1993MNRAS.262..627L

Mon. Not. R. Astron. Soc. **262,** 627–649 (1993)

Merger rates in hierarchical models of galaxy formation

Cedric Lacey[1] and Shaun Cole[2]

[1] *Physics Department, Oxford University, Astrophysics Building, Keble Road, Oxford OX1 3RH*
[2] *Department of Physics, University of Durham, Science Laboratories, South Road, Durham DH1 3LE*

Accepted 1992 November 17. Received 1992 October 29; in original form 1992 June 17

ABSTRACT

We present an analytical description of the merging of virialized haloes which is applicable to any hierarchical model in which structure grows via gravitational instability. The formulae are an extension of the Press–Schechter model. The dependence of the merger rate on halo mass, epoch, the spectrum of initial density fluctuations and the density parameter Ω_0 is explicitly quantified. We calculate the distribution of halo formation times and survival times. We also describe a Monte Carlo method for constructing representative histories of merger events leading to formation of haloes of a prescribed mass.

Applying these results to the age distribution of rich clusters of galaxies, we infer that a high value of the density parameter ($\Omega_0 \gtrsim 0.5$) is required to reproduce the substantial fraction of rich clusters that exhibit significant substructure, if such substructure only persists for a time $0.2\,t_0$ after a merger, where t_0 is the present age of the universe. We also investigate the rate of infall of satellite galaxies into galactic discs, by combining our Monte Carlo technique for halo mergers with an estimate of the time required for dynamical friction to erode the orbits of the baryonic cores of the accreted galaxies. We find that, even for $\Omega_0 = 1$, the infall rate is low (provided that the satellite orbits are not too eccentric), and that we would expect only a modest fraction of stellar discs to be thickened or disrupted by this process.

Key words: galaxies: clustering – galaxies: evolution – galaxies: formation – galaxies: interactions – cosmology: theory – dark matter.

1 INTRODUCTION

It is usually assumed that galaxies, and the large-scale structure that they trace, grew via gravitational instability from small-amplitude Gaussian density fluctuations, generated by physical processes in the very early Universe. In hierarchical models, including the cold dark matter (CDM) model, the amplitude of these fluctuations decreases with increasing scale, resulting in the formation of low-mass objects, which then merge with one another to build up ever more massive structures. This process can be studied by means of *N*-body simulations, but it is important to be able to understand the results of such simulations in simpler terms. In this paper, we present an analytical description of the development of virialized structures which form by dissipationless hierarchical collapse, which is applicable to the dark matter haloes of individual galaxies, and to groups and clusters of galaxies. This description is based on the 'random walk' or 'excursion set' formalism developed by Bond et al. (1991, hereafter BCEK), which involves smoothing the linear density field on various mass scales, and identifying collapsed regions as those above some density threshold. The model predicts a mass function for bound virialized objects which is identical to that derived more heuristically by Press & Schechter (1974). More importantly, the model allows one to make calculations of the merging history of haloes. We compute the halo merger rate, including the dependence on the masses of both haloes involved, on the initial spectrum of density fluctuations, and on the epoch and the value of the cosmological density parameter Ω_0. We also calculate the typical epochs at which haloes of a given present mass formed by mergers of smaller haloes, and how long they typically survive before being merged into much larger structures. This should provide greater insight into the results of the *N*-body simulations, and into the evolution of structure in the real Universe.

The model should also provide a framework for more detailed calculations of galaxy formation. In hierarchical

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

models in which the universe is dominated by dark matter, the formation of luminous galaxies is thought to occur in virialized dark matter haloes, which are the sites where gas is able to cool and collapse to dense cores where star formation can begin. During this process of dissipative collapse and star formation, these haloes are also undergoing mergers with other haloes. Much effort is currently being invested in trying to simulate this process numerically, using combined hydrodynamical and *N*-body codes incorporating rules for star formation and feedback due to energy injection from young stars (e.g. Evrard 1988; Cen et al. 1991; Navarro & Benz 1991). However, limitations of computer time and memory severely limit the dynamic range of these simulations, and also do not allow exploration of a large parameter space. An alternative to these full-scale simulations is to use the analytical description of hierarchical merging developed here, and to model galaxy formation by adding equations for gas cooling, star formation, etc. Such an approach will provide insight into the results from the hydrodynamical simulations, and also complement them, in that a larger dynamic range can be covered (even if only in an approximate way), and a much wider and more systematic investigation of the effects of varying the many uncertain parameters associated with star formation and feedback effects is possible.

There is currently a great deal of interest in values of merger rates for objects of various types. Observations of rich clusters of galaxies show that a significant fraction (~ 30 per cent) of them have substructure, detectable either in their X-ray surface brightness profiles (Jones & Forman 1992) or in the spatial distribution of their galaxies (Dressler & Shectman 1988). This implies that they have only recently formed by merging, and suggests a fairly high current merger rate, ~ 1 per Hubble time. This is what one expects for a high-density universe with $\Omega_0 \sim 1$, in which fluctuations in the linear regime continue to grow, and merging of dark haloes proceeds even at the present epoch. Merging of clusters provides a fairly direct application of our formalism, insofar as the luminous material traces the dark haloes in these systems. We consider this question in Section 4.1, where we use the observed frequency of substructure to estimate a lower bound on the allowed value of Ω_0.

In contrast, observations suggest a much lower merger rate for luminous galaxies. Toomre (1977) estimated a current merger rate ~ 0.1 per Hubble time, based on the number of galaxies observed to have large tidal tails, which were interpreted as being the relics of recent mergers. Simulations show that the merging of two galaxies of comparable mass will produce an elliptical galaxy, provided that both galaxies were mainly stellar beforehand. However, 80–90 per cent of bright galaxies are disc galaxies which have formed their stars over a substantial fraction of the age of the Universe, implying that the average rate of merging between comparable-mass galaxies cannot have exceeded ~ 0.1 per Hubble time. There are strong constraints even on the accretion of small galaxies by large galaxies. If a satellite galaxy made of stars merged with the disc of a spiral galaxy, it would heat the stellar disc, causing it to thicken; yet the discs of spiral galaxies are observed to be quite thin, implying that the amount of mass accreted in this form must be quite small. Toth & Ostriker (1992) argue that this rules out a high-density Universe in which structure forms hierarchically. However, the rate of merging of luminous galaxies is more difficult to calculate than that for haloes, because the luminous cores of baryonic matter need not merge when the surrounding haloes merge, but may instead end up orbiting inside the new large halo. This is thought to be the situation in galaxy clusters. Dynamical friction will eventually cause the baryonic cores to spiral together and merge. We discuss this effect, and its application to accretion of satellite galaxies by spiral discs, in Section 4.2.

Mergers are implicated in various types of unusual activity in galaxies. The most infrared-luminous galaxies seen by *IRAS* appear to be starbursts triggered by the merger of two galaxies (e.g. Sanders et al. 1988). It has also been suggested that there is a link between mergers and the formation of active galactic nuclei, quasars and radio galaxies (e.g. Heckman et al. 1986). Merging has also been proposed as a possible explanation for the surprisingly steep increase in the number density of galaxies on the sky seen as one probes to fainter magnitudes (Rocca-Volmerange & Guiderdoni 1990; Guiderdoni & Rocca-Volmerange 1991; Broadhurst, Ellis & Glazebrook 1992). A merger rate large enough to reconcile the galaxy counts with a high-density Universe would imply that mergers have played a significant role in the formation of the present-day population of galaxies. However, this idea appears to be hard to reconcile with the evidence listed above for a low merger rate for luminous galaxies.

We note that Carlberg (1990a) calculated halo merger rates by manipulating the halo mass function derived originally by Press & Schechter (1974). The formula that he deduced he then employed to model the evolution of the number density of quasars with redshift (Carlberg 1990a), and to constrain Ω_0 and Λ_0 by modelling the evolution of the galaxy merger rate (Carlberg 1990b, 1991). We do not agree with Carlberg's formulae, although our results are likewise consistent with the Press–Schechter mass function. We discuss the differences in Section 5.

The plan of our paper is as follows. The following section, Section 2.1, describes the ideas and reasoning on which the analytic description of hierarchical merging is based. In Section 2.2 we derive the mass function of non-linear bound objects or haloes (equation 2.11), and then in Section 2.3 we derive expressions for the merger probabilities and rates and for halo accretion rates (equation 2.18). In Sections 2.4 and 2.5 we compute expressions for halo survival and formation times. We then describe in Section 3 a Monte Carlo technique that employs these formulae to generate merger histories leading to haloes of a given present mass. We use this technique to extract the distributions of halo formation times and masses, and the distribution of masses of the haloes that were merged with. We then apply these formulae and techniques to two astrophysical problems of current interest. We first, in Section 4.1, estimate a lower bound to Ω_0 by investigating the incidence of recent mergers in rich clusters. Then, in Section 4.2, we make Monte Carlo estimates of the current infall rate of baryonic cores into typical galactic haloes. Combining this with an estimate of the time required for dynamical friction to be efficient at eroding the original orbits of these infalling cores, we estimate the infall rate on to galactic discs present at the centres of these dark matter haloes. We compare our results with those of Carlberg in Section 5, and present our conclusions in Section 6.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

2 ANALYTICAL RESULTS FOR HALO MERGING

2.1 Basic principles

At early times, when the amplitude of density fluctuations is small ($\delta = \Delta\rho/\rho \ll 1$), these perturbations grow according to linear theory: $\delta(\boldsymbol{x}, t) = \delta(\boldsymbol{x}, t_0) D(t)/D(t_0)$, where $D(t)$ is the linear growth factor and $\boldsymbol{x}$ is a comoving coordinate. The density contrast in a given region obeys this simple relation until $\delta(\boldsymbol{x}, t)$ approaches unity, at which point non-linear effects become important, and the region ceases to expand, turns around, and collapses to form a virialized halo. At the point at which the virialized halo forms, the density contrast estimated by linear theory will have reached a critical value δ_c which can be estimated from the evolution of an isolated spherical overdense region. A useful alternative way of viewing this evolution is to consider simply the linear density field $\delta_0(\boldsymbol{x}) \equiv \delta(\boldsymbol{x}, t_0)$ extrapolated to some fixed time t_0 (perhaps the present day), and a critical threshold $\delta_{0c}(t)$ that is progressively lowered with increasing time. Thus we can now identify the regions which will have collapsed to form virialized haloes at time t as those regions in the linear density field for which $\delta_0 \geq \delta_{0c}(t)$. Henceforth, we use δ to denote the extrapolated linear density δ_0, and omit the '0' subscripts on δ and δ_c.

The critical density threshold δ_c can be estimated for any cosmology by considering the growth, turnaround and collapse of a uniform spherical overdense region. If the parameters of this overdense region are selected so as to match the density and expansion rate of a growing-mode linear perturbation of the background universe at early times, then δ_c at a halo formation time $t_{\rm coll}$ can be found from the extrapolated linear amplitude for an idealized spherical perturbation which collapses to a point at time $t_{\rm coll}$. This calculation, performed in the Appendix, yields the threshold

$$\delta_c(t_{\rm coll}) = \frac{3}{2} D(t_0) \left[1 + \left(\frac{t_\Omega}{t_{\rm coll}} \right)^{2/3} \right] \qquad (\Omega_0 < 1)$$
$$= \frac{2(12\pi)^{2/3}}{20} \left(\frac{t_0}{t_{\rm coll}} \right)^{2/3} \qquad (\Omega_0 = 1), \qquad (2.1)$$

where $D(t)$ is an Ω_0-dependent growth factor given in the Appendix, and $t_\Omega = \pi H_0^{-1} \Omega_0 (1-\Omega_0)^{-3/2}$. The result for $\Omega_0 = 1$ can be obtained by taking the limit $t_\Omega/t_0 \to \infty$ of that for $\Omega_0 < 1$. The value of the threshold at the present epoch, $\delta_c(t_0)$, is only weakly dependent on Ω_0: for $\Omega_0 = 1$, $\delta_c(t_0) \approx 1.686$, while, for $\Omega_0 = 0.1$, $\delta_c(t_0) \approx 1.615$. The behaviour of $\delta_c(t)$ as a function of time reflects the fact that, in a flat ($\Omega_0 = 1$) universe, structure continues to grow at all epochs, while, in a low-density ($\Omega_0 < 1$) universe, linear perturbations grow like those in a flat universe for $t \ll t_\Omega$, but stop growing for $t \gtrsim t_\Omega$, when $D(t) \to 1$. Thus the time t_Ω marks the epoch at which new structures cease forming and the existing haloes simply move farther and farther apart without any further mergers. The spherical collapse model also predicts that haloes which have just collapsed and virialized have mean densities $\sim (100\text{–}200)\Omega^{-1}$ times the background density at that time.

In a hierarchical model, in which the rms density fluctuations decrease with increasing scale, the first haloes to form have low masses. These haloes then accrete further material and merge together to produce haloes of progressively larger and larger masses. A remarkably simple analytic description of this history of hierarchical merging has been developed by Cole (1989) and Bond et al. (1991; BCEK hereafter), which appears to be in surprisingly good quantitative agreement with the hierarchical mergers synthesized in cosmological N-body simulations. Their approach is to smooth (average) the linear density field δ over spheres of successively larger masses, and then to assume that the mass of the halo containing a given particle at time t equals the mass M of the largest sphere, placed around the initial position of that particle, within which the average δ exceeds the threshold for collapse, $\delta_c(t)$, calculated as described above. Thus, at each point, one considers the trajectory $\delta(M) = \langle \Delta\rho/\rho \rangle_M$ of the linear density field as a function of the smoothing mass M, and finds the largest M for which $\delta(M)$ crosses through $\delta = \delta_c(t)$. Selection of the largest mass is a way of addressing an issue which pundits call the cloud-in-cloud problem (e.g. Bardeen et al. 1986; Peacock & Heavens 1990). It ensures that the halo so identified will not have been engulfed in a still larger structure, since the surrounding region when averaged on all larger scales will have a mean density below the critical value and so, according to our criterion, will not yet have collapsed.

2.2 The mass function

To demonstrate how this approach of following trajectories $\delta(M)$ can lead to a simple expression for the mass function and the merger rates, it is useful to consider the Fourier decomposition of the linear density field:

$$\delta(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \delta_{\boldsymbol{k}} \exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}). \qquad (2.2)$$

For a Gaussian random field, the different Fourier amplitudes $\delta_{\boldsymbol{k}}$ are independent random variables with random phases. Thus a Gaussian random field is, statistically, completely determined by this power spectrum $\langle |\delta_{\boldsymbol{k}}|^2 \rangle$, which measures the mean square amplitude of the various modes. The density field is smoothed by convolving it with a spherically symmetric window function $W_M(r)$ having some radial extent R, corresponding to a mass $M \sim \rho_0 R^3$. Thus the smoothed field $\delta(M, \boldsymbol{x})$ gives the weighted average of $\delta(\boldsymbol{x})$ over a spherical region of mass M around each point $\boldsymbol{x}$. Applying the convolution to the Fourier series, the smoothed field can be expressed as

$$\delta(M, \boldsymbol{x}) \equiv \int W_M(|\boldsymbol{x}-\boldsymbol{y}|)\, \delta(\boldsymbol{y})\, \mathrm{d}^3\boldsymbol{y} \qquad (2.3)$$
$$= \sum_{\boldsymbol{k}} \delta_{\boldsymbol{k}} \hat{W}_M(k) \exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}),$$

where $\hat{W}_M(k)$ is the Fourier transform of the spatial window function $W_M(r)$:

$$\hat{W}_M(\boldsymbol{k}) = \int W_M(\boldsymbol{x}) \exp(-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x})\, \mathrm{d}^3\boldsymbol{x}. \qquad (2.4)$$

At fixed $\boldsymbol{x}$, equation (2.3) gives the equation of a trajectory $\delta(M)$.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

Many choices are possible for the spatial window function, besides the obvious one where $W_M(r)$ is constant inside a sphere, and zero outside ('top hat filtering'). In general, one wants $W_M(r)$ to be nearly constant at small r, and then to fall off steeply beyond some radius R. The integral of $W_M(r)$ over all space is normalized to unity, so that $W_M(|x-y|)$ acts as a smoothing kernel. Correspondingly, $\hat{W}_M(k)$ tends to unity at small k, and then falls off beyond some value $k_c \sim 1/R \propto M^{-1/3}$, thus heavily suppressing the contribution to $\delta(M)$ from modes of wavelength smaller than the size of the smoothing window. In what follows, it will be convenient to label the mass scale M by the variance

$$S(M)=\sigma^2(M)=\langle|\delta(M,\boldsymbol{x})|^2\rangle = \sum_{\boldsymbol{k}} \langle|\delta_{\boldsymbol{k}}|^2\rangle \hat{W}_M^2(k) \tag{2.5}$$

of the linear density field when smoothed with the window function containing mass M, and to consider the trajectories to be functions of the variable S. Note that, under fairly general conditions, S is a monotonically decreasing function of M. (A sufficient condition is that $\hat{W}_M^2(k)$ be monotonically decreasing with M.) For scale-free initial conditions in which the power spectrum is a pure power law $\langle|\delta_{\boldsymbol{k}}|^2\rangle \propto k^n$, the variance as a function of mass is simply $S \propto M^{-\alpha}$, where $\alpha=(n+3)/3$.

For physically interesting power spectra and window functions, if the smoothing mass scale is sufficiently large then S will tend to zero, and so $\delta(S)$ will also approach zero. As one reduces the scale over which the density field is smoothed, $\delta(S)$ will begin to wander away from zero as progressively shorter and shorter wavelength modes begin to contribute. (Examples of such trajectories can be seen in fig. 3 of BCEK.) In fact, from equation (2.5) it can be seen that the mean square value of $\delta(S)$ is given by $\langle|\delta(S)|^2\rangle = S$. For a given realization of the density field, i.e. a given set of $\delta_{\boldsymbol{k}}$ values, the trajectories at all spatial locations are determined. We do not attempt to relate the trajectories at different locations. Instead, we consider the trajectories at a fixed location $\boldsymbol{x}$ that are obtained for different realizations of the $\delta_{\boldsymbol{k}}$. These trajectories can then be considered as random walks. Statistical properties of the trajectories are calculated by averaging over realizations, which by the usual ergodic theorem corresponds to averaging over spatial locations. The detailed properties of these trajectories depend on the form of the window function chosen, but their description is particularly simple for the case in which $\hat{W}_M(k)$ is a step function in k-space ('sharp k-space filtering'):

$$\hat{W}_M(k)=\begin{cases}1 & k<k_s(M),\\ 0 & k>k_s(M).\end{cases} \tag{2.6}$$

In this case, these wandering trajectories are true Brownian random walks, as each increment to $\delta(S)$ when $S(k_s)$ is increased comes from a new set of Fourier modes in a thin spherical shell in k-space, and thus for a Gaussian random field is uncorrelated with any of the previous steps. The consequence of this is that these trajectories $\delta(S)$ are governed by a simple diffusion equation. If we denote the number density of trajectories at S in the interval δ to $\delta+\mathrm{d}\delta$ as $Q(\delta, S)$, then

$$\frac{\partial Q}{\partial S}=\frac{1}{2}\frac{\partial^2 Q}{\partial \delta^2}. \tag{2.7}$$

We note that the variable S acts like the time variable in this diffusion equation. All the trajectories begin at $S=0$, $\delta=0$ and then diffuse away as S increases. The more complex behaviours exhibited with other choices of window function are explored in BCEK.

We now wish to calculate the fraction of trajectories that are above the threshold $\delta_c(t)$ at some mass scale M but are below this threshold for all larger values of M. This is equivalent to identifying that fraction of the trajectories that have their first upcrossing through the threshold $\delta=\omega=\delta_c(t)$ in the interval S to $S+\mathrm{d}S$, which corresponds (through equation 2.5) to a mass interval M to $M+\mathrm{d}M$. In order to evaluate this, consider placing a barrier at $\omega=\delta_c(t)$ which absorbs trajectories as they attempt to cross through it. The unique solution of the diffusion equation (2.7) with this absorbing boundary condition is

$$Q(\delta, S, \omega)\,\mathrm{d}\delta=\frac{1}{\sqrt{2\pi S}}\left\{\exp\left(-\frac{\delta^2}{2S}\right)-\exp\left[-\frac{(\delta-2\omega)^2}{2S}\right]\right\}\mathrm{d}\delta \tag{2.8}$$

(Chandrasekhar 1943). This follows from the fact that trajectories starting at some point on the line $\delta=\omega$ are equally likely to wander above it as below it. The probability that a particular trajectory will be absorbed by the barrier in the interval S to $S+\mathrm{d}S$ must equal the reduction in the number of trajectories surviving below the barrier. Hence

$$f_S(S,\omega)=-\frac{\partial}{\partial S}\int_{-\infty}^{\omega} Q\,\mathrm{d}\delta = -\left[\frac{1}{2}\frac{\partial Q}{\partial \delta}\right]_{-\infty}^{\omega}, \tag{2.9}$$

where the second line follows from equation (2.7). We use the notation $f_S(S, \omega)$ for the probability density in S, which is a function of both S and ω, i.e. $f_S(S, \omega)\,\mathrm{d}S$ is the probability that a trajectory will have its first upcrossing through the threshold in the interval $(S, S+\mathrm{d}S)$. Substituting the expression in equation (2.8) for Q, we find

$$f_S(S,\omega)\,\mathrm{d}S=\frac{\omega}{(2\pi)^{1/2}S^{3/2}}\exp\left[-\frac{\omega^2}{2S}\right]\mathrm{d}S. \tag{2.10}$$

This expression represents the fraction of mass associated with haloes in the range of M corresponding, through equation (2.5), to the specified range in S. The comoving number density of haloes of mass M present at time t is therefore

$$\frac{\mathrm{d}n}{\mathrm{d}M}(M,t)\,\mathrm{d}M=\frac{\rho_0}{M}f_S(S,\omega)\left|\frac{\mathrm{d}S}{\mathrm{d}M}\right|\mathrm{d}M = \left(\frac{2}{\pi}\right)^{1/2}\frac{\rho_0}{M^2}\frac{\delta_c(t)}{\sigma(M)}\left|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\right|\exp\left[-\frac{\delta_c(t)^2}{2\sigma^2(M)}\right]\mathrm{d}M, \tag{2.11}$$

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

where ρ_0 is the present mean mass density of the universe. This is the well-known expression for the mass function originally proposed by Press & Schechter (1974). Using equation (2.1) to define $\omega = \delta_c(t)$ as a function of time, it can be used to determine the mass spectrum of haloes as a function of time in open or flat cosmologies.

The cumulative mass fraction in haloes above some mass M is given by integrating equation (2.10) from $S=0$, with the result

$$P(>M, t) = P(<S, \omega) = \mathrm{erfc}\left(\frac{\omega}{\sqrt{2S}}\right) = \mathrm{erfc}\left[\frac{\delta_c(t)}{\sqrt{2}\sigma(M)}\right]. \quad (2.12)$$

Thus, if the power spectrum is such that $S=\sigma^2 \to \infty$ as $M \to 0$, then $P(>M) \to 1$ as $M \to 0$, so that all matter is in haloes of some mass.

Let us review some of the assumptions inherent in our derivation of equations (2.11) and (2.12) by considering a cosmological N-body simulation. For each particle in the simulation, we can construct a trajectory $\delta(S)$ by explicitly smoothing the linear density field around its initial location. From these trajectories, we can compute trajectories $M(t)$ for each particle as we have described above. Thus, for any later time, this procedure can be used to tag each particle in the N-body simulation with a mass label, which is our prediction for the mass of halo in which we expect the particle to be incorporated at time t. For equations (2.11) and (2.12) to be exact would require that there be an exact correspondence between this tagged mass and the true halo mass. This correspondence between tagged mass and halo mass is at best approximate, and thus equation (2.11) can only be viewed as an approximation of the true halo mass function. Later, in Section 2.5, we will find that assumption of an exact identity between tagged mass and halo mass can have more serious consequences, and can lead to some mild inconsistencies.

To use equation (2.11) or (2.12) for a particular cosmological model, all one requires is the function $S(M)=\sigma^2(M)$, which is the linear theory variance of the density fluctuations as a function of mass scale. To determine this function, one must select a form of the window function $W_M(r)$, relate the filter radius R or cut-off wavenumber k_s to the mass M, and compute $S(M)$ using equation (2.5). BCEK elected to use a top hat window function

$$W_M(r) = (4\pi R^3/3)^{-1} \begin{cases} 1 & r \le R, \\ 0 & r > R, \end{cases} \quad (2.13)$$

for which the mass can be unambiguously defined as simply the mass enclosed within radius R, $M = 4/3\pi\rho_0 R^3$. The disadvantage of this choice is that for $n \ge 1$ the summation in equation (2.5) diverges. An alternative would be to use the sharp k-space filter (equation 2.6), which always yields a finite $S(M)$, and can be argued to be the consistent choice as it is this window function that results in the diffusion equation (2.7). The problem that arises here is how to relate the cut-off wavenumber k_s to the mass M. Perhaps the most natural way to achieve this is to rescale the spatial form of the window function for this filter,

$$W_M(r) = \frac{(\sin k_s r - k_s r \cos k_s r)}{2\pi^2 r^3}, \quad (2.14)$$

so that $W_M(r)=1$ at $r=0$, and then to integrate over all space to get the volume enclosed. This yields $M = 6\pi^2\rho_0 k_s^{-3}$. For a power-law power spectrum with $2.5 < n < 0.5$, this mass is 2.5–0.7 times larger than that given by the top hat window function with the same S, and the ratio becomes zero for $n > 1$. For scale-free initial conditions, the choice of window function makes no real difference if one measures masses in units of the characteristic mass M_* for which $\sigma(M_*)=\delta_c$; then $S(M) = \delta_c^2 (M/M_*)^{-\alpha}$, where $\alpha = (n+3)/3$. For other models, such as the CDM model, where n varies slowly with mass, the choice of window function will also change the shape of the mass function, but only by a small amount. Thus the ambiguity in the choice of window function for computation of $S(M)$ only really becomes important when relating the mass function to other measurements of the fluctuation amplitude, such as fluctuations in galaxy counts, the galaxy correlation function, or microwave background $\Delta T/T$ measurements. To determine the optimum choice it is probably best to calibrate the mass function against a large N-body simulation or ensemble of simulations in which the power spectrum of density fluctuations is known precisely. We will address this problem in a later paper, where we will also assess the accuracy of this analytical model. Here we will follow BCEK, and take the $S(M)$ relation appropriate for top hat filtering. For the CDM model, we will use the fit to the power spectrum given by Bardeen et al. (1986).

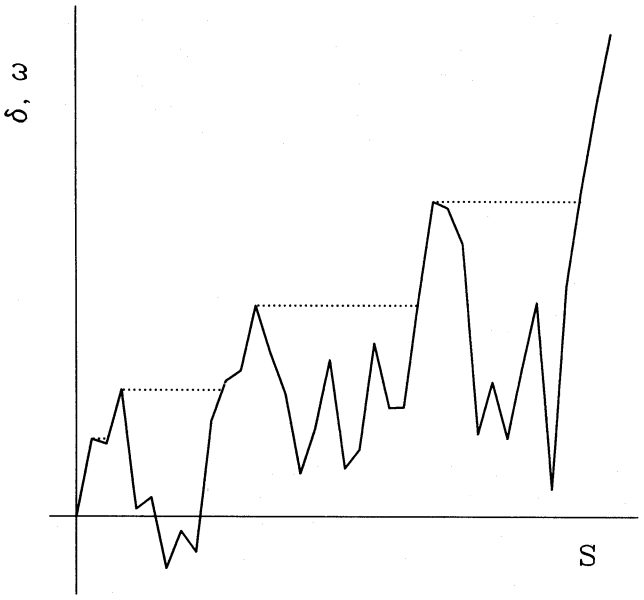

Figure 1. A trajectory $\delta(S)$, and the corresponding halo merger history. The solid line shows the trajectory for the overdensity δ as the smoothing scale is varied. The dotted line shows the trajectory for the halo mass, represented by a function $S(\omega)$. Where δ is increasing with S, the dotted line coincides with the solid line.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

2.3 Halo merger and accretion rates

The great value of viewing the formation of haloes in terms of the trajectories $\delta(S)$ is that it becomes clear how to determine more detailed properties of the merging history of haloes. A given trajectory $\delta(S)$ describes the merger history for a given particle. Fig. 1 shows an example: the solid line is the trajectory $\delta(S)$, while the dotted line shows the merger history $S_{\rm halo}(\omega)$ that is derived from it. The process of following the merging in the normal temporal sequence of increasing t and M corresponds to the process of starting at large ω and S and following the track down and to the left in this figure. Recall that, at each value of the barrier height ω, the halo mass is assumed to correspond to that barrier crossing by the trajectory $\delta(S)$ with the *largest* M, and so the *smallest* S, so that ω versus S for the halo follows $\delta(S)$ when this is increasing, but makes horizontal jumps when $\delta(S)$ decreases. These jumps correspond to sudden jumps in the mass of the halo containing a particle, which we can identify as resulting from merger events. The small steps in $S_{\rm halo}$ corresponding to upward steps in the trajectory of δ versus S correspond to incremental accretion events, adding only a small amount of mass to the halo. Of course, these accretion events are really just mergers with very small haloes, and the distinction made here between 'accretion' and 'merger' events depends on the resolution ΔS with which one looks at the trajectory $\delta(S)$.

We wish to determine the merger probability per unit time for a halo of given mass M at time t. Let us therefore consider the subset of trajectories, depicted in Fig. 2, which make their first upcrossing of a barrier of height ω_2 at S_2 and then continue until they eventually cross a second barrier of height $\omega_1 > \omega_2$ at various values $S_1 > S_2$. These trajectories represent haloes which at the time corresponding to ω_1 have masses corresponding to S_1, and which by the later time corresponding to ω_2 have merged to form a halo of mass corresponding to S_2. The conditional probability $f_{S_1}(S_1, \omega_1|S_2, \omega_2)\,\mathrm{d}S_1$ that one of these trajectories will make its first upcrossing of ω_1 in the interval S_1 to $S_1 + \mathrm{d}S_1$ can be obtained directly from equation (2.10) by noting that this is the same situation as before, but with the source of trajectories moved from the origin to the point (S_2, ω_2). Thus

$$f_{S_1}(S_1, \omega_1|S_2, \omega_2)\,\mathrm{d}S_1 = \frac{(\omega_1-\omega_2)}{(2\pi)^{1/2}(S_1-S_2)^{3/2}} \exp\left[-\frac{(\omega_1-\omega_2)^2}{2(S_1-S_2)}\right]\mathrm{d}S_1 \qquad (2.15)$$

$$(S_1 > S_2,\ \omega_1 > \omega_2),$$

where we have simply made the replacements $S \to S_1 - S_2$ and $\omega \to \omega_1 - \omega_2$ in equation (2.10). This expression can be used directly to yield the mass distribution of the haloes at time t_1 that go on to form haloes of mass M_2 at time t_2. This expression is equivalent to equation (5.1) of BCEK and also to the formula deduced by Bower (1991) by repeated use of the heuristic argument employed by Press & Schechter (1974). Alternatively, we can manipulate equations (2.15) and (2.10) to give the conditional probability that a trajectory with a first upcrossing of ω_1 at S_1 will have a first upcrossing of ω_2 between S_2 and $S_2 + \mathrm{d}S_2$,

$$f_{S_2}(S_2, \omega_2|S_1, \omega_1)\,\mathrm{d}S_2 = \frac{f_{S_1}(S_1, \omega_1|S_2, \omega_2)\,\mathrm{d}S_1 f_{S_2}(S_2, \omega_2)\,\mathrm{d}S_2}{f_{S_1}(S_1, \omega_1)\,\mathrm{d}S_1}$$

$$= \frac{1}{(2\pi)^{1/2}}\left[\frac{S_1}{S_2(S_1-S_2)}\right]^{3/2}\frac{\omega_2(\omega_1-\omega_2)}{\omega_1} \times \exp\left[-\frac{(\omega_2 S_1 - \omega_1 S_2)^2}{2S_1S_2(S_1-S_2)}\right]\mathrm{d}S_2 \qquad (2.16)$$

$$(S_1 > S_2,\ \omega_1 > \omega_2).$$

This yields the conditional probability that a halo of mass M_1 present at time t_1 will have merged to form a halo of mass between M_2 and $M_2 + \mathrm{d}M_2$ at time $t_2 > t_1$. By taking the limit as t_2 tends to t_1 (ω_2 tends to $\omega_1 = \omega$), we can determine a mean transition rate,

$$\frac{\mathrm{d}^2 p}{\mathrm{d}S_2\,\mathrm{d}\omega}(S_1 \to S_2|\omega)\,\mathrm{d}S_2\,\mathrm{d}\omega = \frac{1}{(2\pi)^{1/2}}\left[\frac{S_1}{S_2(S_1-S_2)}\right]^{3/2}\exp\left[-\frac{\omega^2(S_1-S_2)}{2S_1S_2}\right]\mathrm{d}S_2\,\mathrm{d}\omega. \qquad (2.17)$$

Whilst in any finite interval $\Delta\omega$ the change ΔS can be due to the cumulative effects of more than one merger, in an infinitesimal interval $\mathrm{d}\omega$ the entire change ΔS must result from a single merger event. We therefore interpret equation (2.17) as giving the merger rate. It represents the probability that, in the time interval corresponding to $\mathrm{d}\omega$, a halo of mass M_1 will accrete or merge with another halo of mass $\Delta M = M_2 - M_1$, where M_1 and M_2 are related through equation (2.5) to S_1 and S_2. Thus the rate of merging, subdivided according the mass ΔM of the halo that is being merged with, is

$$\frac{\mathrm{d}^2 p}{\mathrm{d}\ln\Delta M\,\mathrm{d}t}(M_1 \to M_2|t) = 2\sigma(M_2)\left|\frac{\mathrm{d}\sigma_2}{\mathrm{d}M_2}\right|\Delta M\left|\frac{\mathrm{d}\omega}{\mathrm{d}t}\right|\frac{\mathrm{d}^2 p}{\mathrm{d}S_2\,\mathrm{d}\omega}(S_1 \to S_2|\omega) \qquad (2.18)$$

$$= \left(\frac{2}{\pi}\right)^{1/2}\frac{1}{t}\left|\frac{\mathrm{d}\ln\delta_{\rm c}}{\mathrm{d}\ln t}\right|\left(\frac{\Delta M}{M_2}\right)\times\left|\frac{\mathrm{d}\ln\sigma_2}{\mathrm{d}\ln M_2}\right|\frac{\delta_{\rm c}(t)}{\sigma_2}\frac{1}{(1-\sigma_2^2/\sigma_1^2)^{3/2}} \times \exp\left[-\frac{\delta_{\rm c}(t)^2}{2}\left(\frac{1}{\sigma_2^2}-\frac{1}{\sigma_1^2}\right)\right],$$

where $\sigma_1 = \sigma(M_1)$ and $\sigma_2 = \sigma(M_2)$. Note that, for an $\Omega_0 = 1$ universe, $\mathrm{d}\ln\delta_{\rm c}/\mathrm{d}\ln t = -2/3$.

The merger rates given by this expression are plotted in Figs 3 and 4. Fig. 3 shows the merger rate by number, for three self-similar models with $n = -2$, -1 and 0, and for the CDM model at $z = 0$ and 1. The CDM model assumes $h \equiv H_0/(100\ \mathrm{km\ s^{-1}\ Mpc^{-1}}) = 0.5$ and a normalization $\sigma_8 \equiv \sigma(8\ h^{-1}\,\mathrm{Mpc}) = 0.5$, which implies a characteristic mass, such that $\sigma(M_*) = \delta_{\rm c}$, of $M_* = 5 \times 10^{12}\ h^{-1}\,\mathrm{M}_\odot$. Fig. 4 shows the fractional rate at which the mass is increased by mergers with other haloes. We see from these figures that mergers with very small haloes dominate numerically

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

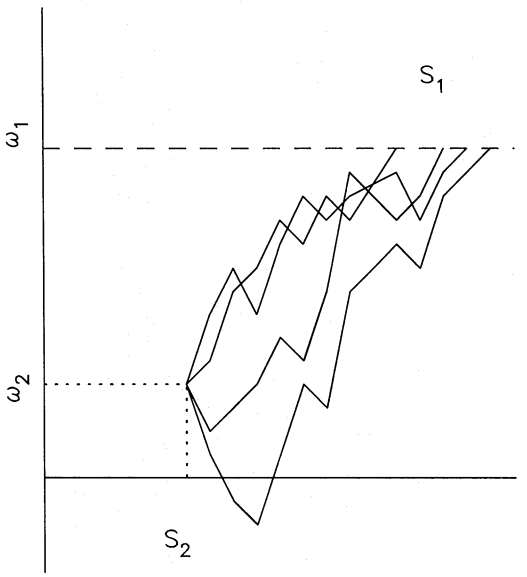

Figure 2. Trajectories of δ versus S for a subset of trajectories that make their first upcrossing of a barrier of height ω_2 at S_2 and then continue until they eventually cross through a second barrier of height $\omega_1 > \omega_2$ at various values S_1.

$[\mathrm{d}^2p/\mathrm{d}\ln\Delta M\,\mathrm{d}t \propto (\Delta M)^{-1/2}$ for $\Delta M \ll M_1]$, reflecting the divergent number of low-mass haloes in the mass function, but that the mass accretion rate is dominated by mergers with haloes of higher masses. For $M_1 \gg M_*$, the total mass accretion rate asymptotes to a constant value of $2/(n+3)(M_1/t)$, and is dominated by the accretion of smaller haloes. For lower masses M_1, the fractional accretion rate is significantly larger, reflected by the increased area under the curves in Fig. 4, but for $M_1 \ll M_*$ most of the increase in mass is due to falling into more massive haloes, $\Delta M > M_1$, rather than to the accretion of smaller systems. For a given value of M_1/M_*, the total mass accretion rate falls as the spectral index n increases. Thus the rapid growth of structure that occurs in the $n=-2$ model is reflected in correspondingly high accretion rates.

2.4 Halo survival times

An interesting quantity that can be calculated from equation (2.16) is the lifetime of a given DM halo. The knowledge of how long a typical halo survives before being incorporated into a much more massive system is of particular importance when attempting to construct physical models of galaxy formation.

Let us consider the fate of all haloes of mass M_1 that exist at time t_1, which we will label, as usual, by S_1 and ω_1. The probability that, by a later time t_2, such a halo will have been incorporated into a new halo of mass greater than M_2 is given by integrating equation (2.16):

$$P(S<S_2, \omega_2|S_1, \omega_1) = \int_0^{S_2} f_{S_2'}(S_2', \omega_2|S_1, \omega_1)\,\mathrm{d}S_2'. \qquad (2.19)$$

Given, however, that the tracks of S versus ω for the halo mass are monotonic (see Fig. 1), this is also the probability that a halo makes the transition from $S<S_2$ to $S>S_2$ at some $\omega>\omega_2$. The probability that the halo will become incorporated into a system of mass larger than that corresponding to S_2 during the time interval corresponding to $\mathrm{d}\omega_2$, which we denote $g_{\omega_2}(S_2, \omega_2|S_1, \omega_1)\,\mathrm{d}\omega_2$, is therefore given by differentiating:

$$g_{\omega_2}(S_2, \omega_2|S_1, \omega_1)\,\mathrm{d}\omega_2 = -\mathrm{d}\omega_2 \frac{\partial}{\partial\omega_2} P(S<S_2, \omega_2|S_1, \omega_1). \qquad (2.20)$$

This yields

$$\begin{aligned} & g_{\omega_2}(S_2, \omega_2|S_1, \omega_1)\,\mathrm{d}\omega_2 \\ & = \left(\frac{2}{\pi}\right)^{1/2} \frac{1}{\omega_1} \left[\frac{S_1}{S_2(S_1-S_2)}\right]^{1/2} \exp\left[\frac{2\omega_2(\omega_1-\omega_2)}{S_1}\right] \\ & \times \left\{ \frac{-S_2(\omega_1-2\omega_2)-S_1(\omega_1-\omega_2)}{S_1} \exp(-X^2) \right. \\ & + \left(\frac{\pi}{2}\right)^{1/2} \left[\frac{S_2(S_1-S_2)}{S_1}\right]^{1/2} \left[1-\frac{(\omega_1-2\omega_2)^2}{S_1}\right] \\ & \left. \times [1+\mathrm{erf}(-X)] \right\} \mathrm{d}\omega_2 \quad (\omega_2<\omega_1, S_2<S_1), \end{aligned} \qquad (2.21)$$

where

$$X \equiv \frac{S_2(\omega_1-2\omega_2)+S_1\omega_2}{[2S_1S_2(S_1-S_2)]^{1/2}}.$$

The corresponding cumulative probability is

$$\begin{aligned} P(S<S_2, \omega_2|S_1, \omega_1) &= P(S_2, \omega>\omega_2|S_1, \omega_1) \\ &= \frac{1}{2}\frac{(\omega_1-2\omega_2)}{\omega_1} \exp\left[\frac{2\omega_2(\omega_1-\omega_2)}{S_1}\right] [1-\mathrm{erf}(X)] \\ &+ \frac{1}{2}[1-\mathrm{erf}(Y)] \quad (S_2<S_1, \omega_2<\omega_1), \end{aligned} \qquad (2.22)$$

where

$$Y \equiv \frac{S_1\omega_2 - S_2\omega_1}{[2S_1S_2(S_1-S_2)]^{1/2}}.$$

We define the survival time t_s of a halo having mass M at time t as the cosmic time by which the mass has doubled due to either accretion or merging. We can use equation (2.21) to determine the distribution of survival times, by setting $S_1 = S(M)$, $S_2 = S(2M)$, $\omega_1 = \omega(t)$ and $\omega_2 = \omega(t_s)$. On the other hand, equation (2.22) gives the fraction of haloes $P(<t_s|M, t)$ with survival times smaller than some value t_s. Fig. 5 shows the distribution of survival times in an $\Omega_0 = 1$ universe, for self-similar models with $n=-2, -1, 0$ and for CDM. These curves show the common feature that for $M \ll M_*$ the distribution of t_s/t is very broad, with a power-

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

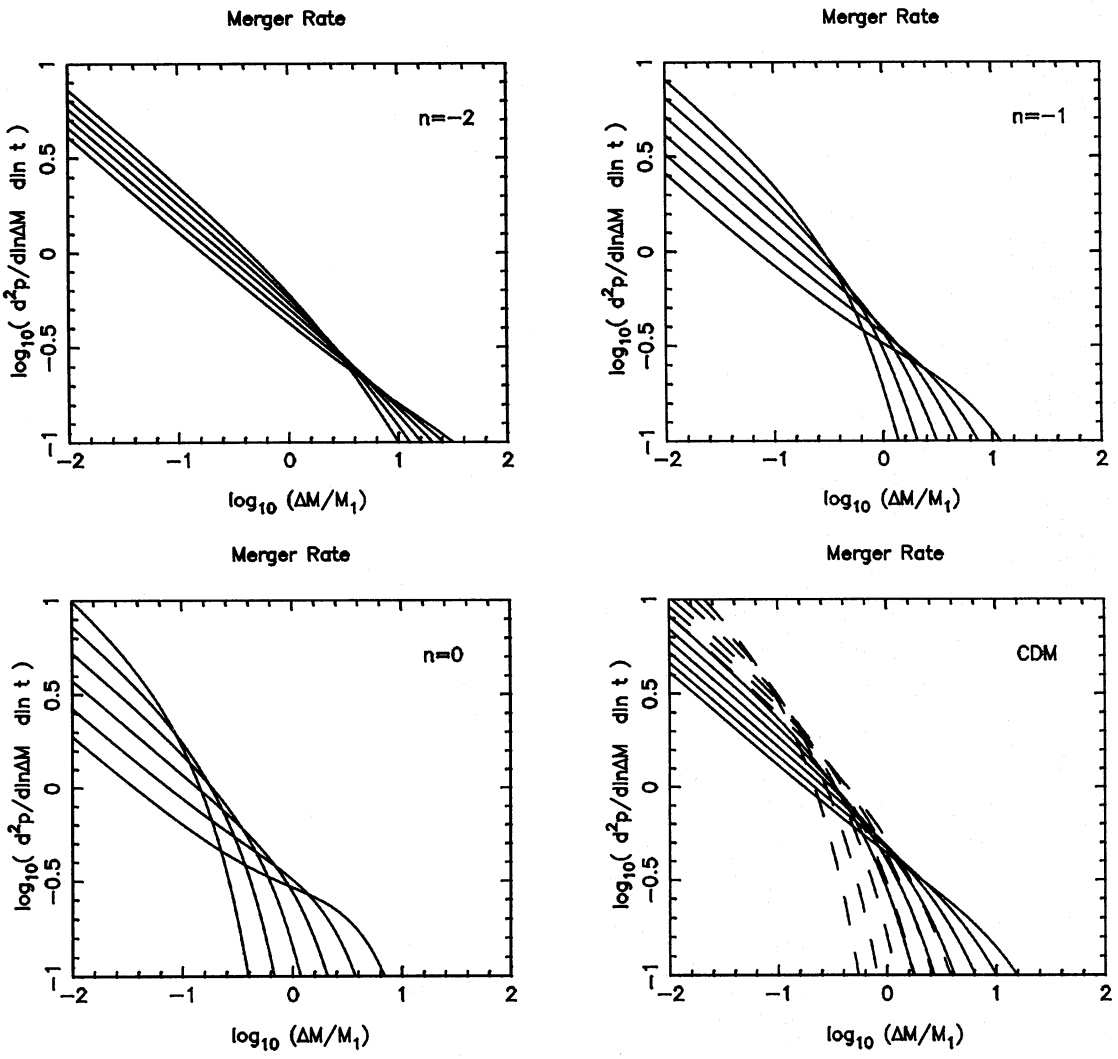

Figure 3. The merger rates given by equation (2.18) for self-similar models and for the CDM model. The quantity plotted, $\mathrm{d}^2p/\mathrm{d}\ln\Delta M\,\mathrm{d}\ln t$, represents the number of haloes of mass ΔM that are accreted in one Hubble time by a halo of mass M_1. The first three panels show the self-similar models and are labelled by n, the spectral index. In these panels, the curve which is highest on the left of the plot is for $M_1 = 16M_*$, and successive curves are for $M_1/M_* = 8$, 4, 2, 1 and 1/2. The fourth panel, which displays merger rates for the CDM model, has curves corresponding to $M_1/10^{13}\,h^{-1}\,\mathrm{M}_\odot = 9.6$, 4.8, 2.4, 1.2, 0.6 and 0.3. The solid curves are for $z = 0$ and the dashed curves correspond to $z = 1$.

law tail extending to larger values, while for $M \gg M_*$ the distribution becomes much narrower, centred around an intermediate value of t_s/t. Thus some low-mass haloes survive for very many Hubble times, but this is very unlikely for high-mass haloes. For the self-similar models with $n > -3$, the limiting behaviour of the median survival time $\hat{t}_s$ can be obtained analytically from equation (2.22): $\hat{t}_s/t \to 2^{3/2}$ for $M \to 0$ and $\hat{t}_s/t \to 2^{3\alpha/2}$ for $M \to \infty$ while, for $n = 0$, $\hat{t}_s/t = 2^{3/2}$, independent of M. Thus, for $n < 0$, the median survival time decreases with increasing mass, while for $n > 0$ the reverse is true. The median survival time is seen to decrease with decreasing n, for a given M/M_*, reflecting the more rapid evolution of structure already remarked on in connection with the merger rate.

The survival time can be compared with the internal dynamical time of a halo. For the spherical collapse model analysed in Appendix A, the internal dynamical time is $T_{\mathrm{dyn}} = (1/2)\,t_{\mathrm{coll}} = (1/2)\,t$, if the halo is modelled as a uniform sphere. Therefore, if $t_s/t < 3/2$, the halo survives for less than one dynamical time after it forms, and so can hardly be said ever to have existed as an equilibrium structure. For $n < -2.6$, more than half of the mass is in haloes with $t_s/t < 3/2$. (This estimate is somewhat sensitive to the adopted value for T_{dyn}/t, however.) For the CDM spectrum, the effective value of n approaches -3 at low masses, suggesting that this may be an important effect during the early stages of structure formation. A second application of these results is to the cooling of gas within dark haloes, where one expects the gas to settle to the centre of the halo only if it has time to cool undisturbed by further mergers. Rees & Ostriker (1977) proposed the approximate criterion $T_{\mathrm{cool}} < T_{\mathrm{dyn}}$ for cooling to be effective, but a more accurate criterion would be $T_{\mathrm{cool}} < (t_s - t)$. We will return to this in a future paper. Cole (1991) has already made investigations along these lines

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

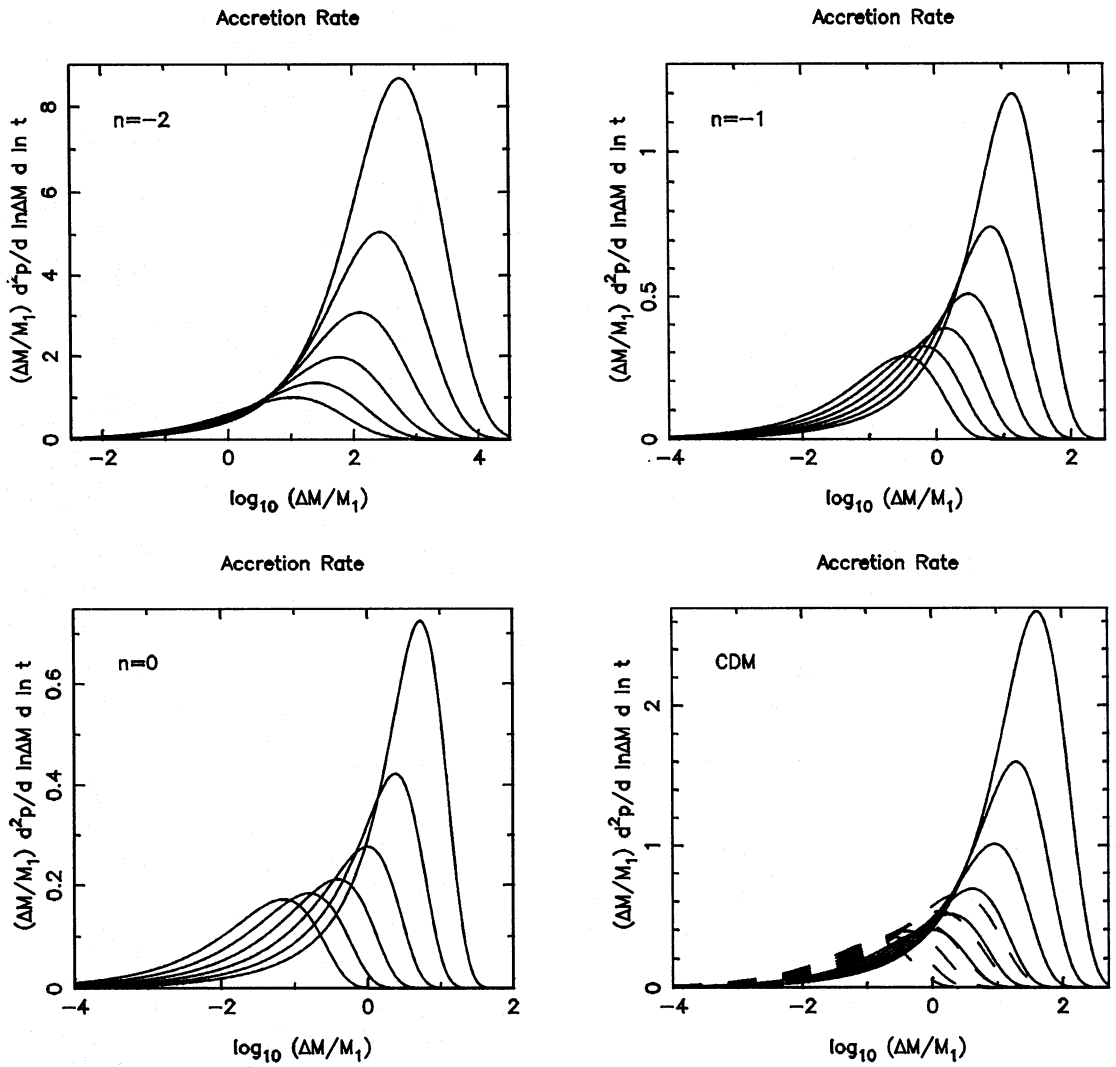

Figure 4. The accretion rates given by equation (2.18) for self-similar models and for the CDM model. The quantity plotted is the fractional mass accreted per Hubble time, $(\Delta M/M_1)\ \mathrm{d}^2p/\mathrm{d}\ \ln\Delta M\ \mathrm{d}\ \ln t$. The first three panels show the self-similar models and are labelled by n, the spectral index. Note the different scales used on the vertical axis. In each panel, the lowest curve at the right is for $M_1 = 16M_*$, and successive curves are for $M_1/M_* = 8$, 4, 2, 1 and 1/2. The fourth panel, which displays the accretion rates for CDM, has curves corresponding to $M_1/10^{13}\ h^{-1}\ \mathrm{M}_\odot = 9.6$, 4.8, 2.4, 1.2, 0.6 and 0.3. The solid curves are for $z = 0$ and the dashed curves correspond to $z = 1$. The net accretion rate in units of M_1/t is proportional to the area under the appropriate curve.

using the 'block model', which is a simple numerical scheme for simulating structure formation.

2.5 Halo formation times: an analytical estimate

The matter making up a halo of mass M at time t was at earlier times distributed in many 'parent haloes' or 'building blocks', which merged hierarchically to produce the current halo. We will define the formation time t_f of the halo as the time when a parent halo appeared which had half or more of the present mass M. Prior to that time, it is not possible to define usefully which of the many parents was the 'main parent'. After that time, the choice of the largest mass parent as the 'main parent' defines a continuous track through the merging tree. A 'merger tree' of this form is depicted schematically in Fig. 6.

Calculation of the distribution of the formation times of haloes defined in this way from the random walks model is somewhat more problematic than the calculation of survival times. We consider several different approaches in turn: two analytical methods in this section, and a Monte Carlo method in Section 3.2.

2.5.1 *Mass-halving time for single trajectories*

Consider first an argument based on following single trajectories. We can determine a *survival time* from a single trajectory $\delta(S)$, by finding the point corresponding to time t and then following it *forward* in time until the mass has *doubled*. If we apply a corresponding procedure going *back* in time, following a trajectory until the mass associated with it has *halved*, we obtain from equation (2.15) the following

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

1993MNRAS.262..627L

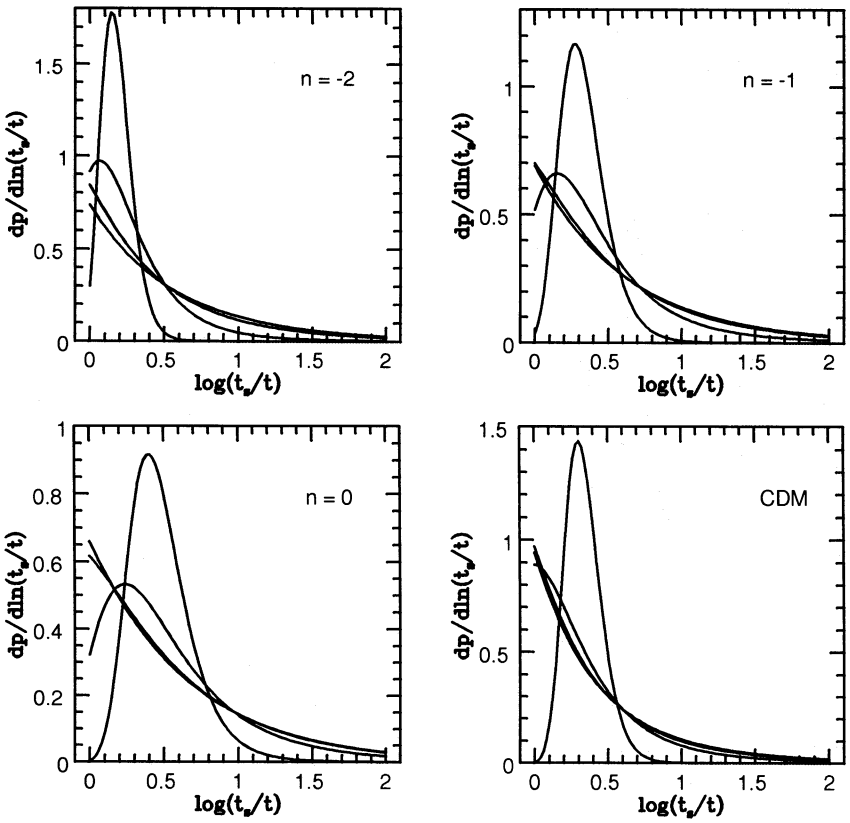

Figure 5. The distribution of survival times of haloes in an $\Omega_0 = 1$ universe. The first three panels show results for power-law power spectra with $n = -2$, -1 and 0 respectively, for masses given by $\nu \equiv \delta_c/\sigma(M) = (M/M_*)^{\alpha/2} = (0.1, 0.3, 1, 3)$, with decreasing ordinate at large t_s corresponding to increasing mass. The fourth panel shows the result for a CDM model at $z = 0$ with $\sigma_8 = 0.5$ and $h = 0.5$, for masses M/h^{-1} $M_\odot = (10^6, 10^9, 10^{12}, 10^{15})$. Again, decreasing ordinate at large t_s corresponds to increasing mass.

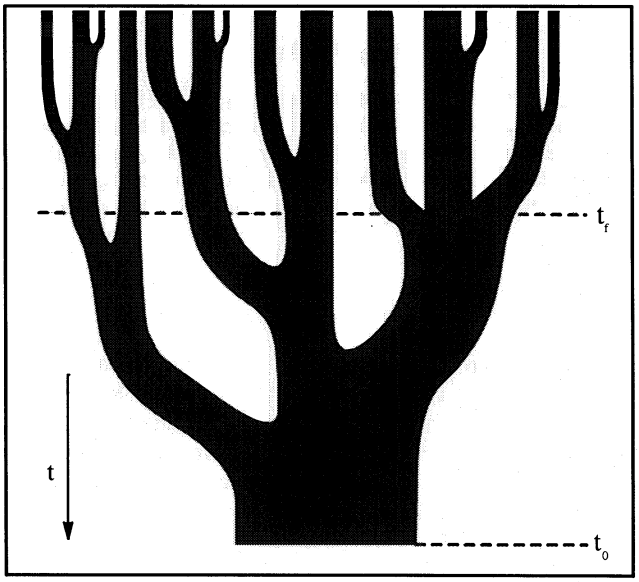

Figure 6. A schematic representation of a 'merger tree' depicting the growth of a halo as the result of a series of mergers. Time increases from top to bottom in this figure and the widths of the branches of the tree represent the masses of the individual parent haloes. A slice through the tree horizontally gives the distribution of masses in the parent haloes at a given time. The present time t_0 and the formation time t_f are marked by horizontal lines, where the formation time is defined as the time at which a parent halo containing in excess of half of the mass of the final halo was first created.

cumulative probability distribution for the time t_h at which this occurs:

$$P(t_h < t_1 | M_2, t_2) = P(M_1 > M_2/2, t_1 | M_2, t_2)$$

$$= \int_{S_2}^{S_h} f_{S_1}(S_1, \omega_1 | S_2, \omega_2)\, \mathrm{d}S_1 \qquad (2.23)$$

$$= \mathrm{erfc}\left[\frac{(\omega_h - \omega_2)}{\sqrt{2(S_h - S_2)}}\right],$$

where $t_1 < t_2$, $M_1 < M_2$ and $S_h = S(M_2/2)$. However, this is *not* the same as the distribution of *halo formation times*, for the reason that, if the halo mass for a trajectory has fallen to some small value at an earlier time, this only means that one of the parent haloes had this small mass; it does not mean that the largest or main parent had a mass below half the current mass. Thus an approach based on single trajectories is inadequate.

2.5.2 *Halo formation times from a counting argument*

In fact, it appears that there is no completely self-consistent way of computing the distribution of formation times from the random walks model, because the correspondence between the halo mass we assign to a particle by analysing its trajectory, $\delta(S)$, and its actual halo mass is only an approximate one. (This will, of course, affect the results on mass

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

functions, merger rates and survival times, as well as formation times, but does not lead to any self-inconsistency in the former cases.) However, two alternative approaches – the analytical counting argument presented in this section, and a Monte Carlo method of generating merging histories presented in Section 3.2 – give rather similar results, so we are encouraged to believe that they may provide a useful approximation to the true answer.

The halo counting argument is as follows. According to the results of Section 2.3, the number density of haloes of mass $(M_1, M_1+\mathrm{d}M_1)$ at time t_1 which are incorporated into haloes of mass $(M_2, M_2+\mathrm{d}M_2)$ at time $t_2 > t_1$ is

$$\mathrm{d}^2 n = \frac{\mathrm{d}n}{\mathrm{d}M_1}(M_1, t_1)\,\mathrm{d}M_1 f_{S_2}(S_2, \omega_2 | S_1, \omega_1)\,\mathrm{d}S_2. \tag{2.24}$$

If we now choose M_1 in the range $M_2/2 < M_1 < M_2$, each of the haloes of mass M_1 at time t_1 must evolve into a distinct halo of mass M_2 at time t_2. (However, not all haloes of mass M_2 have parents of mass M_1 in this range at time t_1 – their largest parent may have mass $< M_2/2$.) Therefore the probability that a halo of mass M_2 at t_2 has a parent at time t_1 with mass in the range $(M_1, M_1+\mathrm{d}M_1)$ is given by dividing equation (2.24) by $(\mathrm{d}n/\mathrm{d}M_2)(M_2, t_2)\,\mathrm{d}M_2$, with the result

$$\frac{\mathrm{d}p}{\mathrm{d}M_1}(M_1, t_1 | M_2, t_2)\,\mathrm{d}M_1 = \left(\frac{M_2}{M_1}\right) f_{S_1}(S_1, \omega_1 | S_2, \omega_2)\,\mathrm{d}S_1 \tag{2.25}$$

$$(M_2/2 < M_1 < M_2,\ t_1 < t_2),$$

where we have used equations (2.11) and (2.16). Integration of equation (2.25) over the range $M_2/2 < M_1 < M_2$ gives the probability that halo M_2 had a parent in this mass range at t_1, which equals the probability that its formation time was earlier than this:

$$P(t_\mathrm{f} < t_1 | M_2, t_2) = P(M_1 > M_2/2, t_1 | M_2, t_2) = \int_{S_2}^{S_\mathrm{h}} \frac{M(S_2)}{M(S_1)} f_{S_1}(S_1, \omega_1 | S_2, \omega_2)\,\mathrm{d}S_1, \tag{2.26}$$

where $S_\mathrm{h} = S(M_2/2)$ as before. This then gives the distribution of formation times. Note that this equation differs from equation (2.23) in containing the weighting factor $M_2/M_1 > 1$ in the integral, which biases the distribution towards earlier formation times.

It is useful to consider this result in more detail for the case of a self-similar model, with $S \propto M^{-\alpha}$. For easier comparison with the results of the next section, we set $t_2 = t_0$ and $M_2 = M_0$. First, we rewrite equation (2.26) in terms of the variables

$$\tilde{S} \equiv (S - S_0)/(S_\mathrm{h} - S_0), \tag{2.27}$$
$$\tilde{\omega} \equiv (\omega - \omega_0)/\sqrt{(S_\mathrm{h} - S_0)},$$

giving

$$P(<t_\mathrm{f}) = P(>\tilde{\omega}_\mathrm{f}) = \int_0^1 \mu(\tilde{S}; S_0, S_\mathrm{h}) K(\tilde{S}, \tilde{\omega}_\mathrm{f})\,\mathrm{d}\tilde{S}, \tag{2.28}$$

where $K(\Delta S, \Delta\omega)\,\mathrm{d}\Delta S$ gives the probability of a change ΔS in a step $\Delta\omega$, from equation (2.15),

$$K(\Delta S, \Delta\omega)\,\mathrm{d}\Delta S = \frac{1}{(2\pi)^{1/2}} \frac{\Delta\omega}{(\Delta S)^{3/2}} \exp\left[-\frac{(\Delta\omega)^2}{2\Delta S}\right] \mathrm{d}\Delta S \tag{2.29}$$

$$(\Delta\omega > 0, \Delta S > 0),$$

and $\mu(\tilde{S})$ is defined to be the function

$$\mu(\tilde{S}; S_0, S_\mathrm{h}) \equiv \frac{M(S_0)}{M(S)} = [1 + (2^\alpha - 1)\tilde{S}]^{1/\alpha}, \tag{2.30}$$

where the second equality only applies for the self-similar model. The reason for choosing these variables is that $P(>\tilde{\omega}_\mathrm{f})$ then depends on the power spectrum only through the function $\mu(\tilde{S})$ in the range $0 \le \tilde{S} \le 1$, which in turn depends only very weakly on α. The distribution of formation times, when expressed in terms of these scaled variables, is therefore almost identical for different values of the spectral index n in the range of physical interest, and, by extension, should be almost identical for any slowly varying power spectrum (such as CDM). Results for the distribution in $\tilde{\omega}_\mathrm{f}$ for $n = -2, -1, 0$ and 1 are shown in Fig. 7. For comparison, this figure also shows the distribution of mass-halving times for single trajectories from equation (2.23). The formation time t_f is then related to $\tilde{\omega}_\mathrm{f}$ by the equation

$$\delta_\mathrm{c}(t_\mathrm{f}) = \delta_\mathrm{c}(t_0) + \tilde{\omega}_\mathrm{f}\sqrt{\sigma^2(M_0/2) - \sigma^2(M_0)}, \tag{2.31}$$

which is easily solved by making use of the relations (2.1). For the specific case that $\Omega_0 = 1$ and $S \propto M^{-\alpha}$, this gives

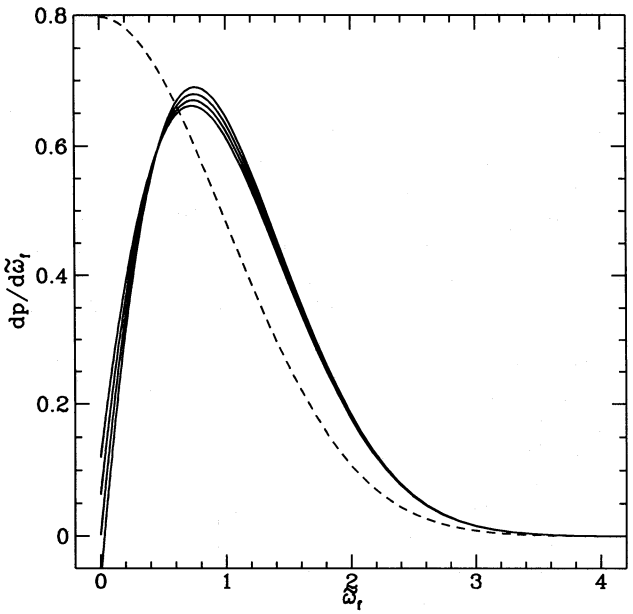

Figure 7. The distribution of halo formation epochs $\tilde{\omega}_\mathrm{f}$ calculated from equation (2.28), for power-law power spectra with $n = -2$, -1, 0 and 1 (solid curves, from bottom to top at the peak). The dashed curve shows the distribution of mass-halving times from equation (2.23). $\tilde{\omega}_\mathrm{f}$ is related to the formation time t_f by equation (2.31). For $\Omega_0 = 1$, $\tilde{\omega}_\mathrm{f} \propto z_\mathrm{f}$, so this is a scaled version of the distribution of formation redshifts.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

$$1+z_f=\left(\frac{t_0}{t_f}\right)^{2/3}=1+\tilde{\omega}_f\sqrt{2^\alpha-1}\left(\frac{M_0}{M_*}\right)^{-\alpha/2}\qquad(\Omega_0=1),\quad(2.32)$$

so that $z_f\propto\tilde{\omega}_f$ at fixed mass.

For the special case $n=0$ ($\alpha=1$), the distribution of $\tilde{\omega}_f$ can be obtained from equation (2.28) analytically:

$$\frac{dp}{d\tilde{\omega}_f}d\tilde{\omega}_f=-\frac{\partial}{\partial\tilde{\omega}_f}P(>\tilde{\omega}_f)\,d\tilde{\omega}_f=2\tilde{\omega}_f\,\mathrm{erfc}(\tilde{\omega}_f/\sqrt{2})\,d\tilde{\omega}_f\qquad(n=0).\quad(2.33)$$

For other values of n, it is calculated numerically.

There is a problem with the $\tilde{\omega}_f$-distributions shown in Fig. 7: for $n>0$, they go slightly negative close to $\tilde{\omega}_f=0$. For example, the $n=1$ curve is negative for $\tilde{\omega}_f<0.03$, falling to $dp/d\tilde{\omega}_f=-0.07$ at $\tilde{\omega}_f=0$. Clearly, this is unacceptable behaviour for a probability density. This inconsistency arises because our derivation involved counting haloes, which will only give consistent probabilities if particles that our analysis of trajectories tags as having halo mass M really are spatially grouped into objects of mass M. The peculiar behaviour found for $dp/d\tilde{\omega}_f$ presumably results from the fact that this identification is not exact. This problem did not arise in calculating the distribution of survival times, because that involved only counting trajectories, and did not explicitly require the 'tagged' mass to be the same as the true mass. In Section 3.2 we will present a calculation of formation times which avoids the pitfall of negative probabilities, based on Monte Carlo simulations of halo merger histories. Discussion of the results on formation times is postponed to that section.

3 MONTE CARLO GENERATION OF HALO MERGER HISTORIES

3.1 The method

In Section 2 we considered the trajectory of the overdensity $\delta(M)$ at a point as the smoothing mass M was varied. The mass of the halo containing the particle at time t was then identified as the largest mass at which the trajectory upcrossed through the barrier $\delta=\omega=\delta_c(t)$. By varying the barrier height ω with t, one can then generate the history of the way in which the halo mass for a particle varies with time, which represents the sequence of halo mergers for this particle. However, using the results of Section 2.3, we can generate random walks for the mass history $M(t)$ more directly, without having to generate random walks for $\delta(M)$ first. This we now describe.

Consider the situation where one starts with a particle which is in a halo of mass M_0 at time t_0. As before, it is more convenient to take $S(M)=\sigma^2(M)$ as our 'mass' variable, and $\omega(t)=\delta_c(t)$ as our 'time' variable. We now consider trajectories which run back in time from $(M_0,\ t_0)$, i.e. in the direction of *decreasing* M and t, or *increasing* S and ω. For a step of size $\Delta\omega>0$ in ω, the probability distribution for the change $\Delta S>0$ in S is given by the function $K(\Delta S,\Delta\omega)d(\Delta S)$ in equation (2.29). The new halo mass at time $t(\omega+\Delta\omega)<t(\omega)$ is then simply $M(S+\Delta S)<M(S)$. The change in mass $\Delta M=M(S)-M(S+\Delta S)$ (going forward in time) has occurred through mergers with one or more other haloes. By taking many steps drawn from this probability distribution, we generate a possible mass history for a particle in a halo of given present mass M_0. The set of all such histories represents both a sum over the different histories of particles in the same present halo, and a sum over all haloes of the same present mass. Some examples of such histories are shown in Fig. 8, for an $n=-2$ power spectrum in an $\Omega_0=1$ universe.

Some features of these random walks can be seen directly from equation (2.29). For $\Delta S\gg(\Delta\omega)^2$ we have $K(\Delta S,\Delta\omega)\propto\Delta\omega/(\Delta S)^{3/2}$, so the probability of a given change ΔS is proportional to the interval $\Delta\omega$, implying that we are in the regime where the change in S (or in mass) is due essentially to a single merger event which has only a small probability of occurring in that time interval. In the opposite regime, $\Delta S\ll(\Delta\omega)^2$, the change in S typically results from summing the effects of several mergers occurring in that time interval. If the fractional mass change is small, $\Delta M/M\ll1$, then $\Delta S\propto\Delta M$, so that the merger probability in the single-merger regime varies approximately as $p(\Delta M,\ \Delta t)d(\Delta M)\propto\Delta t/(\Delta M)^{3/2}d(\Delta M)$. Thus, in the merger history of an average particle, there has been an effectively infinite number of mergers with other haloes of infinitesimal mass ($\Delta M\to0$), but the net increase in mass has been dominated by mergers with more massive haloes, with the amount of mass accreted in haloes with mass $\sim\Delta M$ scaling roughly as $(\Delta M)^{1/2}$, up to $\Delta M\sim M$. These results closely reflect those derived for accretion rates from equation (2.18). The choice of step-size $\Delta\omega$ in the random walks determines the mass resolution: if we want to resolve single mergers of mass $\Delta M_c\ll M$, then we require $(\Delta\omega)^2\lesssim|d\ln\sigma^2/d\ln M|(\Delta M_c/M)S$.

When we trace the history of a halo back in time, it splits into branches of successively smaller mass, representing a hierarchy of mergers. The mass histories generated in the way just described correspond to choosing randomly among

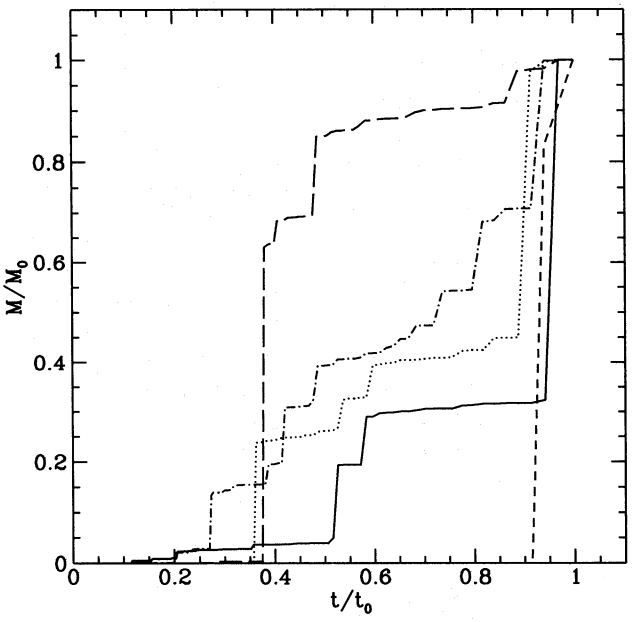

Figure 8. Random trajectories for the halo mass of a particle, generated using equation (2.29), for an $n=-2$ power spectrum in an $\Omega_0=1$ universe. All the trajectories are constrained to have the same halo mass M_0 at time t_0. The initial mass was chosen to be $M_0=M_*$, and a constant step-size $\Delta\omega=0.02$ was used.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

all the possible branches, with the probability of being on a particular branch being proportional to its mass. However, if we look at all the branches at a time $t_1 < t_0$, then one of them will have a larger mass than any of the others. At times when the largest mass branch exceeds $M_0/2$, we can define this branch to be the 'main trunk' of the merging tree. Moving down the merger tree, the main trunk grows continually by accreting smaller haloes, until it becomes the current halo at time t_0. The main trunk is the 'main parent' of Section 2.5. As in Section 2.5, we define the 'formation time' t_f of the current halo as the time when the main trunk first appeared, and its 'formation mass', $M_f \geq M_0/2$, as the mass of the main trunk at that time.

We can generate mass histories for the *main parent* halo in the interval $t_f \leq t \leq t_0$, $M_f \leq M \leq M_0$ by modifying the procedure described above with an extra assumption. Suppose we start on the main trunk. In a single step $\Delta\omega$, the halo mass falls from M to $M - \Delta M$, which represents a *splitting* of the halo into precursors of mass $M - \Delta M$ and ΔM. If $\Delta M > M/2$, then our trajectory has chosen the smaller precursor, which is a side branch off the main trunk. We should therefore discard this trajectory and choose another one. This is equivalent to choosing ΔS from the distribution of equation (2.29), but with the constraint $\Delta S < (\Delta S)_{max} \equiv S(M/2) - S(M)$, so that at each step we have $\Delta M < M/2$, and ΔM is the mass of the smaller halo accreted on to the main parent halo. $K(\Delta S, \Delta\omega)$ is then normalized over the range $0 < \Delta S < (\Delta S)_{max}$ by dividing by $\mathrm{erfc}[\Delta\omega/(2\Delta S_{max})]$.

This procedure would be exact if there were an exact correspondence between the halo mass we estimate for a particle by filtering on various scales, and its actual halo mass, but, since this correspondence is only approximate, alternative procedures for generating main-trunk trajectories can give somewhat different results. For instance, instead of restricting the range in ΔS in each step, we could generate steps ΔS without any restriction, and then choose the new mass to be the larger of ΔM and $M - \Delta M$. This gives results which are similar overall, but which differ in detail in their numerical values.

3.2 Halo formation times revisited

We thus generate 'main parent' trajectories $S(\omega)$, starting at (S_0, ω_0) and ending when the trajectory crosses $S = S_h \equiv S(M_0/2)$. The process of averaging over many such Monte Carlo merger histories provides an alternative estimate of the distribution of halo formation times to the analytical counting argument in Section 2.5. In carrying out the averaging, each trajectory is given a weight proportional to $1/M_f$, corresponding to the assignment of equal weight to each main trunk.

The distribution of formation epochs given by this method for power-law power spectra, $P(k) \propto k^n$, with $n = -2, -1, 0$ and 1, is shown in Fig. 9. This should be compared with Fig. 7, which shows the corresponding distributions calculated using the analytical argument. As in Section 2.5, it is convenient to use the scaled variables $\tilde{S}$ and $\tilde{\omega}$ defined in equation (2.27), in terms of which the random walks start at $\tilde{S} = 0$, $\tilde{\omega} = 0$ and end when the trajectory crosses $\tilde{S} = 1$. For a power-law power spectrum, $\Delta\tilde{S}_{max} = 1 + (2^\alpha - 1)\tilde{S}$, and $M/M_0 = [1 + (2^\alpha - 1)\tilde{S}]^{-1/\alpha}$, so that results are independent of the current mass M_0 when one works in terms of $\tilde{S}$, $\tilde{\omega}$ and

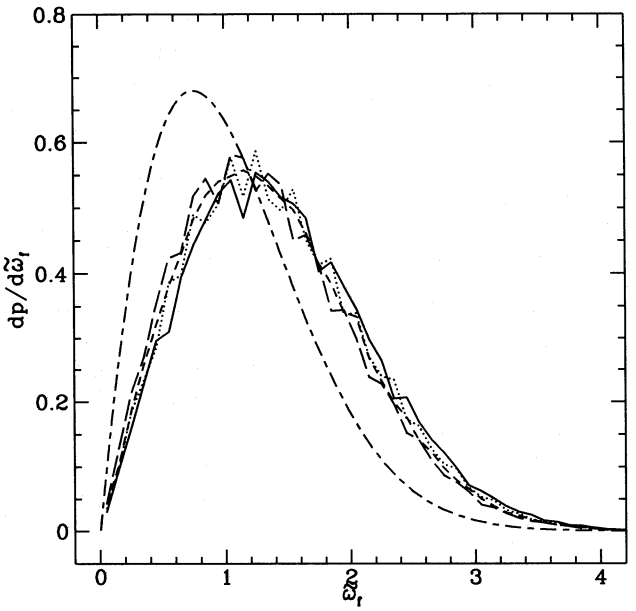

Figure 9. The distribution of 'formation epochs' $\tilde{\omega}_f$ of haloes with a given current mass, for power-law power spectra with $n = -2$ (solid curve), $n = -1$ (dotted), $n = 0$ (short-dashed) and $n = 1$ (long-dashed), based on Monte Carlo simulations of halo merger histories. $\tilde{\omega}_f$ is related to t_f by equation (2.31). For $\Omega_0 = 1$, $\tilde{\omega}_f \propto z_f$. The curves are based on 10^6 random trajectories with mass resolution $\Delta M/M_0 = 10^{-4}$. Also shown (smooth short-dashed–long-dashed curve) is the analytical distribution for $n = 0$, copied from Fig. 7.

M/M_0. In fact, as the figures show, the results in terms of these scaled variables are also very similar for different values of n, and, by extension, for any power spectrum that is slowly varying. The reason for the similarity is that the random walks only depend on the form of $\sigma(M)$ in the range $M_0/2 < M < M_0$, and this dependence is in lowest order absorbed into these variables. Recall that $\tilde{\omega}_f$ is related to the formation time by equation (2.31) or (2.32), so that, for $\Omega_0 = 1$, $z_f = \sqrt{2^\alpha - 1}(M_0/M_*)^{-\alpha/2}\tilde{\omega}_f$, and the $\tilde{\omega}_f$-distribution is just a linearly scaled version of the distribution of formation redshifts, with a scaling factor that depends on mass and on the shape of the power spectrum.

Comparing the analytical and Monte Carlo results in Figs 7 and 9, we see that, in both cases, the $\tilde{\omega}_f$-distributions rise roughly linearly at small $\tilde{\omega}_f$ and decline exponentially at large $\tilde{\omega}_f$, and are insensitive to the value of n. However, they differ somewhat in detail. For $n = 0$, the 10th, 50th and 90th percentiles of $\tilde{\omega}_f$ for the analytical distribution are 0.35, 0.97 and 1.89 respectively, while for the numerical distribution they are 0.54, 1.34 and 2.38: ~40 per cent larger. For other values of n, the comparison is similar. This provides some indication of the likely errors in our approach. Note that, by construction, the probability distributions generated by the Monte Carlo method are necessarily positive, unlike those generated by the analytical method.

As a test, we also computed the distribution of mass-halving times t_h that one obtains from Monte Carlo mass trajectories *without* the constraint $\Delta M < M/2$ at each step, i.e. not always following the main parent. These Monte Carlo results were identical to the analytical results from equation (2.23) (also shown in Fig. 7), as they should be.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

Whichever method one uses, the physical interpretation of these results is that, of the haloes existing at a given time, those of lower masses on average formed earlier than those of high masses, if $n > -3$, with the typical formation redshift scaling as $z_f \propto \sigma(M)$ for $\Omega_0 = 1$. This is analogous to the assumption, made in many simple studies of galaxy formation, that haloes of mass M typically collapse at redshift $1+z \approx \sigma(M)/\delta_{cr}$. In the latter case, however, one does not apply the constraint that the halo still exists at the present day.

We have also calculated the distribution of halo formation times predicted for a CDM power spectrum with $\Omega_0 = 1$, $H_0 = 50$ km s^{-1} Mpc^{-1} and $\sigma_8 = 0.5$. The results are shown in Fig. 10, this time in terms of physical variables. The median formation times for $M_0/(h^{-1}\,M_\odot) = (10^6, 10^9, 10^{12}, 10^{15})$ are $t_f/t_0 = (0.17, 0.22, 0.35, 0.71)$. Again, we see that low-mass haloes form significantly earlier than high-mass haloes, an effect which is likely to be important in models of galaxy formation.

3.3 Other results

Using the same Monte Carlo trajectories as for the formation times, with the same weighting, we can also compute the distribution of formation masses M_f, and the distribution of masses ΔM of the smaller haloes accreted on to the main parent since its mass exceeded M_f. The results for power-law power spectra are shown in Figs 11 and 12. Fig. 11 shows that most haloes have formation masses M_f only slightly above the minimum value $M_0/2$. For $n = 0$, the 10th, 50th and 90th percentiles of the distribution are $M_f/M_0 = (0.50, 0.53, 0.67)$, and the results are very similar for other values of n. Most of the results are not therefore very sensitive to the $1/M_f$ weighting of the trajectories. The fraction of mass accreted in haloes of various masses ΔM since time t_f, shown in Fig. 12 (where $f_M(<\Delta M)$ is the cumulative mass fraction accreted in haloes with masses $<\Delta M$), shows the expected

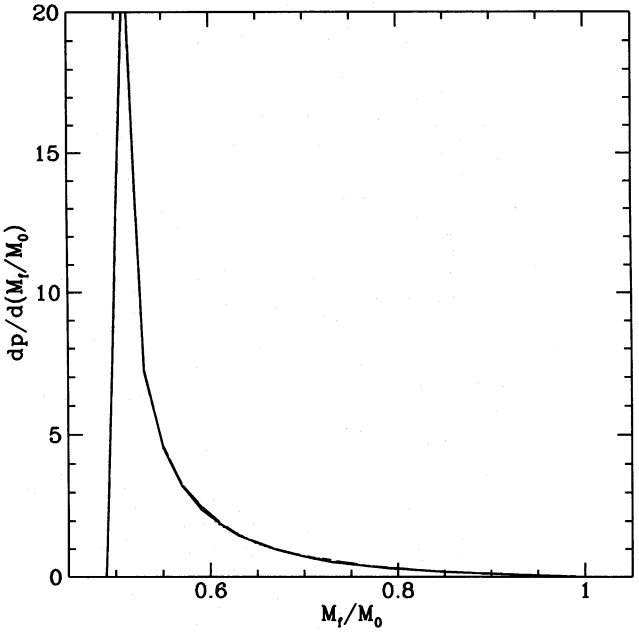

Figure 11. The distribution of 'formation masses' M_f of haloes with a given current mass M_0, for power-law power spectra with $n = -2$ (solid curve), $n = -1$ (dotted), $n = 0$ (short-dashed) and $n = 1$ (long-dashed), based on Monte Carlo simulations of halo merger histories. The parameters are the same as in Fig. 9.

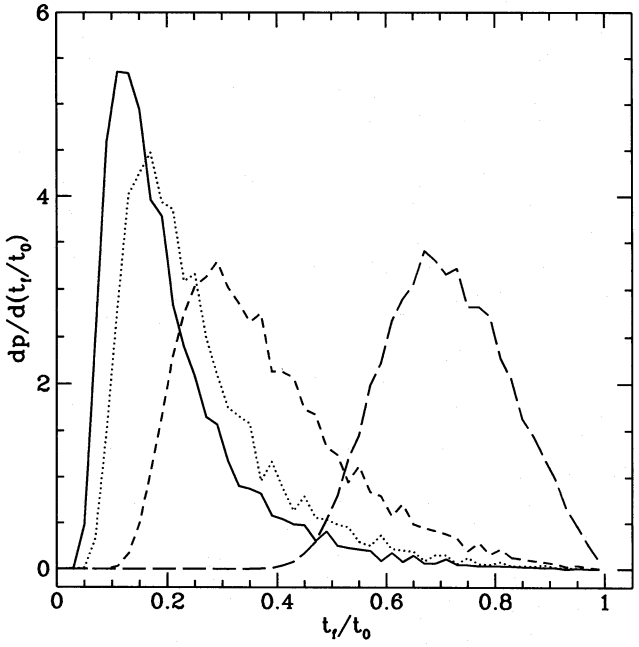

Figure 10. The distribution of formation times t_f of haloes in a CDM universe with $\Omega_0 = 1$, $H_0 = 50$ km s^{-1} Mpc^{-1} and $\sigma_8 = 0.5$. The solid, dotted, short-dashed and long-dashed curves are for present halo masses $M_0/(h^{-1}\,M_\odot) = (10^6, 10^9, 10^{12}, 10^{15})$ respectively. The curves are based on 10^6 random trajectories with mass resolution $\Delta M/M_0 = 10^{-4}$.

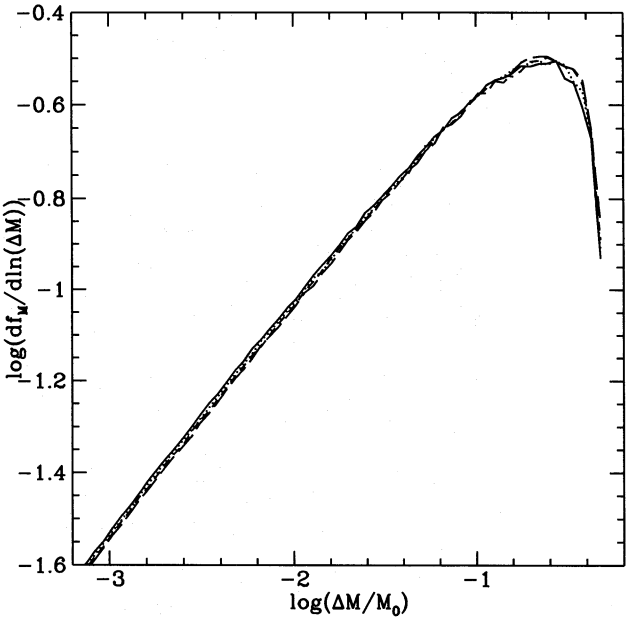

Figure 12. The fraction of mass accreted in haloes of mass ΔM by haloes of given current mass M_0, since time t_f, for power-law power spectra with $n = -2$ (solid curve), $n = -1$ (dotted), $n = 0$ (short-dashed) and $n = 1$ (long-dashed). Curves are obtained from Monte Carlo simulations of halo merger histories, with the parameters the same as in Fig. 9.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

behaviour $df_M/d\ln \Delta M \propto (\Delta M)^{1/2}$ at low masses, peaking at $\Delta M/M_0 \approx 1/4$, and falling to zero at $\Delta M/M_0 = 1/2$. The (10th, 50th, 90th) percentiles for the mass accretion, for $n=0$, are $\Delta M/M_0 = (3.9\times10^{-3}, 8.1\times10^{-2}, 3.2\times10^{-1})$.

One can generalize the procedure described above to generate a complete 'merger tree', describing the merger histories of all of the small haloes which are the building blocks for a given current halo. For each step back in time, a branch of this tree forks into two smaller branches. To follow the history of the main parent when $M > M_0/2$, we chose the branch with the larger mass at each fork. We can, however, generate a history for the branch with the smaller mass, by starting a trajectory at $(\Delta M, t)$, using the same equation (2.29) to generate steps, and so on, as this branch in turn splits into smaller branches. This procedure is only approximate, as it ignores correlations between different branches, which are likely to exist at some level. The total number of branches is infinite, but if one only follows branches down to some minimum halo mass $M_{\rm min}$ then the number becomes finite. This procedure is similar to that in the 'block model' described by Cole & Kaiser (1988) and Cole (1989, 1991), but has the advantage that, in the block model, each step corresponds to a factor of 2 increase in halo mass, while for the procedure described here the mass resolution can be made as small as one pleases, subject to limitations of computer time. These merger trees can be used to study the evolution of the baryonic components of galaxies through the combined effects of cooling of the gas and merging of the haloes; this has already been done by Cole (1991) using the block model. We plan to return to this avenue of study in a future paper.

4 APPLICATIONS

4.1 Merging galaxy clusters

We can apply our results on halo merging directly to groups and clusters of galaxies, insofar as the galaxies or X-ray emitting intergalactic gas trace the distribution of dark matter in these systems. Here we will attempt to derive a constraint on the density parameter Ω_0 by calculating the fraction of massive haloes that have undergone recent mergers and comparing this number to the fraction of rich clusters that observations suggest have recently formed by merging.

The observational evidence that suggests that a rich cluster has recently merged with another system is the presence of substructure in the density distribution. Such substructure is expected to be erased by violent relaxation on a time-scale of order the cluster's dynamical time and so its presence indicates that the merger occurred very recently. In studies of rich clusters, a significant fraction of clusters have been found to exhibit low-contrast substructure. Geller & Beers (1982) found that about 40 per cent of rich clusters have more than 1 maximum in their galactic surface density distribution. Dressler & Shectman (1988) concluded that 30–40 per cent of clusters exhibit significant substructure when analysed using a combination of position and velocity information. More recently, Forman & Jones (1990) and Jones & Forman (1992) have classified a sample of 208 X-ray bright clusters according to the structure of their *Einstein* X-ray surface brightness maps. About 20 per cent of their sample show 'double' or 'complex' structure, and we will adopt this number as a fairly conservative lower bound on the fraction of clusters that have recently formed by merging. The time interval during which these mergers occurred should be roughly the past dynamical time, but is hard to estimate accurately. In order to display this uncertainty, we have chosen simply to adopt the values of $0.2\,t_0$ and $0.5\,t_0$, where t_0 is the current age of the Universe, as estimates of the relaxation time required for the substructure to be erased. For comparison, for a truncated isothermal sphere with a mean density given by equation (A15), the orbital period of a particle on a circular orbit at the half-mass radius is $0.5\,t_0$, while the time for 1 cycle in radius for a radial orbit with apocentre at the half-mass radius is about $0.2\,t_0$.

We now make use of the distribution of lifetimes (equation 2.26) that we estimated analytically, in order to predict, as a function of the density parameter Ω_0, the fraction of rich clusters that have formed recently by merging. By requiring this fraction to be at least 20 per cent we will set a lower limit on Ω_0. Before we can employ the formula (2.26) to compare models with different Ω_0, we must decide how to set the amplitude of the density fluctuations in each model. We have chosen to do this by using the datum that the observed number density of rich clusters (richness class $R>1$) is $n_{\rm rich} \sim 6.0\times10^{-6}\ h^3\ {\rm Mpc}^{-3}$ (e.g. Batuski et al. 1989), and also that their typical velocity dispersion is approximately 760 km s^{-1} (Struble & Rood 1987). Taking this one-dimensional velocity dispersion to imply a circular velocity $V_c = (GM/r)^{1/2} = \sqrt{2}\times 760$ km s^{-1} and using the halo overdensity given by the spherical collapse model detailed in Appendix A, we deduce a cluster mass of $M_{\rm rich} = 3.1\times10^{14}\ h^{-1}\ {\rm M}_\odot$ for $\Omega_0 = 1$, increasing to $M_{\rm rich} = 4.2\times10^{14}\ h^{-1}\ {\rm M}_\odot$ for $\Omega_0 = 0.1$. We then normalize the power spectrum of density fluctuations by requiring that the number density predicted by equation (2.11),

$$n_{\rm rich} = \int_{M_{\rm rich}}^{\infty} \frac{{\rm d}n}{{\rm d}M}(M, t_0)\,{\rm d}M, \tag{4.1}$$

be equal to the observed number density. For low values of the normalization parameter σ_8 (the rms fluctuation of the mass contained within spheres of radius $8\ h^{-1}$ Mpc), $n_{\rm rich}$ is very small, because very few haloes of mass greater than $M_{\rm rich}$ exist. As one increases σ_8, more haloes with $M > M_{\rm rich}$ form and $n_{\rm rich}$ increases; then as σ_8 is increased still further these haloes merge together and $n_{\rm rich}$ begins to decrease. Here we adopt the lower value of σ_8 for which $n_{\rm rich}$ equals the observed abundance. For the CDM power spectrum normalized in this way we find $\sigma_8 = 0.5$ for $\Omega_0 = 1$, rising to $\sigma_8 = 2.2$ for $\Omega_0 = 0.1$. It is interesting to note that for sufficiently low Ω_0 the maximum value $n_{\rm rich}$ as one varies σ_8 will be less than the observed cluster abundance. In fact, for both the CDM and $n=-2$ spectra we find that the correct cluster abundance cannot be reproduced for $\Omega_0 \lesssim 0.1$.

We can now use equation (2.26) to estimate the fraction, F_s, of these clusters that formed, i.e. assembled at least half of their mass, in the past $0.2\,t_0$ and $0.5\,t_0$ yr. It is this fraction which we expect to exhibit measurable substructure. Figs 13(a) and (b) show this prediction for CDM and for an $n=-2$ spectrum respectively. On the mass scale of rich clusters, the CDM spectrum of fluctuations has an effective slope of $n=-1.1$ for $\Omega_0 = 1$, falling to $n=-1.8$ for

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

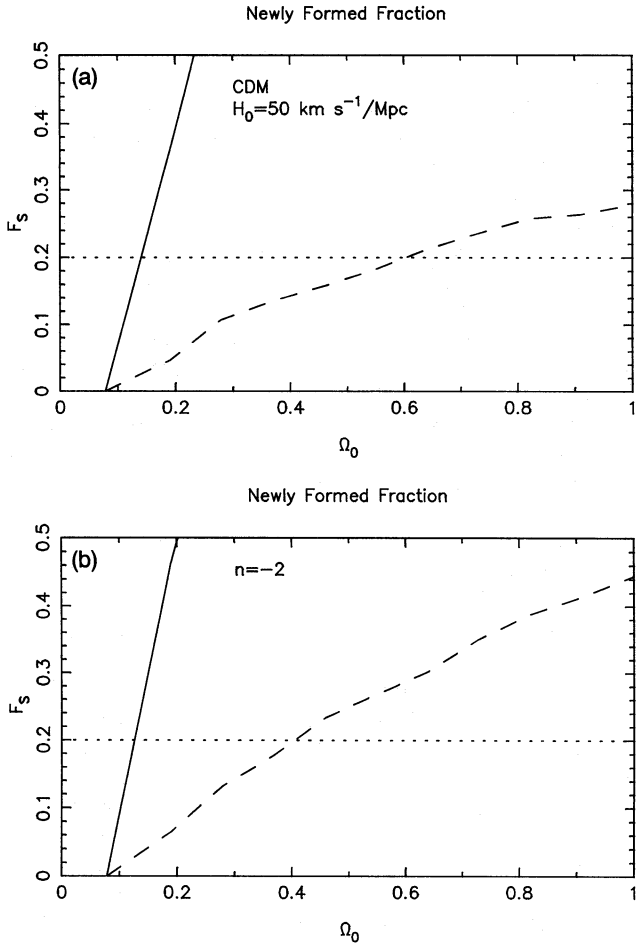

Figure 13. In both (a) and (b), the solid and broken curves indicate the fraction of rich clusters that have accreted more than 50 per cent of their mass in the past $0.5\,t_0$ and $0.2\,t_0$ yr respectively. Here t_0 is the present age of the Universe. For the purposes of these figures, a rich cluster has been defined to be any halo with a one-dimensional velocity dispersion greater than 760 km s^{-1}. In (a) the CDM spectrum of density fluctuations appropriate for $H_0 = 50$ km s^{-1} Mpc^{-1} and for each choice of Ω_0 was adopted, while in (b) an $n = -2$ power spectrum was used. The amplitude of the fluctuations is set so that for each value of Ω_0 the number density of rich clusters equals $6 \times 10^{-6}\ h^3$ Mpc^{-3}. The horizontal line at $F_s = 20$ per cent indicates an estimate of the fraction of rich clusters that are observed to have major substructure (see text). If the presence of substructure in this 20 per cent of clusters implies that they are younger than $0.2\,t_0$, then one concludes that Ω_0 is greater than 0.6 for the CDM spectrum or 0.4 for the $n = -2$ spectrum.

$\Omega_0 = 0.1$, and thus it is to be expected that the $n = -2$ model will evolve more rapidly than the CDM model and consequently have a larger fraction of young clusters. However, this dependence is relatively weak, and for the shorter estimate of the relaxation time we deduce that the limits on the density parameter are $\Omega_0 \gtrsim 0.6$ and 0.4 for the CDM and $n = -2$ models respectively. The longer estimate of the relaxation time results in the much weaker limit $\Omega_0 \gtrsim 0.1$ for both models. If we had used the Monte Carlo results of Section 3 for the distribution of formation times, we would have obtained somewhat more stringent (i.e. larger) lower bands on Ω_0. Clearly, the limit that one infers on Ω_0 depends crucially on the time-scale over which substructure persists. Thus, before this calculation can be used to determine a robust estimate of Ω_0, N-body simulations will have to be used to investigate how long merger-induced substructure persists.

These limits are comparable to the limit $\Omega_0 \gtrsim 0.5$ deduced from the same data by Richstone, Loeb & Turner (1992). However, their analysis ignored the mass distribution of the clusters and the influence of the spectral index on the rate at which the cluster population evolves. Their analysis also differed from ours in adopting a larger mass for a typical rich cluster of $1.0 \times 10^{15}\ h^{-1}\ \mathrm{M}_\odot$, and in assuming that substructure would be erased in a relaxation time of only $0.1/H_0$. If we adopt these more extreme numbers, then even in a flat $\Omega_0 = 1$ universe only 11 per cent of rich clusters would be expected to exhibit substructure.

4.2 Accretion of baryonic cores in dark haloes, and merging of luminous galaxies

We have so far been considering the merging of dark matter haloes. This does not, however, directly tell us about the merging of visible galaxies, which consist of cores of baryonic material (stars + gas) sitting within these dark haloes. In the standard picture (White & Rees 1978), these cores form when gas is able to cool within a dark halo and condense to the centre. However, when haloes merge, the baryonic cores they contain, being more compact, may avoid merging with each other, and end up orbiting within the new combined halo, so that a halo formed by many mergers may contain many distinct baryonic cores. For baryonic cores that are composed mainly of stars, the results on N-body simulations of mergers between spherical galaxies without haloes summarized by Aarseth & Fall (1980) indicate that two cores will merge at the pericentre of their relative orbit only if their separation and relative velocity there satisfy $R_{\mathrm{peri}} \lesssim R_*$ and $V_{\mathrm{peri}} \lesssim V_*$, where R_* and V_* are the characteristic radius and internal velocity of the stellar cores. The conditions for the merging of baryonic cores that are mainly gaseous have not been investigated, but presumably will also require that the cores overlap at pericentre. In either case, if the orbits of the cores within the combined halo do not initially satisfy these conditions, then merging of the cores will only occur after dynamical friction or gas-dynamical drag has eroded the orbits and brought the cores close together. This process may be slow compared to the rate at which haloes are built up by hierarchical merging. This appears to be the case in clusters of galaxies, where many luminous galaxies orbit within a single cluster halo.

A full calculation of the rate of merging of luminous galaxies would require the following of the hierarchy of halo merging through several stages, starting with the haloes in which the baryons originally condensed into cores. This is quite complicated, and is postponed to a future paper. Here we present a simpler calculation, of a large halo accreting smaller haloes, in which each small halo of mass ΔM is *assumed* to contain all of its baryons in a *single* core of mass $\Delta M_{\mathrm{b}} = f_{\mathrm{b}} \Delta M$. We use the halo mass trajectories described in Section 3 to follow the merger history of a halo from the time t_{f} when it had half or more of its current mass M_0, up to the present time t_0. After the merging of each of the accreted haloes with the main halo, the baryonic core which the

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

accreted halo contained is assumed to orbit inside the new combined halo, with dynamical friction against the dark matter background gradually dragging it into the centre. This calculation actually gives an upper limit to the accretion of baryonic cores by a halo, since, if the baryons in the accreted halo are distributed in several cores rather than one, the dynamical friction time for each core will be larger.

To compute the dynamical friction rate, we assume that the halo can be modelled as a singular isothermal sphere with radius R_H and circular velocity V_c. We assume that the main halo of mass M formed by merging with a satellite of mass ΔM at time t_{mg} has a mean density $3\pi/(Gt_{mg}^2)$, as given by the spherical collapse model described in Appendix A, so that $R_H = V_c t_{mg}/(2\pi)$. The dynamical friction time T_{df} is then given by equation (B4), derived in Appendix B, with satellite mass $M_S = f_b \Delta M$:

$$T_{df} \approx \frac{\epsilon^{0.78}}{0.855(2\pi) f_b \ln\Lambda} \left(\frac{r_{ci}}{R_H}\right)^2 \left(\frac{M}{\Delta M}\right) t_{mg} \qquad (\epsilon \gtrsim 10^{-2}), \quad (4.2)$$

so that the baryonic core sinks to the centre of the halo at time $t_{df} = t_{mg} + T_{df}$. (The effect of subsequent mergers on the halo structure is ignored in this estimate.) In the above formula, we take $\ln\Lambda = \ln[M/(f_b \Delta M)]$, r_{ci} is the radius of the circular orbit with the same energy as the initial satellite orbit, and the quantity $0 \le \epsilon \le 1$ measures the eccentricity of this orbit. ϵ is defined as the ratio of the angular momentum to that for a circular orbit of the same energy, and is equal to 1 for a circular orbit and 0 for a radial orbit. Equation (4.2) provides a good fit to the eccentricity dependence over the range $10^{-2} \lesssim \epsilon \lesssim 1$. We see that eccentric orbits decay much faster than circular ones.

The appropriate values to use for r_{ci} and ϵ are somewhat uncertain, although they could be estimated from N-body simulations. We will assume $r_{ci}/R_H = 1$ for the initial radius, i.e. assume that the satellite core starts at the edge of the main halo, and consider two cases for the initial eccentricity: $\epsilon = 1$, corresponding to circular orbits, and $\epsilon = 0.2$, corresponding to very elongated orbits with pericentre-to-apocentre distance ratio $r_{min}/r_{max} = 0.05$. For a small halo falling into a much larger one, the assumption that the satellite starts near the edge seems a reasonable one. For mergers of comparable-mass objects, the dynamical friction time becomes comparable to the orbital time, so there may be no clear separation between the halo merging and dynamical friction processes.

With the above assumptions, we can compute the current average rate at which baryonic lumps of mass ΔM_b are being accreted to the centres of haloes of mass M_0 by averaging over a small time interval around t_0. The results, for a CDM spectrum with $\Omega_0 = 1$, $\sigma_8 = 0.5$, $h = 0.5$, and a baryon fraction $f_b = 0.1$, are shown in Fig. 14 for $\epsilon = (1, 0.2)$ and

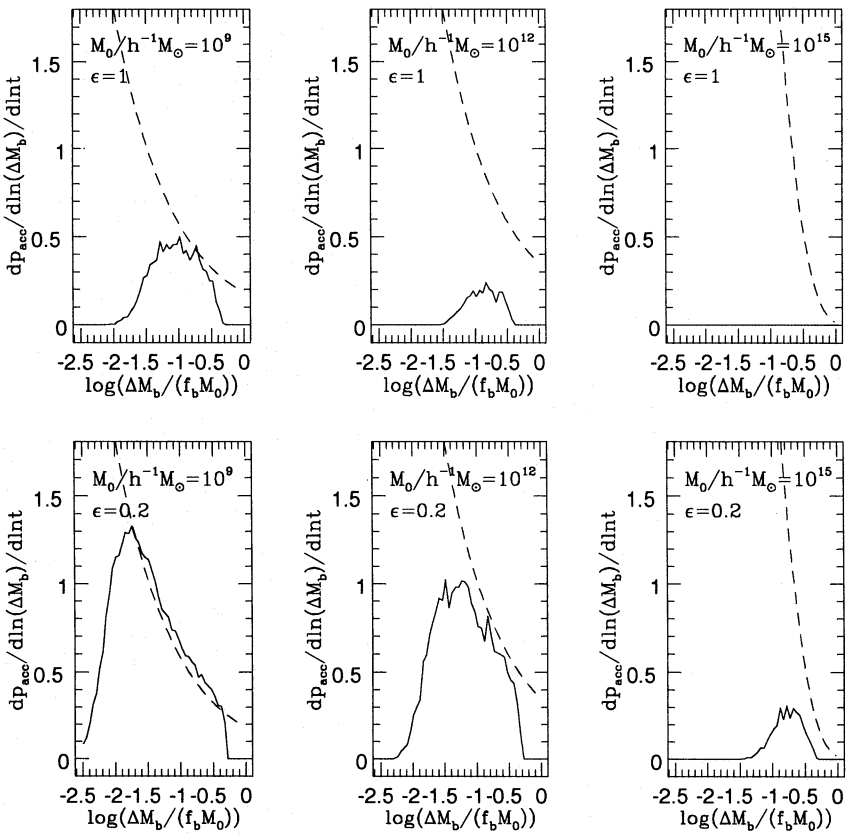

Figure 14. The present rate of accretion of baryonic cores to the centres of haloes by dynamical friction, for an $\Omega_0 = 1$ CDM universe with $\sigma_8 = 0.5$, $h = 0.5$ and $f_b = 0.1$. The solid lines show the accretion rates of cores of mass ΔM_b obtained from Monte Carlo simulations with 10^5 trajectories, based on the assumption that there is only 1 baryonic core in each accreted halo. The dashed curves show the rate of accretion of haloes of mass $\Delta M = \Delta M_b/f_b$ derived from equation (2.18). From left to right, the panels show results for $M_0/h^{-1}\,M_\odot = 10^9, 10^{12}, 10^{15}$, with $\epsilon = 1$ (circular orbits) for the top row and $\epsilon = 0.2$ (eccentric orbits) in the bottom row.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

$M_0/h^{-1}\ \mathrm{M}_\odot = (10^9, 10^{12}, 10^{15})$. Also shown in the same plots, by a dashed line, is the current halo merger rate at the corresponding mass, calculated from equation (2.18). There are several features to be noted. The first is that the accretion rates for baryonic cores all cut off above a mass $\Delta M_\mathrm{b} = (1/2) f_\mathrm{b} M_0$. This is a simple consequence of the fact that the baryonic cores currently being accreted are associated with *past* halo mergers, and by construction these have $\Delta M \leq (1/2) M_0$. The second point is that the accretion rates for the cores all fall to zero at low masses, even though the corresponding halo merger rates continue to rise [as $(\Delta M)^{-1/2}$]. This is because the dynamical friction time scales as $T_\mathrm{df} \propto (\Delta M/M)^{-1}$, which for low masses ΔM becomes larger than the time for which the halo has existed. Thus dynamical friction favours accretion of higher mass cores. The third point is that the rate of accretion of baryonic cores is a larger fraction of the halo merger rate for low halo mass M_0 and low values of ϵ. Recalling that lower mass haloes on average formed earlier than high-mass haloes (e.g. see Fig. 10), we see that this results from the dynamical friction time-scale T_df being a smaller fraction of the time $t_0 - t_\mathrm{f}$ for which the halo has existed. When this ratio becomes sufficiently small, the time lag produced by dynamical friction between merging of the haloes and accretion of the baryonic core becomes unimportant, so the accretion rate of cores at ΔM_b becomes equal to the halo merger rate at $\Delta M = \Delta M_\mathrm{b}/f_\mathrm{b}$, over some range of ΔM, as can be seen in the plot for $M_0 = 10^9\ h^{-1}\ \mathrm{M}_\odot$, $\epsilon = 0.2$. In fact, since in our calculation the largest possible halo mass which can be accreted in a single merger is $\Delta M/M \leq 1/2$, haloes that form too late do not have time to accrete any cores by dynamical friction; using equation (4.2), the condition $t_\mathrm{mg} + T_\mathrm{df} < t_0$ translates to $t_\mathrm{f} < t_\mathrm{mg} < 0.45\, t_0$ for $\epsilon = 1$ and $t_\mathrm{f} < t_\mathrm{mg} < 0.74\, t_0$ for $\epsilon = 0.2$. The typical $\Delta M/M$ of accreted haloes is well below the maximum value of 1/2, and these mergers can happen at any time between t_f and t_0, so that most of the baryonic cores may not be accreted even if the above conditions on t_f are met. This explains why the accretion rates for cores are so small for $M_0 = 10^{15}\ h^{-1}\ \mathrm{M}_\odot$ haloes in Fig. 14.

We cannot reach any firm conclusions about the merging of luminous galaxies from the above calculation, because we do not know to what degree it is true that the accreted haloes and the main halo contain all their baryons in single cores when they merge. If we none the less assume this to be the case, and identify typical luminous galaxies with haloes of mass $M_0 \sim 10^{12}\ h^{-1}\ \mathrm{M}_\odot$, then, for circular starting orbits, we infer from Fig. 14 a merger rate ~ 0.2 per $\ln \Delta M_\mathrm{b}$ per $\ln t$ between comparable-mass galaxies, less by a factor of ~ 3 than the corresponding halo merger rate. This is similar to the merger rate ~ 0.1 per Hubble time estimated by Toomre (1977). On the other hand, for eccentric orbits, the inferred merger rate for luminous galaxies would be similar to that for the corresponding haloes, and would be rather high compared to this observational estimate.

Recently, Navarro & White (in preparation) have performed a series of full *N*-body/hydrodynamical simulations of the hierarchical formation of individual galaxies. The calculation we have performed here is particularly relevant to their simulations, as the dynamic range of their calculation means they are unable to resolve substructure in the infalling baryonic core and so, like us, are forced to assume that infalling haloes have only one core. We note that our assumptions about the range of initial angular momenta and the fates of the infalling satellites are in qualitative agreement with their results. Navarro & White find that the merging of the dark matter haloes precedes the merging of baryonic cores by anything from 1 to 10 dynamical times, and that the factor that determines how many orbits an infalling satellite makes before spiralling in and merging is its initial orbital angular momentum.

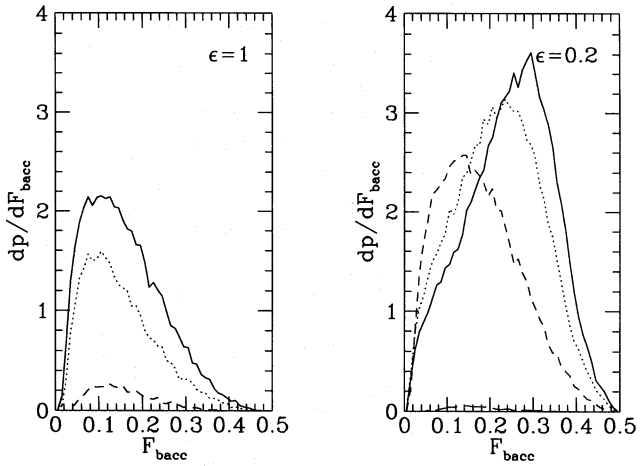

Figure 15. The fraction of baryons accreted in cores by dynamical friction since halo formation, for haloes of a given current mass M_0. The distributions are computed for the same model as in Fig. 14. The solid, dotted, short-dashed and long-dashed lines are for $M_0/h^{-1}\ \mathrm{M}_\odot = (10^6, 10^9, 10^{12}, 10^{15})$ respectively. The left panel is for $\epsilon = 1$, and the right for $\epsilon = 0.2$. Note that the probability distributions have a delta-function contribution at $F_\mathrm{bacc} = 0$ which is not shown in the figures.

Another quantity that we can compute from the dynamical friction calculation is the fraction F_bacc of all the baryons in the final halo of mass M_0 that have been accreted to the centre in baryonic cores since the halo formation time. The results are shown in Fig. 15, for the same parameters as in Fig. 14, for current masses $M_0/h^{-1}\ \mathrm{M}_\odot = (10^6, 10^9, 10^{12}, 10^{15})$. For the cases where the amount of accretion is small, the mean value of F_bacc obtained from Fig. 15 is significantly less than that which would be obtained from Fig. 14 by integrating over mass and multiplying by t, because the accretion rate of cores is increasing with time at t_0. The results for the fraction of baryons accreted are interesting because of the problem of the survival of thin galactic discs. If a spiral galaxy accretes a satellite galaxy, then as the satellite sinks into the disc it will cause heating of the stars, causing the disc to thicken. Toth & Ostriker (1992) calculated that a typical spiral disc cannot have accreted more than 10 per cent of its mass in small satellites since the time when most of the stars formed, without making discs thicker than is observed. They argued that such a low accretion fraction for most galaxies is a problem for $\Omega_0 = 1$ cosmologies in which structure forms by hierarchical clustering. Our results for haloes of mass $M_0 \sim 10^{12}\ h^{-1}\ \mathrm{M}_\odot$ are that, for $\epsilon = 1$, 96 per cent will have accreted no baryonic cores at all since their formation times, while for $\epsilon = 0.2$ this fraction is reduced to 37 per cent. While it is not clear that this result can be carried over directly to accretion of satellites by spiral galaxy discs (for the same reasons as given above), it does

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

suggest that the thinness of discs may not be a problem, provided that the orbits of accreted satellites are not too eccentric.

5 COMPARISON WITH PREVIOUS WORK

Carlberg (1990a) derived analytical expressions for the merger rates both of dark haloes and of the luminous galaxies within them. Our results do not agree with his. Consider first the halo merger rate, for which our result is formula (2.18), giving the rate as a function of the masses, M and ΔM, of both haloes involved. Carlberg considered two different ways to estimate a net merger rate for haloes of roughly equal mass M, based on the Press–Schechter mass function. His equation (3) is simply the rate of change of the number density of haloes of mass M. Since the total number of haloes of mass M can be both depleted by mergers producing more massive haloes and replenished by mergers of lower mass haloes, this rate cannot be equated directly with the merger rate at a particular mass. In fact, it yields a negative rate in the regime where the number of haloes of mass M is increasing with time. This undesirable property led Carlberg to adopt the rate of change of his equation (4) as a better expression for the merger rate. The assumption he made to derive his formula was that any increase in the fraction of mass locked up in haloes of mass greater than M is the result of mergers involving haloes in the mass range $M/2$ to M. This assumption need not be satisfied. It is quite possible for this mass fraction to increase due to objects of mass greater than M accreting low-mass haloes with masses much less than M. Typically, the mass function of haloes is broad, which can result in a large fraction of this mass increase being due to the accretion of such low-mass haloes. Thus mergers between haloes with masses outside the range $M/2$ to M can contribute to an increase in the mass fraction in haloes of mass greater than M. This process would then cause Carlberg's expression to overestimate the true merger rate.

We will now compare Carlberg's merger rate directly with the merger rate we have calculated. Carlberg calculated a rate at which mergers are occurring between haloes in the mass range $M/2$ to M, and then divided by an expression for the number density of haloes in this same mass range, to get a rate per halo. Carlberg's result can be rewritten as a rate per halo per Hubble time,

$$R_{\rm mg}(M/2<M'<M,t) \equiv \frac{1}{n_{\rm H}(M/2<M'<M,t)}\, t\frac{{\rm d}n_{\rm mg}(M/2<M'<M)}{{\rm d}t} \qquad (5.1)$$

$$= \left|\frac{{\rm d}\ln\delta_{\rm c}}{{\rm d}\ln t}\right| \Big/ \left|\frac{{\rm d}\ln\sigma}{{\rm d}\ln M}\right|.$$

This is equivalent to equation (6) of Carlberg (1990a), when we omit his factor representing the probability for the luminous components within a halo to merge, except that we have generalized his expression to $\Omega_0 \neq 1$. Thus Carlberg predicted a merger rate per halo which depends on mass only through the spectral index of the density fluctuations.

We can calculate the quantity $R_{\rm mg}$ defined in equation (5.1) directly using our formula (2.18) for the merger rate and equation (2.11) for the number of haloes as a function of mass. The expression we require is

$$R_{\rm mg}(M/2<M'<M,t) = \frac{1}{\left[\int_{M_1=M/2}^{M_1=M}\frac{{\rm d}n}{{\rm d}M_1}(M_1,t)\,{\rm d}M_1\right]}\,\frac{1}{2}\int_{M_1=M/2}^{M_1=M}\frac{{\rm d}n}{{\rm d}M_1}(M_1,t) \times \int_{\Delta M=M/2}^{\Delta M=M} t\frac{{\rm d}^2p}{{\rm d}\ln\Delta M\,{\rm d}t}(M_1\rightarrow M_1+\Delta M|t)\frac{{\rm d}\Delta M}{\Delta M}\,{\rm d}M_1. \qquad (5.2)$$

The integrals over ΔM and M_1 are both over the range $M/2$ to M, so that we count only mergers between pairs of haloes in this mass range, and the factor of $1/2$ is included so that we count the number of pairs of merging haloes, as does Carlberg. In Fig. 16, we compare these rates for the cases of $\Omega_0 = 1$ and 0.2 and scale-free initial conditions with $n = -2$ and -1. Note that our equation (2.18) gives the merger rate for all values of ΔM, while Carlberg *assumed* that all mergers have $M/2 < \Delta M < M$. We see from Fig. 16 that, for objects of a fixed mass M, the merger rate is low at early times, and then rises as these objects become more numerous. The rate then slowly declines again as the mass of the objects becomes small compared to the growing characteristic mass $M_*(t)$. Carlberg's expression for the merger rate is too large by a factor of 10 or more, even compared to the peak value given by our formula. It overestimates the merger rate by an even larger factor at high redshift for fixed M, and for halo masses large compared to $M_*(t)$.

Comparison of the merger rates for different values of Ω_0 in Fig. 16 shows that the current merger rate is lower for the $\Omega_0 = 0.2$ model than for the $\Omega_0 = 1$ model, by about a factor of 2, when both are normalized to have the same value of M_* at $z = 0$. The shape of the redshift dependence is different for the two values of Ω_0, because the past evolution of the characteristic mass $M_*(t)$ is different. Carlberg proposed using the *slope* of the dependence of merger rate on redshift in the range $0 < z < 0.5$ as a means of estimating both Ω_0 (Carlberg 1990b) and the cosmological constant Λ_0 (Carlberg 1991). Carlberg was considering the merging of *luminous galaxies* for this test. We would only remark that, for merging of the *dark* haloes, the slope of the merger rate near $z = 0$ does not appear to be especially sensitive to Ω_0, unless one is considering masses $M \gg M_*(t_0)$.

Carlberg also computed an expression for the merger rate of luminous galaxies, by multiplying the halo merger rate by a factor $P_{\rm mg}$ representing the probability for the luminous cores to have a relative velocity small enough to allow merging. He implicitly assumed that the cores start out on orbits that lead them to collide within 1 orbit ($R_{\rm peri} \lesssim R_*$), which need not be the case. Thus, in Carlberg's picture, merging of the cores occurs either immediately after merging of the haloes, or not at all. Carlberg did not allow for the possibility that orbital decay by dynamical friction may lead to merging of the cores after some time delay. Further, in computing $P_{\rm mg}$, Carlberg equated the distribution of pericentric relative velocities of cores to the relative velocity distribution averaged over all

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

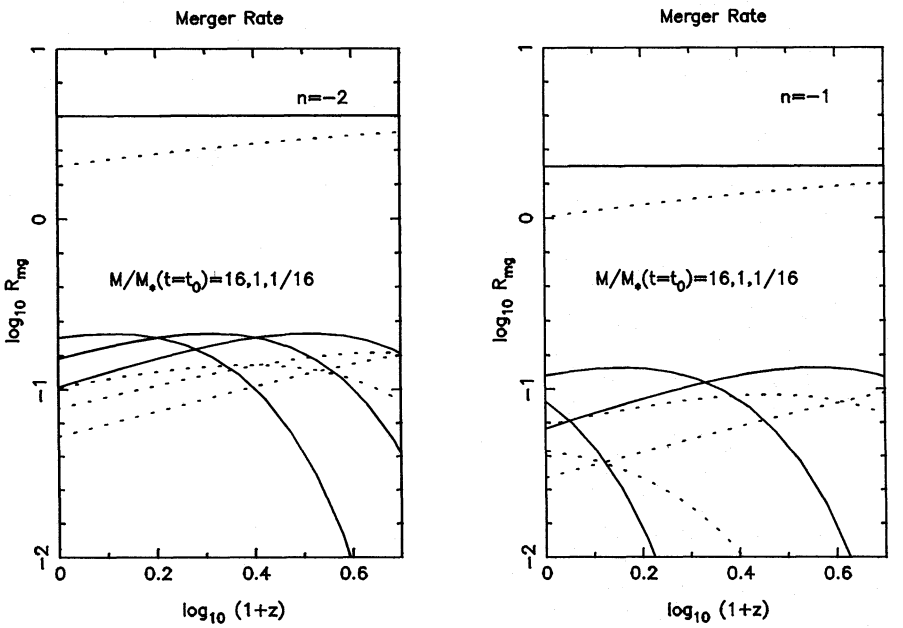

Figure 16. The fractional merger rate R_{mg} as a function of redshift, z. R_{mg} is defined as the number of mergers between haloes in the mass range $M/2$ to M per Hubble time at the specified redshift, relative to the number density of haloes in the same mass range. The solid lines show results for $\Omega_0 = 1$, and dotted lines for $\Omega_0 = 0.2$, for scale-free initial conditions with spectral index $n = -2$ in the left panel and $n = -1$ in the right panel. The two sets of three curves at the bottom of each figure are our predictions, for the masses indicated, given by integrating equation (2.18) and averaging over haloes in the mass range $M/2$ to M. The merger rates that peak at the highest redshift correspond to the lowest mass haloes. The masses are given in units of $M_*(t = t_0)$, the characteristic mass at the present epoch. The pairs of curves at the top of each figure show the rates given by the formula of Carlberg (1990a), which do not depend on mass for scale-free power spectra.

particles of a certain separation (~ 100 kpc). This neglects the fact that pericentric velocities will be different in haloes of different masses. The typical relative velocity that Carlberg inferred is then quite large compared to the internal velocities of galaxies, leading to a small P_{mg}, and a large suppression of the rate of merging of luminous galaxies relative to that of their haloes. This tends to compensate for his overestimate of the halo merger rate. Carlberg's calculation also takes no account of the fact that two haloes that merge may each contain multiple luminous galaxies.

6 CONCLUSIONS

We have derived an analytical expression for the merger rate of virialized haloes (equation 2.18), which is applicable to any hierarchical model in which structure grows via gravitational instability. This formula is based on the reformulated and extended Press–Schechter theory presented by Bond et al. (1991), and quantifies how the merger rate depends on halo mass, epoch, the initial spectrum of density fluctuations and on the density parameter Ω_0. In all cases, mergers with very tiny haloes dominate in number, but the increase in mass by merging is dominated by infall into larger haloes when the halo mass is small, and by accretion of smaller haloes when the halo mass is large. We have also derived expressions which give estimates of the formation and survival times of haloes, these being defined respectively as the time when half of the halo's mass was assembled, and the time when the halo will double in mass by merging. These quantities are of great importance when constructing realistic models of galaxy formation, as they define the time-span over which gas bound to the halo is able to cool, condense and perhaps be converted into stars. On average, the low-mass haloes existing at a given time are found to have formed earlier than the high-mass haloes. High-mass haloes typically survive for less than a few times the age of the Universe at that time, while low-mass haloes cover a wider range of survival times, up to quite high values. These results reflect the fact that, at any time, the higher mass haloes are tending to be built up by merging, while the low-mass haloes are disappearing by merging into these larger systems.

The validity of these formulae relies on equating of the mass of the halo within which a particle is found with the mass of the largest spherical region centred on the particle whose mean linear overdensity, calculated using the window function (2.14), is greater than the threshold value δ_c. The correspondence between these two masses is far from perfect, as is illustrated in figs 6 and 7 of BCEK. However, at least over the limited range of halo masses that has been probed in N-body simulations, the formulae for the mass function (2.11) and the conditional distribution (2.15) have been found to agree remarkably well with the results of simulations. We therefore expect similar agreement with simulations for the formulae derived here, at least for masses comparable to or greater than M^*. In our next paper, we will make a more detailed comparison of these results with N-body simulations. The particular choice of window function we make is motivated by analytical convenience – with sharp k-space filtering, the trajectory of mean overdensity at a point as a function of filtering mass is a Brownian random walk, resulting in analytically tractable expressions for the mass function, merger rates, and other properties.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

We have also presented a Monte Carlo method, based on our expression for the merger rate, that enables representative merger histories to be generated for haloes of varying mass. Being able to construct a complete description of the formation path of a given halo is often the most direct and simple way to address complicated problems. For instance, given a set of rules for star formation, one could use this method to determine where and when the stars form, and where they end up. We expect this technique to be of great value when attempting to model galaxy formation, and in studying the evolution of the galaxy luminosity function. In this paper, we have employed this method to obtain an alternative estimate of the distribution of halo formation times, which is in reasonable agreement with that found analytically, and of the distribution of the masses of the haloes accreted since halo formation.

We have applied our formalism to the merging of galaxy clusters. The large fraction of clusters observed to have substructure indicates recent formation by merging. This seems to require a fairly high-density Universe, with $\Omega_0 \gtrsim 0.5$, if this substructure lasts for less than $0.2\,t_0$, with t_0 being the present age of the Universe, but only requires $\Omega_0 \gtrsim 0.1$ if the substructure lasts as long as $0.5\,t_0$.

Calculation of the merger rate of luminous galaxies is a non-trivial extension of the calculation of the merger rate of their dark haloes. When the haloes merge, the baryonic cores comprising the luminous galaxies are left on orbits in the new merged halo, and can merge with each other only if their separations and relative velocities are reduced to small values by dynamical friction. A completely self-consistent calculation would require the generation of the complete hierarchy of mergers leading to a given present-day halo, and the application of the dynamical friction criterion at each stage to see which cores merge. In this paper, we have carried out only a more limited, exploratory calculation. We used the Monte Carlo method to generate histories of the accretion of small haloes, each assumed to contain a single baryonic core, by a large halo, and used an estimate of the time-scale for dynamical friction to erode the orbits of the baryonic cores to estimate the rate of accretion of baryonic lumps on to a central galaxy. This calculation indicates that, if following halo merging the cores start off on nearly circular orbits near the edge of the new combined halo, dynamical friction can slow the baryonic accretion rate down to a small fraction of the halo merger rate, at least for higher halo masses. The accretion of smaller baryonic cores is also effectively suppressed. However, if the initial orbits of the baryonic cores are very eccentric, then dynamical friction acts faster, and the rate of accretion of baryonic cores is closer to the halo merger rate. For halo masses around $10^{12}\,M_\odot$, with a CDM power spectrum with $\Omega_0 = 1$, and nearly circular orbits, this calculation suggests that a central galaxy will in most cases accrete very little mass in baryonic cores. This may resolve the problem raised by Toth & Ostriker (1992) of how the discs of spiral galaxies remain thin, despite merging of haloes. We will make a more detailed study of the merging of luminous galaxies in a future paper.

ACKNOWLEDGMENT

CGL is supported by a SERC Advanced Fellowship.

REFERENCES

Aarseth S. J., Fall S. M., 1980, ApJ, 236, 43
Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15
Batuski D. J., Bahcall N. A., Olowin R. P., Burns J. O., 1989, ApJ, 341, 599
Binney J., Tremaine S., 1987, Galactic Dynamics. Princeton Univ. Press, Princeton, NJ
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440 (BCEK)
Bower R. J., 1991, MNRAS, 248, 332
Broadhurst T. J., Ellis R. S., Glazebrook K., 1992, Nat, 355, 55
Carlberg R. G., 1990a, ApJ, 350, 505
Carlberg R. G., 1990b, ApJ, 359, L1
Carlberg R. G., 1991, ApJ, 375, 429
Cen R. Y., Ostriker J. R., Spergel D. N., Turok N., 1991, ApJ, 383, 1
Chandrasekhar S., 1943, Rev. Mod. Phys., 15, 2; reprinted, 1954, in Wax N., ed., Selected Papers on Noise and Stochastic Processes. Dover, New York, p. 3
Cole S., 1989, PhD thesis, Univ. Cambridge
Cole S., 1991, ApJ, 367, 45
Cole D., Kaiser N., 1988, MNRAS, 233, 637
Dressler A., Shectman S. A., 1988, AJ, 95, 985
Evrard A. E., 1988, MNRAS, 235, 911
Forman W., Jones C., 1990, in Oegerle W. R., Fitchett M. J., Danly L., eds, Clusters of Galaxies. Cambridge Univ. Press, Cambridge, p. 257
Geller M. J., Beers T. C., 1982, PASP, 94, 421
Guiderdoni B., Rocca-Volmerange B., 1991, A&A, 252, 435
Heckman T. M., Smith E. P., Baum S. A., van Breugel W. J. M., Miley G. K., Illingworth G. D., Bothun G. D., Balick B., 1986, ApJ, 311, 526
Jones C., Forman W., 1992, in Fabian A. C., ed., Proc. NATO-ASI Series, Clusters and Superclusters of Galaxies. Kluwer, Dordrecht, p. 49
Navarro J. F., Benz W., 1991, ApJ, 380, 320
Peacock J. A., Heavens A. F., 1990, MNRAS, 243, 133
Peebles P. J. E., 1980, The Large-Scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Press W. H., Schechter P., 1974, ApJ, 187, 425
Rees M. J., Ostriker J. P., 1977, MNRAS, 179, 541
Richstone D., Loeb A., Turner E. L., 1992, ApJ, 393, 477
Rocca-Volmerange B., Guiderdoni B., 1990, MNRAS, 247, 166
Sanders D. B., Soifer B. T., Elias J. H., Madore B. F., Matthews K., Neugebauer G., Scoville N. Z., 1988, ApJ, 325, 74
Struble M. F., Rood H. J., 1987, ApJS, 63, 543
Toomre A., 1977, in Tinsley B. M., Larson R. B., eds, The Evolution of Galaxies and Stellar Populations. Yale Univ. Observatory, New Haven, Connecticut, p. 401
Toth G., Ostriker J. P., 1992, ApJ, 389, 5
White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

APPENDIX A: CRITICAL OVERDENSITY FOR COLLAPSE FOR $\Omega_0 < 1$

In order to calculate the critical linear theory overdensity $\delta_c(t_{coll})$ corresponding to the time at which a region collapses, let us consider evolution of a uniformly overdense spherical region embedded in an open Friedman–Robertson–Walker universe.

In an open universe, the radius a of an unperturbed spherical region can be shown to evolve with time according to the pair of parametric equations

$$a = A(\cosh\eta - 1), \qquad t = B(\sinh\eta - \eta) \tag{A1}$$

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

(see e.g. Peebles 1980, section 19), where the constants A and B are related by

$$A^3=(4\pi/3)\,G\rho_0 a_0^3 B^2. \tag{A2}$$

The values of the constants A and B are determined by the present value of the Hubble constant $H_0=(\dot{a}/a)_0$, the density parameter $\Omega_0=8\pi G\rho_0/(3H_0^2)$, and by choosing the present value of the radius a_0. Together these constraints imply

$$\eta_0=\cosh^{-1}(2/\Omega_0-1), \tag{A3}$$
$$B=1/2H_0^{-1}\Omega_0(1-\Omega_0)^{-3/2}.$$

The expansion and eventual recollapse of a perturbed overdense region containing the same mass M as the region of the background universe with which it is being compared are described by the parametric equations

$$a_p=A_p(1-\cos\theta), \qquad t=B_p(\theta-\sin\theta), \tag{A4}$$

where the constants A_p and B_p are related to A and B by the constraint

$$\frac{A_p^3}{B_p^2}=\frac{A^3}{B^2}=GM. \tag{A5}$$

(For a general spherical perturbation, t in equation A4 is replaced by $t-T_p$; the choice $T_p=0$ corresponds to the selection of a pure growing-mode perturbation.) As time progresses, the perturbed region expands less rapidly than the background universe, and a density contrast given by

$$1+\delta=a^3/a_p^3 \tag{A6}$$

develops, where $\delta=\Delta\rho/\rho=(\rho_p-\rho)/\rho$. The region collapses to a singularity, $a_p=0$, when $\theta=2\pi$, which by (A4) occurs at

$$t_{\rm coll}=2\pi B_p. \tag{A7}$$

We must now match the solutions for the perturbed region and the background by considering their early evolution in the limit of η, $\theta\to 0$. In this limit, equations (A1) and (A4) can be written as the Taylor expansions

$$a=A\left(\frac{\eta^2}{2!}+\frac{\eta^2}{4!}+\ldots\right), \qquad t=B\left(\frac{\eta^3}{3!}+\frac{\eta^5}{5!}+\ldots\right) \tag{A8}$$

and

$$a_p=A_p\left(\frac{\theta^2}{2!}-\frac{\theta^4}{4!}+\ldots\right), \qquad t=B_p\left(\frac{\theta^3}{3!}-\frac{\theta^5}{5!}+\ldots\right). \tag{A9}$$

Eliminating η and θ, we find

$$a=A\left(\frac{t}{B}\right)^{2/3}\frac{6^{2/3}}{20}\left[1+\frac{6^{2/3}}{20}\left(\frac{t}{B}\right)^{2/3}+\ldots\right] \tag{A10}$$

and

$$a_p=A_p\left(\frac{t}{B_p}\right)^{2/3}\frac{6^{2/3}}{20}\left[1-\frac{6^{2/3}}{20}\left(\frac{t}{B_p}\right)^{2/3}+\ldots\right]. \tag{A11}$$

Substitution of these two expressions into equation (A6) and use of equation (A5) give, to leading order,

$$\delta=\frac{3\times 6^{2/3}}{20}\left[\left(\frac{1}{B_p}\right)^{2/3}+\left(\frac{1}{B}\right)^{2/3}\right]t^{2/3} \qquad (t\to 0)$$
$$=\frac{3(12\pi)^{2/3}}{20}\left[\left(\frac{1}{t_{\rm coll}}\right)^{2/3}+\left(\frac{1}{t_\Omega}\right)^{2/3}\right]t^{2/3}. \tag{A12}$$

In the second line, we have used equation (A7) to eliminate B_p, and defined $t_\Omega=2\pi B=\pi H_0^{-1}\Omega_0(1-\Omega_0)^{-3/2}$ to eliminate B. Equation (A12) shows that, at early times, the density perturbation grows as $\delta\propto t^{2/3}$, which is the linear growing-mode solution. At later times, $t\gtrsim t_\Omega$, the linear perturbation behaviour departs from this; the exact linear solution is $\delta\propto D(t)$, where

$$D(t)=\frac{3\sinh\eta(\sinh\eta-\eta)}{(\cosh\eta-1)^2}-2 \tag{A13}$$

(see e.g. Peebles 1980, section 11), which varies as $D(t)\approx(12\pi)^{2/3}/10\,(t/t_\Omega)^{2/3}$ for $t\ll t_\Omega$, and $D(t)\to 1$ for $t\gg t_\Omega$. Comparing equations (A12) and (A13), we find that the exact linear behaviour for a spherical perturbation which collapses at time $t_{\rm coll}$ is

$$\delta=\frac{3}{2}D(t)\left[1+\left(\frac{t_\Omega}{t_{\rm coll}}\right)^{2/3}\right]. \tag{A14}$$

The setting of $t=t_0$ in the above expression then gives the extrapolated linear overdensity at time t_0 for a perturbation which collapses at time $t_{\rm coll}$.

When the spherical perturbation collapses, we assume that it reaches virial equilibrium at the time $t_{\rm coll}$ when formally $a_p\to 0$, at a radius which is half of its radius at maximum expansion ($\theta=\pi$). Using equations (A1)–(A6), the ratio of the halo density to the background density at the virialization time can be shown to be

$$\left(\frac{\rho}{\bar{\rho}}\right)_{\rm vir}=\left(\frac{2\pi}{\sinh\eta_{\rm coll}-\eta_{\rm coll}}\right)^2(\cosh\eta_{\rm coll}-1)^3, \tag{A15}$$

where $\eta_{\rm coll}$ is given by solving equation (A1) with $t=t_{\rm coll}$. On the other hand, comparison of the halo density to the critical density $\rho_c=3H(t)^2/(8\pi G)$ at virialization gives

$$\left(\frac{\rho}{\rho_c}\right)_{\rm vir}=8\pi^2\left[\frac{(\cosh\eta_{\rm coll}-1)^2}{\sinh\eta_{\rm coll}(\sinh\eta_{\rm coll}-\eta_{\rm coll})}\right]^2. \tag{A16}$$

The right-hand side of equation (A16) has the limiting values $18\pi^2$ for $\eta_{\rm coll}\ll 1\,(t_{\rm coll}\ll t_\Omega)$ and $8\pi^2$ for $\eta_{\rm coll}\gg 1\,(t_{\rm coll}\gg t_\Omega)$. Thus the halo density at collapse is always of order 100–200 times the critical density.

APPENDIX B: DYNAMICAL FRICTION TIME IN AN ISOTHERMAL HALO

We model each dark matter halo as a singular isothermal sphere, with circular velocity V_c and density $\rho_H=V_c^2/(4\pi Gr^2)$, truncated at a radius R_H. Within the truncation radius, the 1D velocity dispersion is taken to have a constant value $\sigma=V_c/\sqrt{2}$. For a satellite of mass M_S orbiting within the halo, the force exerted on it due to dynamical friction against the background halo particles is

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System

1993MNRAS.262..627L

$$\boldsymbol{F}_{\rm df} = -4\pi G^2 \ln\Lambda \rho_{\rm H} M_{\rm S}^2 B(v/\sqrt{2}\sigma)\frac{\boldsymbol{v}}{v^3} \tag{B1}$$

(e.g. Binney & Tremaine 1987, section 7.1), where $\boldsymbol{v}$ is the satellite velocity

$$B(X) \equiv \mathrm{erf}(X) - \frac{2X}{\sqrt{\pi}}\mathrm{e}^{-X^2}$$

and $\ln\Lambda \approx \ln(rv^2/GM_{\rm S})$ is the usual Coulomb logarithm (treating the satellite as a point mass).

The decay of the satellite orbit due to dynamical friction was calculated in the orbit-average approximation. The satellite orbit is described by E and J, the energy and angular momentum per unit mass, in terms of which the radial and tangential components of the velocity are given by

$$v_r = \sqrt{2(E-\Phi(r)) - J^2/r^2}, \tag{B2}$$

$$v_\theta = J/r,$$

where Φ is the gravitational potential. The instantaneous rates of change of E and J are $\dot{E} = -v|F_{\rm df}|/M_{\rm S}$ and $\dot{J} = -(rv_\theta/v)|F_{\rm df}|/M_{\rm S}$. The orbit-averaged rates of change are given by

$$\langle \dot{Q} \rangle = \left(\frac{1}{\int_{r_{\rm min}}^{r_{\rm max}} \frac{\mathrm{d}r}{v_r}} \right) \int_{r_{\rm min}}^{r_{\rm max}} \dot{Q} \frac{\mathrm{d}r}{v_r}, \tag{B3}$$

where Q is E or J, and $r_{\rm min}$ and $r_{\rm max}$ are the radial turning points of the orbit, given by solving equation (B2) for $v_r = 0$. The orbit-averaged equations $\mathrm{d}E/\mathrm{d}t = \langle \dot{E} \rangle$ and $\mathrm{d}J/\mathrm{d}t = \langle \dot{J} \rangle$ are then integrated to find the time $T_{\rm df}$ for the satellite orbit to shrink to $r = 0$, in terms of the initial values of E and J. This is a reasonable approximation provided that $T_{\rm df}$ is much greater than the time for a single orbit. The value of $\ln\Lambda$ is taken to be constant and equal to its initial value.

For a singular isothermal sphere, the dynamical friction time can be written as

$$T_{\rm df} = \frac{f(\epsilon) V_{\rm c} r_{\rm ci}^2}{2GM_{\rm S} B(1) \ln\Lambda} \tag{B4}$$

$$= \frac{f(\epsilon)}{2B(1)\ln\Lambda} \left(\frac{r_{\rm ci}}{R_{\rm H}}\right)^2 \left(\frac{R_{\rm H}}{V_{\rm c}}\right) \left(\frac{M_{\rm H}}{M_{\rm S}}\right).$$

Here, $r_{\rm ci}(E)$ is the radius of a circular orbit in the halo with the same energy E as the actual orbit, while the 'circularity' $\epsilon \equiv J/J_{\rm c}(E)$ is the angular momentum relative to that for a circular orbit with the same energy. Thus $0 \le \epsilon \le 1$, with $\epsilon = 1$ for a circular orbit and $\epsilon = 0$ for a radial orbit. For a circular orbit, $f(\epsilon) = 1$. For eccentric orbits, the function $f(\epsilon)$ was found by numerical integration of the orbit-averaged equations: for the range $10^{-2} \le \epsilon \le 1$, the function was found to be fitted to an accuracy of better than 3 per cent by the expression $f(\epsilon) \approx \epsilon^{0.78}$. We take $\ln\Lambda \approx \ln(M_{\rm H}/M_{\rm S})$.

Downloaded from https://academic.oup.com/mnras/article/262/3/627/1022789 by guest on 07 February 2022

© Royal Astronomical Society • Provided by the NASA Astrophysics Data System