at an early stage to form the dark mass; of the remainder, half is still uncondensed and the other half constitutes the luminous content of galaxies.

In Section 2 we first review the theory of 'self-similar' gravitational clustering in an expanding universe, which provides a model for the behaviour of the dark mass. The clustering builds up in a hierarchical fashion; the smaller scale virialized systems quickly merging into an amorphous whole when they are incorporated in a larger bound cluster. A detailed discussion of the disruption process is given in the Appendix. The fate of the residual gas within these transient potential wells is discussed in Section 3. Luminous galaxies up to a certain limiting size can form when this gas cools and becomes sufficiently concentrated in the centre of the potential well to be self-gravitating and liable to fragmentation. When a halo is disrupted in a larger system the luminous galaxy in its core can preserve its identity because dissipation has made it more concentrated than the surrounding dark material. Finally (Section 4) we present specific models for the evolution of the observed system of galaxies and clusters and for the galactic luminosity function, and we discuss the general implications of the proposed scenario.

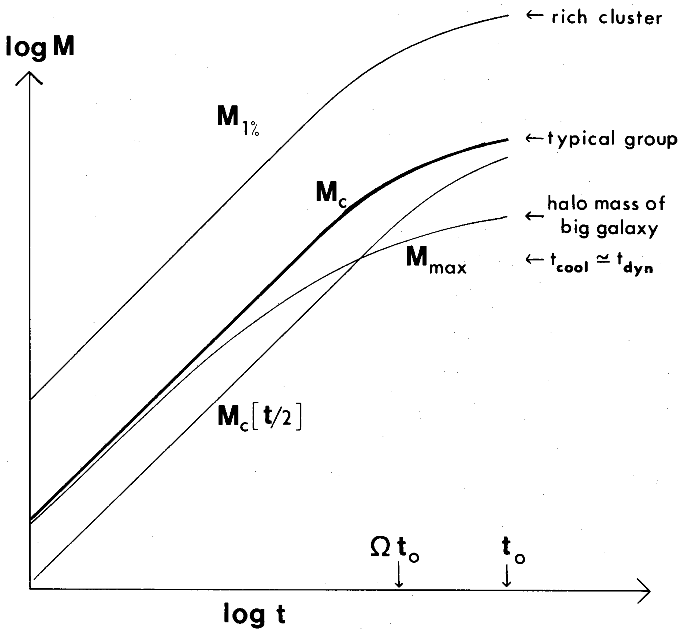

Figure 1. This diagram shows in a schematic way how 'dark' matter, assumed to have condensed into (stellar mass?) objects by, or soon after, the recombination time $t_{\rm rec}$, becomes clustered on progressively larger scales. A universe with present age t_0 and $\Omega < 1$ is assumed. The heavy line shows the typical mass scale $M_c(t)$ for which $t = t_{\rm dyn}$, where $t_{\rm dyn}$ is twice the turn-around time. If the inhomogeneities present at $t_{\rm rec}$ have amplitudes $\propto M^{-\alpha}$, then $M_c \propto t^{2/3\alpha} \propto (1+z)^{-1/\alpha}$ until $t \simeq \Omega t_0$, but the mass scale of clustering tends to grow more slowly thereafter (see (2.7)–(2.11) in text). Most bound units with $t_{\rm dyn} \lesssim t/2$ will already have been incorporated into this scale of the hierarchy. There will be a spread in mass scales with a given overdensity (i.e. a given turn-around time); the line $M_{1\ \rm per\,cent}(t)$, displaced *vertically* with respect to $M_c(t)$ by an amount that depends on α, indicates the mass such that 1 per cent of material at time t is bound in collapsed units $\geqslant M_{1\ \rm per\,cent}$. The curves can be normalized so that $M_c(t_0)$ and $M_{1\ \rm per\,cent}(t_0)$ fit the observed galaxy distribution. At any stage of the hierarchy, residual gas can condense in these halos to form luminous galaxies provided that it can cool before the halos are disrupted by merging into larger systems (see Section 3). The requirement $(t_{\rm cool} + t_{\rm dyn}) \leqslant t$ sets a characteristic upper limit $M_{\rm max}(t)$ to the halo masses in which galaxies can have condensed by time t.

2 Dynamical aspects

2.1 GROWTH AND 'TURN-AROUND' OF DENSITY PERTURBATIONS: GRAVITATIONAL CLUSTERING

A viewpoint which is commonly adopted is that the present inhomogeneities in the Universe developed from initial density perturbations with a power-law spectrum such that, at some early time t_i (e.g. the recombination epoch $t_{\rm rec}$, at a redshift $z \simeq 1000$) the characteristic amplitude on relevant mass scales M was of the form

$$\left\langle\left(\frac{\rho_i - \bar{\rho}_i}{\bar{\rho}_i}\right)^2\right\rangle^{1/2} = \left\langle\left(\frac{\delta\rho}{\rho}\right)^2\right\rangle_i^{1/2} \propto M^{-\alpha} \tag{2.1}$$

with $\alpha > 0$. If pressure effects are unimportant, then the density perturbation on each mass scale grows as $(\delta\rho/\rho) \propto (1+z)^{-1}$ until the amplitude becomes non-linear and the irregularity separates out into a non-expanding bound system (or until $(1+z) \lesssim \Omega^{-1}$, when linear growth terminates).

For a strictly spherical perturbation, there is a simple expression relating the turn-around time $t_{\rm turn}$ to the initial density perturbation (Sunyaev 1971; Gunn & Gott 1972). Provided that the initial redshift z_i satisfies $(1+z_i) \gg \Omega^{-1}$, then the Hubble constant H_i at the initial epoch is related to H_0 by

$$H_i^2 = \Omega H_0^2 (1+z_i)^3. \tag{2.2}$$

Defining ρ_{ci} to be $3H_i^2/8\pi G$, one finds

$$\frac{\rho_{ci} - \bar{\rho}_i}{\rho_{ci}} = \frac{1-\Omega}{\Omega(1+z_i)}. \tag{2.3}$$

Clearly, only perturbations with overdensities larger than this can ever become bound systems. For such perturbations, the turn-around time is related to the initial density ρ_i by

$$t_{\text{turn}} = \frac{\pi}{2H_i}\left(\frac{\rho_{ci}}{\rho_i - \rho_{ci}}\right)^{3/2}. \tag{2.4}$$

For any law of the form (2.1) there will be a mass scale m_i on which the initial fluctuations are of order unity; and so long as pressure effects are unimportant, the growth of density perturbations on scales $\gg m_i$ will proceed in an analogous fashion to the gravitational clustering of a set of point masses m_i which are initially expanding with the mean Hubble flow. The gravitational clustering of point masses was considered in an interesting paper by Press & Schechter (1974) and much of the present discussion is based on their work.

Laws of the form (2.1) relate the typical overdensity relative to the *mean* density to the mass scale; however, it is the overdensity relative to the *critical* density which determines the turn-around time and the final density. For $\Omega < 1$, this distinction is important for mass scales which turn around when $(1+z) \lesssim \Omega^{-1}$. From (2.2)–(2.4) we find that, for scales just turning around now (i.e. $t_{\text{turn}} = t_0$),

$$\frac{\rho_i - \bar{\rho}}{\rho_i - \rho_{ci}} = 1 + (\Omega^{-1} - 1)\left(\frac{2H_0 t_0 \Omega^{1/2}}{\pi}\right)^{2/3}. \tag{2.5}$$

$H_0 t_0$ is a slowly-varying function of Ω (see, e.g. Gott *et al.* (1974)); for our chosen value $\Omega = 0.2$,

$$\left(\frac{\rho_i - \bar{\rho}}{\rho_i - \rho_{ci}}\right) = 2.55. \tag{2.6}$$

Assuming a self-similar clustering process and a law of the form (2.1) the physical characteristics of condensations depend on mass M according to the relations

$$t_{\text{turn}}/t_0 = (M/M_0)^{3\alpha/2} x^{-3/2} \tag{2.7}$$

$$\rho/\rho_0 = (M/M_0)^{-3\alpha} x^3 \tag{2.8}$$

$$r/r_0 = (M/M_0)^{1/3+\alpha} x^{-1} \tag{2.9}$$

$$\overline{v^2}/\overline{v_0^2} = T/T_0 = (M/M_0)^{2/3-\alpha} x \tag{2.10}$$

where, for a universe with $\Omega = 0.2$

$$x = [2.55 - 1.55(M/M_0)^{\alpha}]. \tag{2.11}$$

In these formulae, the suffix zero denotes the values appropriate to a mass M_0 for which $t_{\text{turn}} = t_0$. In (2.10) $\overline{v^2}$ is the virial velocity dispersion and T the corresponding temperature.

Statements such as (2.7)–(2.10) apply in an 'average' sense only. In fact there will be a spectrum of initial overdensities associated with regions of a given mass M; and therefore a spread in turn-around times for a given mass. A straightforward extension of the methods of Press & Schechter (1974) leads to the following expression for the mass spectrum of discrete bound systems at any given time (*cf.* Gott & Turner 1977)

$$N(M)\,dM \propto M^{\alpha-2} \exp\,[-(M/M_c)^{2\alpha}]\,dM \tag{2.12}$$

where M_c is a typical mass turning around at time t. At times corresponding to $(1+z) > \Omega^{-1}$, M_c increases as $t^{2/3\alpha}$, the basic point masses becoming incorporated in progressively larger units. In a low density ($\Omega < 1$) universe, the clustering process 'freezes' when $(1+z) \lesssim \Omega^{-1}$ and M_c levels off (compare equations (2.7) (2.11), and Fig. 1).

2.2 NORMALIZATION TO THE GALAXY COVARIANCE FUNCTION

If this process of gravitational clustering is an adequate model for what happened in the actual Universe, then the fiducial quantities in (2.7)–(2.10) can be determined from empirical estimates of the mass scale which is just turning around at the present time. The covariance function data provide the basis for such a determination, if we assume that the galaxies are distributed in the same way as gravitating matter in general. (This assumption is consistent with our general picture.) The empirical covariance function can be fitted by the power-law

$$\xi(r) = A(hr)^{-\gamma} \tag{2.13}$$

where the best (though somewhat uncertain) values of the parameters are $\gamma \simeq 1.77$, and $A = 4.5 \times 10^{44}$ cgs units (Peebles 1974).

The mean density within a sphere of radius r centred on a galaxy is given by

$$\langle\rho\rangle \simeq \frac{3}{3-\gamma}\,\bar{\rho}\xi(r) \quad (\xi(r) > 1). \tag{2.14}$$

Now the mean density within an object which is now at turn-around ($t_{\text{turn}} = t_0 = 2 \times 10^{17} g^{-1/2} h^{-1}$ s) is $\rho_{\text{turn}} = 5.5\,\rho_{\text{co}}\,g$, where $\rho_{\text{co}} = \Omega^{-1}\bar{\rho} = 3H_0^2/8\pi G$. The number g is 1 for $\Omega = 1$ and $\sim(1.5)^{-2}$ for $\Omega \ll 1$ (so that $\rho_{\text{co}}\,g$ is the critical density in an Einstein–de Sitter universe with the same *age,* rather than the same H_0, as the actual Universe). The radius of a typical enhancement which is now at turn-around is therefore given by

$$A(hr_{\text{turn}})^{-\gamma} = 5.5\left(\frac{3-\gamma}{3}\right)\Omega^{-1}g \tag{2.15}$$

i.e.

$$r_{\text{turn}} = h^{-1}A^{1/\gamma}\left(5.5\,\frac{3-\gamma}{3}\,\Omega^{-1}g\right)^{-1/\gamma} \tag{2.16}$$

and the corresponding mass is

$$M_{\text{turn}} = {}^4\!/\!_3\,\pi g \cdot 5.5\,\rho_{\text{c0}}\,r_{\text{turn}}{}^3. \tag{2.17}$$

For $\Omega = 0.2$ and $h = 0.5$ numerical values are

$$M_{\text{turn}\,0} = 5.5 \times 10^{13} M_\odot \tag{2.18}$$

$$r_{\text{turn}\,0} = 4.4\ \text{Mpc}. \tag{2.19}$$

If a sphere of mass $M_{\text{turn}\,0}$ virialized at a radius $\frac{1}{2}r_{\text{turn}\,0}$, its velocity dispersion would be

$$\langle v_0^2\rangle^{1/2} = 385\ \text{km/s}. \tag{2.20}$$

These correspond to the dimensions of a typical sparse group of galaxies and we adopt them as fiducial values in what follows. Given these values, the spread in masses implied by (2.12) is sufficient to encompass rich clusters (*cf.* Section 4). The uncertainties in the procedure leading to (2.18)–(2.20) stem not only from imprecision in the data (2.13) but – more importantly – from the oversimplified model we have adopted for the actual non-linear clustering process. For the latter reason, there are also some grounds for doubting the precise applicability of the scaling relations (2.7)–(2.10).

2.3 NON-DISSIPATIVE HIERARCHICAL CLUSTERING

In so far as the clustering process can be modelled by a hierarchy with discrete steps (each separated by a factor 2 in t) we expect the following picture. Those units which turn around at a given time, t, will have a characteristic mass $M(t)$ (equation (2.7)) and a density of order $5.5\,\rho_c(t)$. These units will each be composed of three or four subunits which have just collapsed and so have turn-around times of order $t/2$ and densities $\sim 44\,\rho_c(t)$. At time $2t$, these units will in turn have collapsed and relaxed and will themselves be grouped in threes and fours, these new groups having binding energies per unit mass larger by a factor $2^{4/9\alpha-2/3}$ than those of their predecessors. However, during (or soon after) their collapse the units will have destroyed their internal substructure; *a three or four member group which is more tightly bound than its constituent members will be transformed by relaxation effects into an amorphous system in $\lesssim 1$ crossing time.* This important result can be shown by rough analytic estimates (see Appendix) and is also found in N-body simulations (White 1976b; Aarseth, Gott & Turner 1978, in preparation).

We thus expect that a 'snapshot' of the Universe taken at time t would show the dark material to be clustered on a characteristic mass scale, which has been in existence for $\lesssim 1$ crossing time. A few subunits may be distinguishable within each subcluster, but *any* finer structure within the subunits would have been erased by collisions, mergers and tidal effects. A second snapshop taken at time $2t$ would show the same picture on the next level of the hierarchy: the masses $M(t)$ would now have collapsed and virialized, their substructure being erased in the process. Notice that in this picture one never expects to see clusters with more than a few members, nor to see clusters with distinct substructure and short crossing times. Because rich clusters do exist, and because many observed groups and clusters have $t_{\text{crossing}} \ll t_0$, a purely dissipationless picture cannot describe the formation of both galaxies and clusters (*cf.* White 1977).

At the present epoch, collapsed clusters have a characteristic scale $\sim 1.5\times 10^{13}h^{-1}M_\odot$, with a tail towards higher masses (*cf.* Section 4). We wish to argue that the bulk of cosmic matter may indeed have behaved as described above, being distributed smoothly on scales $< 10^{13}h^{-1}M_\odot$. We suggest that this material constitutes the so-called 'missing mass' in clusters and the extensive halos of isolated galaxies; we further suggest that *all* the luminous matter seen in galaxies formed from residual gas that settled within the potential wells provided by the dark material at each stage of the clustering process and then collapsed to form stars. We must therefore next consider how the gravitational field of the clusters dark matter would influence any remaining gas. The main new features in the problem are dissipation and cooling.

3 The fate of gas

3.1 COOLING

Ionized gas can be supported by its own pressure in a potential well characterized by a mass M and radius r if its temperature T is given by

$$3\,kT = \frac{GMm_{\rm p}}{r} \tag{3.1}$$

(we assume that the gas mass is $<M$, so we can ignore its self-gravitation). Gas within the potential well will be heated to $\sim T$ either by one strong shock or by a succession of weak ones during the violent relaxation that accompanies formation of the halo.

For the masses and radii relevant to galaxies, the temperatures (3.1) are well above 10^4 K, so any shocked or pressure-supported gas will indeed be ionized. It will then (even if it is pure H and He) cool radiatively: it cannot remain in equilibrium for more than a cooling timescale unless it can draw on a further supply of energy.

The cooling rate due to bremsstrahlung, recombination and collisionally-excited line emission, can be written as $\Lambda(T)\, n_{\rm e} n_{\rm H}$ erg/s cm^6; and the cooling time is

$$t_{\rm cool} \simeq \frac{3\,kTm_{\rm p}}{\rho_{\rm gas}\Lambda\gamma(T)}. \tag{3.2}$$

Of obvious interest is the ratio of $t_{\rm cool}$ to the Hubble time. Also of interest is its relation to the dynamical or formation timescale $t_{\rm dyn}$, which is equal to twice the turn-around time $t_{\rm turn}$ for the corresponding potential well. If $t_{\rm cool} \lesssim t_{\rm dyn}$, any gas within the potential well must cool and collapse to the centre, probably fragmenting into stars. If $t_{\rm cool} > t_{\rm dyn}$, the gas will be pressure-supported and will contract quasi-statically towards the centre, until eventually it becomes self-gravitating and able to fragment. Clearly this quasi-static contraction will not yet have proceeded far unless $t_{\rm cool} \lesssim H_0^{-1}$. If the potential wells are due to agglomerations of dark material with typical mass $M(t)$ (*cf.* (2.7)) then they may themselves collide and merge into a mass on the next stage of the hierarchy after a further time $\sim t_{\rm dyn}$. During these mergers the gas will be shock-heated and transformed into a single hotter and more rarified cloud, for which $t_{\rm cool}/t_{\rm dyn}$ is even larger than it was before (and thus more unfavourable for condensation and fragmentation). Thus if $t_{\rm cool}/t_{\rm dyn} > 1$, condensation will be possible only in that small fraction of cases when a mass survives for an unusually long time before being incorporated into a larger system. A precise criterion would depend on the profile of the potential well and the detailed kinematics and dynamics, but we can make the approximate statement that gas can accumulate, cool (and possibly fragment) within clusters of dark mass provided that

$$t_{\rm cool} \lesssim H^{-1}, \tag{3.3a}$$

and

$$t_{\rm cool} \lesssim t_{\rm dyn} \tag{3.3b}$$

but this can happen only in a small fraction of cases when (3.3b) is violated.

3.2 FRAGMENTATION INTO LUMINOUS GALAXIES

If the luminous content of galaxies (as opposed to their halos) forms by the collapse of gas within pre-existing potential wells, conditions (3.3) set an upper limit to the possible luminous mass. The argument is similar to that of Rees & Ostriker (1977) and Silk (1977), except that these authors considered *self*-gravitating gas clouds.

Condition (3.3) is *necessary* for fragmentation, but additional requirements must be satisfied if the gas is actually to be able to fragment into stars. Let us suppose that F is the fraction of cosmic material which is gaseous at any stage, and that a gas mass FM settles within the potential well due to a mass M of radius r. Even if the gas can cool, it will not fragment until it has accumulated in a region small enough for its local density (and hence self-gravity) to dominate that of the background halo. If it remained in a spherical cloud and the halo material had uniform density, this would require contraction to a radius such that

$$r_{\text{gas}} \lesssim F^{1/3} r. \tag{3.4}$$

This criterion is obviously a very rough one. If cooling instabilities allow the gas to become inhomogeneous, or if it collapses to a disc, then it may become liable to fragmentation even if r_{gas} exceeds the limit given by (3.4). On the other hand, if the halo material were centrally condensed, r_{gas} would need to be smaller to achieve a sufficient density enhancement (e.g. we would require $r_{\text{gas}} \lesssim Fr$ if the halo had an 'isothermal' r^{-2} density gradient with very small core radius).

A condition such as (3.4) in any case merely determines when fragmentation can *start*. As Larson and others have emphasized, the timescale for star formation is likely to be related to the free-fall timescale by some factor of order unity, but the amount of further gaseous dissipation and central concentration that develops *after* (3.4) is satisfied is exceedingly sensitive to this factor. We interpret (3.4) as defining the radius at which the gas-dynamical collapse calculations of Larson (1974a) first become applicable. The further contraction is likely to be greater when $t_{\text{cool}} \simeq t_{\text{dyn}}$ than when cooling and fragmentation can happen almost instantaneously. Once star formation has been triggered, there is an extra energy input (from young stars, supernovae, etc) which may be able to eject most of the gas before more than some fraction f has turned into stars (Larson 1974b).

3.3 GALAXY FORMATION DURING THE HIERARCHICAL CLUSTERING

We assume that, provided (3.3) holds, the potential wells which exist at any stage of the hierarchy will accumulate a core of luminous stars of total mass fFM. We further assume (*cf.* Larson 1974b) that f scales with the binding energy, so that

$$f \propto M/r. \tag{3.5}$$

The IMF of the luminous stars will be assumed to be the same for all values of M and all stages of the hierarchy. (It could have the form which – with suitable assumptions about the *rate* of gas–star conversion – accords with the observed colours and stellar populations of spirals and ellipticals.)

In this picture, low-mass systems of luminous stars will form early, before the characteristic mass M_{turn} of the hierarchical clustering has attained a high value. The low-mass halos that existed at these early times will by now have lost their identity and merged into larger amorphous systems, *but the luminous material that condensed in their centres may nevertheless have survived to the present day in identifiable stellar systems.* Provided that these luminous cores have become sufficiently concentrated during their cooling and fragmentation phases, they will be able to survive the violent relaxation which destroys their halos. The timescales for their subsequent dynamical evolution are then much larger than t_{dyn} (see Appendix).

When the halos of the first small galaxies are disrupted to form bigger units, the residual gas may again be able to cool and collapse to form a larger central galaxy. The model thus naturally predicts the existence of small satellites around big galaxies. If there is sufficient time before incorporation of the halo in a yet larger unit the central galaxy may swallow some of its larger satellites in the manner envisaged by Tremaine (1976), thus increasing its own luminosity relative to that of its satellite system.

The luminous cores thus survive as fossils of the earlier stages of the hierarchy and the luminosity function is *not* of the 'synchronic' form (2.12), but rather must be estimated by assuming that the number of galaxies which formed via condensation into halos of mass between M and $2M$ varies as M^{-1} (since a constant fraction ~ ½ of all the mass in the Universe would at some stage have been incorporated in bound systems in any given mass range). A specific model for the luminosity function is constructed in Section 4.

The formation of the dark material may have resulted in the injection of pregalactic metals into the residual gas. Further, much of the gas is effectively recycled at each stage of the hierarchy, and so may be enriched. The X-ray detection of Fe in some rich clusters is thus no embarrassment to our scheme. Moreover the progressive enhancement provides a further reason why the more massive – and hence more recently formed – galaxies have higher metal abundance (*cf.* Larson 1974a, b). Note that differences in the star formation *rate* will affect the present stellar population and mass to light ratio of galaxies; a more efficient early conversion of gas to stars leads to a higher present M/L for a given IMF. Although we suggest that smaller galaxies condensed first, less efficient star formation could account for the persistence of gas (and for the presence of young stars) in some such systems.

4 A specific model

4.1 CHOICE OF PARAMETERS F AND f

In this section, we show that, if the amplitude of the clustering is normalized to agree with the covariance function, as discussed in Section 2 (equations (2.18)–(2.20)) then the processes described in Section 3 can lead to a system of galaxies whose luminosity function and characteristic parameters are consistent with observation.

The velocity v_0 (equation (2.20)) corresponds to a temperature

$$T_0 = 3.5 \times 10^6\,\mathrm{K}. \tag{4.1}$$

The temperature of typical bound condensations formed at earlier times will vary according to the scaling law (2.10). The corresponding scaling law for the cooling time is

$$\left(\frac{t_{\mathrm{cool}}}{t_{\mathrm{cool}\,0}}\right) = \frac{\Lambda(T_0)}{\Lambda(T)}\left(\frac{T}{T_0}\right)\left(\frac{\rho_0}{\rho}\right) \tag{4.2}$$

where the density and temperature are scaled according to (2.8) and (2.10) respectively. In what follows we consider two possible chemical compositions for the gas which formed the last generation of galaxies: a metal-free mixture of H and He, in the ratio 10:1 by number; and a mixture enriched to 10 per cent of the 'cosmic' metal abundance. Cooling curves for these two cases, assuming collisional ionization, are taken from Cox & Tucker (1969) and Raymond, Cox & Smith (1976).

The parameters F and f describing the amount of residual gas and its rate of conversion into stars can be determined from the observed properties of rich clusters of galaxies, since – on our hypothesis – the mass ratios of dark matter, luminous stars and gas in such systems should have their universal values. The fraction F_i of cosmic matter which remained gaseous after formation of the dark material is

$$F_i = \frac{\text{mass of gas + luminous mass in galaxies}}{\text{total mass}}. \tag{4.3}$$

The fractional depletion of this gas during subsequent star formation sets the value of f_{max} in the scaling law

$$f = f_{\text{max}} \left(\frac{M}{M_{\text{max}}} \right)^{2/3-\alpha} \tag{4.4}$$

which is derived from (3.5) under the assumption that all galaxies form sufficiently early for the x dependence in the scaling laws to be unimportant (we verify this later). M_{max} is then the maximum halo mass in which a luminous core can form and f_{max} is the corresponding efficiency for conversion of the contained gas into stars. Approximating the clustering by a discrete hierarchical process of N stages, we then have

$$\prod_{\substack{N \text{ stages of} \\ \text{hierarchy}}} (1 - \tfrac{1}{2} f_n) = \prod_{n=1,N} \left(1 - \frac{f_{\text{max}}}{2} (2^{2/3-4/9\alpha})^{n-1} \right)$$

$$= \frac{\text{gaseous mass}}{\text{gaseous mass + luminous mass}}. \tag{4.5}$$

The factor ½ in front of f_n corresponds to the assumption that half of the total mass participates in each stage of the hierarchy. (At early times, half the matter will be in overdense regions of mass M, and half in underdense regions.)

On the basis of observations, we take $F_i = 0.2$, and a value of 0.5 for the RHS of (4.5). In what follows, we also assume $h = 0.5$; and treat α, for the moment, as a free parameter.

4.2 THE MAXIMUM MASS AND LUMINOSITY FOR GALAXIES

In Fig. 2 we plot the dependence on α of various critical halo masses. The minimum halo mass, M_{min}, in which a luminous core can form is set when $\alpha \gtrsim 1/3$ by the condition that the formation time be later than recombination, and when $\alpha \lesssim 1/3$ by the condition that the virial temperature be $\gtrsim 10^4$ K. These masses define the lower limit of the hierarchy. They are quite strongly α dependent, but are $\lesssim 10^9 M_\odot$ for α in the range 0.05–0.6. (Under our assumptions, this limiting mass corresponds to the very low limiting luminosity $\sim 5 \times 10^6 f_i L_\odot$.) The maximum halo mass, M_{max}, in which a luminous core can form is set by the condition $t_{\text{cool}} = t_0$, and is of order $10^{13} M_\odot$ for all interesting values of α. The corresponding cores will have luminosities $L_{\text{max}} \sim 5 \times 10^{10} f_{\text{max}} L_\odot$. In fact, however, there will be a depletion of luminous cores when $t_{\text{cool}} > t_{\text{dyn}}$. We see from Fig. 2 that this condition gives typical masses (and luminosities) ~ 3 times lower than the condition $t_{\text{cool}} < t_0$.

We see from Fig. 2 that the inclusion of cooling due to metals does not affect these critical masses substantially. At $z \gtrsim 20$, Compton cooling on the microwave background would dominate radiative cooling. This has not been allowed for in the calculations leading to Fig. 2, but in fact even radiative cooling is always efficient enough to guarantee $t_{\text{cool}} < t_{\text{dyn}}$ throughout the domain where Compton cooling is important, so no conclusions are altered by its omission.

The line $t_{\text{dyn}} = t_0$ in Fig. 2 denotes the typical mass of collapsed clusters. This mass is

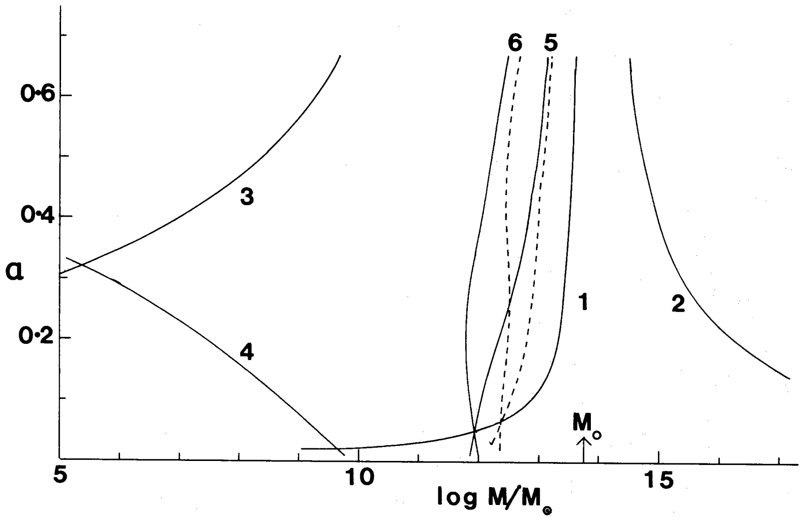

Figure 2. Various characteristic masses are here plotted, as a function of α (equation (2.1)), for the case when $\Omega = 0.2$, $h = 0.5$ and F_f (the fraction of cosmic matter still in gaseous form) is 0.1. The amplitude of (2.1) is normalized by choosing the scale mass M_0 to have the value (2.19). The curves are as follows: (1) The typical mass for which $t_{\rm dyn} = t_0$ (i.e. $t_{\rm turn} = t_0/2$). (2) The mass $M_{1\,\rm per\,cent}$ such that 1 per cent of the material is in units with $t_{\rm dyn} < t_0$. (3) The typical mass for which $t_{\rm dyn} = t_{\rm rec}$. (4) The typical mass for which $T_{\rm virial} = 10^4\,$K. This sets the minimum halo mass within which gas can condense if H and He are the dominant cooling agents. (5) The typical mass for which $t_{\rm cool} = t_0$. The luminous cores of galaxies will have a characteristic maximum mass $0.1/f$ times this value (see Table 1). (6) The typical mass for which $t_{\rm cool} = t_{\rm dyn}$. The continuous curves 5 and 6 correspond to cooling by H and He alone, and the dashed curves to the inclusion of heavy elements with 10 per cent of their 'cosmic' abundance.

related to M_c (equation (2.12)). If, as we assume in what follows, our normalization procedure picks out the mass scale such that half of all galaxies are at present in turning condensations of mass $>M_0$, one can show that (2.12) may be written more precisely as

$$N(M)\,dM \propto M^{\alpha-2} \exp\,[-0.23(M/M_c)^{2\alpha}]\,dM \tag{4.6}$$

where M_c is now the mass given by the line $t_{\rm dyn} = t_0$ in Fig. 2. In this case, 1 per cent of galaxies are in collapsed condensations of mass larger than that given by the line $t_{\rm dyn\,1\,per\,cent} = t_0$. (These masses are in fact related to M_0 by

$$M_c = 0.81^{1/\alpha} M_0 \quad \text{and} \quad M_{1\,\rm per\,cent} = 3.10^{1/\alpha} M_0;$$

the corresponding cluster luminosities are simply $L_c = {}^1\!/_{200} M_c$ and $L_{1\,\rm per\,cent} = {}^1\!/_{200} M_{1\,\rm per\,cent}$.) According to Gott & Turner (1977), the requirement that the distribution of cluster mass be broad enough to account for rich clusters as well as small groups leads to a value of α close to ⅓. Press & Schechter (1974) argued from numerical simulations that the effective α could not be larger than ½. Fig. 2 shows quite clearly that values of α much less than ⅓ are ruled out by the observed lack of collapsed systems significantly more massive than $10^{15} M_\odot$.

In Table 1 we present our calculated values for various characteristic parameters of the hierarchy for four different values of α; where two values are given for any quantity the upper value corresponds to the assumption of a metal-free gas mixture and the lower to the assumption of an enriched mixture. The number of stages in the equivalent discrete hierarchy, N, was calculated from the formula

$$N = \mathrm{Int}\left[\frac{3\alpha}{2}\,\frac{\log\,(M_{\rm max}/M_{\rm min})}{\log 2} + 1\right]. \tag{4.7}$$

Table 1.

α	N	$\log(M_{\max}/M_{\odot})$	$\log(L_{\max}/L_{\odot})$	$\log(r_{\max}/1\,\text{kpc})$	$\log(v_{\max}/1\,\text{km/s})$	$(t_{\text{dyn}\max}/t_0)$	$\log(M_c/M_{\odot})$	$\log(L_c/L_{\odot})$
⅔	12	13.13	9.87	2.44	2.73	0.18	13.60	11.30
		13.21	9.95	2.54	2.72	0.23		
½	23	12.96	10.02	2.40	2.66	0.19	13.56	11.26
		13.07	10.13	2.52	2.66	0.25		
⅓	12	12.67	10.05	2.35	2.54	0.22	13.48	11.18
	13	12.92	10.30	2.55	2.57	0.33		
⅙	4	12.22	9.85	2.35	2.32	0.37	13.18	10.88
	5	12.71	10.33	2.64	2.42	0.59		

This is the number which we used in (4.5) to derive a value for $f_{\max}$. We see from Table 1 that $L_{\max}$ is very insensitive to α or to the metal content of the gas; further, it is very close to the corresponding scale in the observed galaxy luminosity function. This agreement is one of the strong points of our scheme. The radius $r_{\max}$ in Table 1 is that of a homogeneous sphere with the same mass and energy as a typical halo of mass $M_{\max}$. The large values found for $r_{\max}$ substantiate our earlier claim that the luminous parts of galaxies are too concentrated for galaxies and clusters to fit on a continuous hierarchy. They are, however, consistent with our contention that condition (3.4) specifies the initial radius for a gaseous collapse model of the type proposed by Larson (1974a). As expected the corresponding (three-dimensional) halo velocity dispersions, $v_{\max}$, are similar to those observed in the luminous part of large elliptical galaxies, and we note that the formation timescales $t_{\text{dyn}\max}$ corresponding to $M_{\max}$ are sufficiently small that our neglect of the x dependence in the scaling law for f (equation (4.4)) is indeed justified.

4.3 A MODEL FOR THE LUMINOSITY FUNCTION

Under our assumptions, the mass to light ratio is the same for all clusters, so the cluster mass function (4.6) can be converted directly to a *cluster* luminosity function

$$N(L)\,dL \propto L^{\alpha-2} \exp\,[-0.23(L/L_c)^{2\alpha}]\,dL. \tag{4.8}$$

To calculate the *galactic* luminosity function we need to consider the joint mass-density distribution of halos which turn around at any stage of the clustering process. This distribution is

$$N(m,\rho)\,dm\,d\rho \propto m^{-2}\,dm \exp\,(-0.23m^{2\alpha}\rho^{2/3})\,m^{\alpha}\,d(\rho^{1/3}) \tag{4.9}$$

where m is the mass in units of $M_{\max}$ and ρ is the density in units of the corresponding typical scale density $\rho_{\max}$.

The luminosity of a galaxy, L, is proportional to $fM \propto M^2/r \propto M^{5/3}\rho^{1/3}$, where we neglect the slow variation of F (from 0.2 to 0.1 over the whole range of the hierarchy). This gives

$$m = l^{3/5}\rho^{-1/5} \tag{4.10}$$

where $l = L/L_{\max}$. Thus the joint luminosity-density distribution is

$$N(l,\rho)\,dl\,d\rho \propto l^{(3\alpha-8)/5}\rho^{-(3\alpha+7)/15} \exp\,(-0.23\,l^{6\alpha/5}\rho^{(10-6\alpha)/15})\,dl\,d\rho. \tag{4.11}$$

A luminosity function can be calculated by integrating this function over those densities at each luminosity for which $t_{\rm cool} < t_0$. We make the approximation $t_{\rm cool} \propto T/\rho$, which is reasonable in the temperature range of interest, since $\Lambda(T)$ is fairly flat there. This gives

$$t_{\rm cool}/t_0 = l^{2/5}\rho^{-4/5}, \tag{4.12}$$

the condition $t_{\rm cool} < t_0$ then giving $\rho > l^2$. Thus

$$\phi(l)\,dl \propto l^{(3\alpha-8)/5}\,dl \int_{l^2}^{\infty} d\rho\,\rho^{-(3\alpha+7)/15} \exp\left(-0.23\,l^{6\alpha/5}\rho^{(10-6\alpha)/15}\right). \tag{4.13}$$

Making the substitution $Y = \rho^{(8-3\alpha)/15} l^{(24\alpha-9\alpha^2)/(25-15\alpha)}$ in the integral reduces this to

$$\phi(l)\,dl \propto l^{-(8-3\alpha)/(5-3\alpha)}\,dl \int_{l^{(80-6\alpha-9\alpha^2)/(75-45\alpha)}}^{\infty} dY \exp\left(-0.23\,Y^{(10-6\alpha)/(8-3\alpha)}\right). \tag{4.14}$$

An adequate approximation to this expression over the whole range of l is obtained by setting the integral equal to the value of the integrand at the lower limit, yielding the final result

$$\phi(L)\,dL \propto L^{-(8-3\alpha)/(5-3\alpha)} \exp\left[-0.23(L/L_{\rm max})^{(20+6\alpha)/15}\right]. \tag{4.15}$$

Because of the disruption of halos for which $t_{\rm dyn} < t_{\rm cool} < t_0$ before the contained gas has time to cool, the actual luminosity function will become somewhat steeper than (4.15) for a small range of luminosities below $L_{\rm max}$ (*cf.* Fig. 2). For $\alpha = 1/3$, (4.15) yields a power-law of slope -1.75 at the faint end. This is slightly steeper than Abell's (1975) value of -1.625, and substantially steeper than Schechter's (1976) value of -1.25. We note that our predicted slope would be flattened by any process which made low-mass galaxies relatively more vulnerable to disruption. (E.g. star formation may occur more rapidly when $t_{\rm cool} \ll t_{\rm dyn}$, making low-mass galaxies less centrally condensed than high-mass galaxies relative to the halos in which they form.)

With these provisos, the form and scale of our luminosity function compare quite well, at least at the bright end, with those of the fitting functions empirically derived by Schechter (1976). The agreement is as good as could be expected, given the schematic nature of the theory; note that some effects which may influence the upper end of the luminosity function have been neglected (Ostriker & Tremaine 1975; White 1976a). We stress that for a given α, the predicted *cluster* luminosity function (equation (4.8)) does not have the same shape or scale as $\phi(L)$.

5 Conclusions

There is much evidence that ~80 per cent of the matter in the Universe is not in gas or luminous stars, but is now in some dark form. On the (almost mandatory) assumption that this dark material condensed before or soon after recombination and clustered gravitationally on progressively larger scales, we have argued as follows:

(i) The dark material must now be in amorphous units whose mass spectrum spans the range from massive galactic halos to rich clusters of galaxies.

(ii) The luminous inner part of galaxies *cannot* have formed by purely dissipationless clustering. Rather, it most probably condensed from residual gas lying in the transient potential wells provided by the dark matter. By the present time, half this residual gas has been incorporated into luminous galaxies, the rest (perhaps enriched with heavy elements) remains uncondensed in intergalactic space. An upper limit to galactic luminosities is set by the requirement that the gas should have time to settle in a potential well, cool and fragment into luminous stars. This limit agrees adequately with the masses, luminosities and radii of large galaxies. The existence of giant galaxies surrounded by satellites, embedded in a common dark halo, is a natural consequence of our model. The model also explains why, in clusters such as Coma, the masses of the largest galaxies are so much less than that of the system as a whole.

(iii) On somewhat specific accumptions, a luminosity function can be derived which agrees reasonably well with observation.

Acknowledgments

We acknowledge helpful discussions with M. Fall, J. Gunn, B. Jones, J. Ostriker and J. Silk.

References

Abell, G. O., 1975. In *Galaxies and the Universe*, ed. Sandage, A. *et al.*, Chicago University Press.

Alladin, S. M., Potdar, A. & Sastry, K. S., 1974. In *Dynamics of stellar systems*, ed. Hayli, A., D. Reidel, Dordrecht, Holland.

Binney, J. J., 1977. *Astrophys. J.*, **215**, 483.

Cox, D. O. & Tucker, W. H., 1969. *Astrophys. J.*, **157**, 1157.

Gott, J. R., Gunn, J. E., Schramm, D. N. & Tinsley, B. M., 1974. *Astrophys. J.*, **194**, 543.

Gott, J. R. & Turner, E. L., 1977. *Astrophys. J.*, **216**, 357.

Gunn, J. E. & Gott, J. R., 1972. *Astrophys. J.*, **176**, 1.

Larson, R. B., 1974a. *Mon. Not. R. astr. Soc.*, **166**, 585.

Larson, R. B., 1974b. *Mon. Not. R. astr. Soc.*, **169**, 229.

Ostriker, J. P. & Tremaine, S. D., 1975. *Astrophys. J.*, **202**, L113.

Peebles, P. J. E., 1974. *Astrophys. J.*, **189**, L51.

Press, W. H. & Schechter, P., 1974. *Astrophys. J.*, **187**, 425.

Raymond, J. C., Cox, D. P. & Smith, B. W., 1976. *Astrophys. J.*, **204**, 290.

Rees, M. J., 1977. In *Evolution of galaxies and stellar populations*, p. 339, eds Larson, R. B. & Tinsley, B. M., Yale University Observatory Publications.

Rees, M. J. & Ostriker, J. P., 1977. *Mon. Not. R. astr. Soc.*, **179**, 451.

Schechter, P., 1976. *Astrophys. J.*, **203**, 297.

Silk, J. I., 1977. *Astrophys. J.*, **211**, 638.

Spitzer, L., 1958. *Astrophys. J.*, **127**, 17.

Spitzer, L. & Chevalier, R. A., 1973. *Astrophys. J.*, **183**, 565.

Sunyaev, R. A., 1971. *Astr. Astrophys.*, **12**, 190.

Tremaine, S. D., 1976. *Astrophys. J.*, **203**, 72.

Toomre, A., 1977. In *Evolution of galaxies and stellar populations*, p. 401, eds Larson, R. B. & Tinsley, B. M., Yale University Observatory Publications.

van Albada, T. S. & von Gorkom, J. H., 1977. *Astr. Astrophys.*, **54**, 121.

White, S. D. M., 1976a. *Mon. Not. R. astr. Soc.*, **174**, 19.

White, S. D. M., 1976b. *Mon. Not. R. astr. Soc.*, **177**, 717.

White, S. D. M., 1977. *Comm. Astrophys.*, 7, 95.

White, S. D. M. & Sharp, N. A., 1977. *Nature*, **269**, 395.

Appendix: the disruption of substructure

When a bound unit first collapses in any dissipationless clustering process, it will be extremely inhomogeneous, being composed of a spectrum of smaller lumps which collapsed on shorter timescales than the system as a whole. A graphic example of this is given in White

(1976b). As the system collapses the subunits interact violently and lose their identity. This ironing out of substructure combines elements of at least three different dynamical processes. Encounters between sublumps give rise to strong tidal forces which increase the internal energy of individual lumps at the expense of their orbital motion through the system; this effect leads to the tidal evaporation and disruption of the lumps in the manner discussed by Spitzer (1958) for star clusters. The transfer of energy from orbital motion to internal motions during an encounter between two lumps can, however, result in their becoming bound to each other and merging into a single more diffuse object. This stickiness has been investigated in the context of galaxy–galaxy encounters by Alladin, Potdar & Sastry (1974), Toomre (1977) and van Albada & von Gorkom (1977). The third process which contributes to the destruction of substructure is dynamical friction; heavy subunits can rapidly give up their kinetic energy of motion through the cluster both to lighter subunits and to individual particles and as a result they settle to the centre of the system where they can disrupt and merge more easily. Elements of all these processes are discernible in the rapid destruction of subclustering in N-body simulations (*cf.* White 1976b and Aarseth, Gott & Turner 1978, in preparation). *In these simulations the substructure of any bound unit is rubbed out almost as soon as it collapses.* The rough analytic arguments given in this Appendix show that this important result should still be valid when the numbers of distinct particles making up the units and sub-units are far higher than can be simulated by N-body methods.

Spitzer (1958) shows that the change in internal energy of a lump of mass m_1 in an impulsive encounter at impact parameter D, pericentric distance p, velocity difference at infinity V_∞ and pericentric velocity difference V_p with another lump of mass m_2 is approximately

$$\Delta U_1 \simeq 4G^2 m_1 m_2^2 r_1^2/3p^4 V_\text{p}^2 \gtrsim 4G^2 m_1 m_2^2 r_1^2/3D^4 V_\infty^2 \tag{A1}$$

where, following Spitzer & Chevalier (1973), we take r_1 to be the half-mass radius of the lump. Assuming the system as a whole to have mass M, half-mass radius R and velocity dispersion V, we take $V_\infty = \sqrt{2}V$ in (A1) and integrate over all possible impacts to get:

$$\frac{dU_1}{dt} = 4.2\,\frac{G^2 m_1 r_1^2}{V}\int_0^\infty dm_2 n(m_2)\, m_2^2 [r_1^2 + r_2^2]^{-1} \tag{A2}$$

where $n(m_2)\,dm_2$ is the number density of lumps in the range $(m_2, m_2 + dm_2)$ and where we have used a minor variation of the prescription of Spitzer & Chevalier (1973) to deal with the lower limit of the integration over impact parameters. Adopting $3M/8\pi R^3$ as a typical density for the system as a whole we find

$$\frac{dU_1}{dt} = 0.50\,\frac{G^2 m_1 r_1^2 QM}{VR^3}\left\langle \frac{m}{r^2 + r_1^2}\right\rangle \tag{A3}$$

where Q is the fraction of the mass of the system in lumps, and angular brackets denote a mass-weighted average over the lumps. If we assume $U_1 = Gm_1^2/4r_1$ (a good approximation for all likely density profiles for the lumps) and a formation time for the whole system $T_\text{dyn} = 2\pi R/V$, we find that the disruption time is given by

$$t_{\text{dis}\,1} = \frac{U_1}{dU_1/dt} = \frac{1}{25}\,\frac{1}{Q}\,\frac{m_1}{\langle m[1 + (r/r_1)^2]^{-1}\rangle}\,\frac{R}{r_1}\,T_\text{dyn}. \tag{A4}$$

This clearly suggests that any object in which much of the mass is in fairly diffuse sublumps will destroy its substructure during, or shortly after, its collapse. In our hierarchy we expect $r/R \sim (m/4M)^{1/3}$ giving $t_{\text{dis}} \simeq 0.13\,Q^{-1}(m/M)^{-1/3}\,T_{\text{dyn}}$, and so all but the central regions of any sublumps will be disrupted when a unit collapses.

The approximations leading to (A4) neglect the orbital energy loss of the lumps. In fact two lumps are quite likely to capture each other during an encounter and to merge rapidly thereafter. We now estimate the timescale on which such mergers take place under the simplifying assumption that all lumps are similar. Defining v to be the internal velocity dispersion of a lump ($v^2 = Gm/2r$) we further assume a velocity-dependent capture cross-section

$$\sigma(V_\infty) = \begin{cases} 0 & V_\infty > v \\ 4\pi r^2[1 + 4(v^2/V_\infty^2)] & V_\infty < v \end{cases}. \tag{A5}$$

The normalization and cut-off of this cross-section are suggested by *N*-body experiments (White 1978, in preparation); its velocity dependence merely accounts for gravitational focusing. Comparison of (A5) with the results of Alladin *et al.* (1974) suggest that it is a conservative criterion and underestimates the capture efficiency at low velocities. Assuming that the lumps have typical number density $3MQ/8\pi mR^3$ and a Gaussian distribution of orbital velocities with dispersion V, the expected time for a lump to capture another lump is given by

$$\frac{1}{t_{\text{cap}}} = \frac{3MQ}{8\pi mR^3}\int_0^{v/V} dx\left(\frac{27}{4\pi}\right)^{1/2} x^2 \exp\left(-3x^2/4\right) 4\pi r^2\left(1 + 4\frac{v^2}{V^2}x^{-2}\right) Vx. \tag{A6}$$

This leads to

$$t_{\text{cap}} = 0.032\,\frac{1}{Q}\,\frac{mr^{-2}}{MR^{-2}}\left(\frac{v}{V}\right)^{-4} T_{\text{dyn}}, \tag{A7}$$

where we have assumed $v \lesssim V$ in approximating the integral over the velocity distribution. To apply this to substructure in our hierarchy, we take $r/R \simeq (m/4M)^{1/3}$, as before, and find $t_{\text{cap}} = 0.032Q^{-1}(m/M)^{-1}\,T_{\text{dyn}}$, suggesting that merging of subunits will occur rapidly as any system collapses.

Any lump which escapes the initial violent relaxation of a system unscathed will subsequently experience dynamical friction as it moves through the cluster and loses its orbital energy to lighter objects. For a typical cluster model, the time for this friction to bring a lump into the cluster centre is (*cf.* White 1977)

$$t_{\text{fric}} \simeq \frac{1}{8\ln(R/r)}\,\frac{M}{m}\,T_{\text{dyn}}. \tag{A8}$$

Clearly all but the smallest lumps will quickly spiral to the centre where they will be incorporated in the general density profile of the system.

It is clear from equations (A4), (A7) and (A8) that the bound collapsed systems which form in a dissipationless clustering hierarchy cannot long retain any substructure. Thus in our model the halo of dark material around any luminous galaxy core must disrupt and merge with other halos as soon as it becomes part of a larger unit. Provided that the violent processes accompanying this initial merging do not bring the luminous cores into orbits too close to the centre of the final object, the dynamical timescales (equations (A4), (A7) and (A8)) are too long for the cores to undergo significant further evolution before the new system is incorporated in a yet larger object. When this happens the group of galaxy cores and its common halo are broken up together. Since the luminous galaxy cores form highly condensed subsystems within their halos, we expect that after the halos have merged the cores will be more concentrated to the cluster centre than the dark matter, but will still be well separated. In our model this concentration can explain why binary galaxies and small groups of galaxies appear to have lower mass to light ratios than rich clusters. We stress, however, that our understanding of the clustering and gas condensation processes is not good enough for us to be able to specify the exact conditions under which luminous cores can survive the disruption of their halos without merging at the centre of the resulting object. It is clearly necessary for the pregalactic gas to contract until it is self-gravitating and it seems probable that its final radius needs to be significantly less than the core radius of its halo. In any other situation it is difficult to escape the arguments of White & Sharp (1977) against the existence of isothermal halos in binary systems, and of White (1977) and Sections 1 and 2 against dissipationless galaxy formation in general.