Mon. Not. R. Astron. Soc. **364,** 823–832 (2005) doi:10.1111/j.1365-2966.2005.09602.x

Fast identification of bound structures in large *N*-body simulations

J. Weller,[1⋆] J. P. Ostriker,[1,2] P. Bode[2] and L. Shaw[1]
[1]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA*
[2]*Princeton University Observatory, Princeton NJ 08544-1001, USA*

Accepted 2005 September 9. Received 2005 September 1; in original form 2004 May 27

ABSTRACT

We present an algorithm that is designed to allow the efficient identification and preliminary dynamical analysis of thousands of structures and substructures in large *N*-body simulations. First, we utilize a refined density gradient system (based on DENMAX) to identify the structures and then apply an iterative approximate method to identify unbound particles, allowing fast calculation of bound substructures. After producing a catalogue of separate energetically bound substructures, we check to see which of these are energetically bound to adjacent substructures. For such bound complex subhaloes, we combine components and check if additional free particles are also bound to the union, repeating the process iteratively until no further changes are found. Thus, our subhaloes can contain more than one density maximum, but the scheme is stable: starting with a small smoothing length initially produces small structures that must be combined later and starting with a large smoothing length produces large structures within which sub-substructure is found. We apply this algorithm to three simulations. Two that are using the TPM algorithm by Bode, Ostriker & Xu and one on a simulated halo by Diemand, Moore & Stadel. For all these haloes, we find about 5–8 per cent of the mass in substructures.

Key words: methods: *N*-body simulations – methods: numerical – galaxies: clusters: general – galaxies: haloes – dark matter.

1 INTRODUCTION

Until recently, observational extragalactic astronomy has been based primarily on the study of galaxies and clusters of galaxies. The theoretical constructs in the standard Λ cold dark matter (CDM) paradigm for structure formation that are most closely associated with these phenomena are 'haloes' of dark matter and the 'subhaloes' within them. In this bottom-up picture, all self-gravitating virialized objects comprise accumulated smaller objects and these latter objects, hierarchically, comprise still smaller ones *ad infinitum*, assembled through 'merger trees'. Thus, a close examination of any representative object should show the undigested remnant cores of previously ingested objects, tidal streamers of debris shredded from the outer parts of these same subhaloes and the relatively smooth background material, which contains the somewhat phase-mixed accumulation of all the digested tidal effluvia. A closer and closer analysis in phase space would allow identification of components added at earlier and earlier times.

Thus, 'identification of substructure', even if perfect tools were available, requires some intellectual precision in the dynamical definitions of what is meant by 'subhaloes'. Until recently, the lack of sufficiently accurate computations made this issue moot, but now investigators have begun this analysis, using a variety of defined terms. We will provide our own definitions later in this section.

Historically, it was impossible to produce galaxy-size haloes in dense clusters with dark matter simulations (White 1976; van Kampen 1995; Summers, Davis & Evrard 1995; Moore, Katz & Lake 1996). This was mainly due to the limited mass and force resolution of the simulations used and was commonly known as the *overmerging problem*. The major causes of this problem were premature tidal disruption due to inadequate force resolution and two-particle evaporation for haloes with a small number of particles (Klypin et al. 1999). However, the combination of an increase in computing power and the invention of more efficient algorithms has led to promising developments over recent years that have overcome the numerical problems (Ghigna et al. 1998; Klypin et al. 1999; Moore et al. 1999; Okamoto & Habe 1999; Bode, Ostriker & Xu 2000; Ghigna et al. 2000; Springel et al. 2001; Kravtsov et al. 2004; De Lucia et al. 2004). Besides the numerical insufficiencies that can destroy substructures, there are also physical reasons for the destruction of structure, which are tightly connected to the numerical problems. First, there is dynamical friction, which drives the subhalo to the halo centre where it can be disrupted and merge with the central object. Secondly, there is tidal stripping when the tidal force from the halo on the subhalo is larger than the gravitational force holding the subhalo together. Thirdly, there may be shock heating which occurs during the close passage of two subhaloes and more dominantly on passing of a subhalo near the halo centre; this

⋆E-mail: jochen.weller@ucl.ac.uk

effect is believed to be less prominent than the first two (Moore et al. 1996; Gnedin, Hernquist & Ostriker 1999; Klypin et al. 1999).

Klypin et al. (1999) estimated that a force softening of $\epsilon = 3\,h^{-1}$ kpc and a mass resolution below $m_{\rm p} = 10^9\,h^{-1}\,{\rm M}_\odot$ would be sufficient to identify a substructure of mass $10^{11}\,h^{-1}\,{\rm M}_\odot$ with at least 30 particles at a distance $70\,h^{-1}$ kpc from the centre of a $10^{14}\,h^{-1}\,{\rm M}_\odot$ cluster. Needless to say, higher resolution would be even better. The usual approach to obtain such resolution is to take a cluster from a cosmological N-body simulation and resimulate it at higher resolution with inclusion of the long-distance (tidal) gravitational fields. However, if one wants to address the problem of substructure in a statistical and cosmological context, then one needs fairly large simulation boxes. Thus, with existing computing power, one cannot currently use much higher resolution than given above.

Our goal is to design algorithms that can be used to analyse structure/substructure in very large simulations such as 'light cone radians' of the Virgo group (Evrard et al. 2002) rather than individual very high resolution simulations of clusters. Besides the noted numerical difficulties, the identification of structures and substructures in large N-body simulations is a long-standing problem of principle. This has been addressed in the past by many different methods, mainly geometrical rather than physical (Huchra & Geller 1982; Davis et al. 1985; Bertschinger & Gelb 1991; Warren et al. 1992; Gelb & Bertschinger 1994; Lacey & Cole 1994; Stadel et al. 1997; Weinberg, Hernquist & Katz 1997; Eisenstein & Hut 1998; Klypin et al. 1999; Springel et al. 2001). Many methods exploit to some extent the Friends-Of-Friends (FOF; Huchra & Geller 1982; Davis et al. 1985; Lacey & Cole 1994) or the DENMAX (Bertschinger & Gelb 1991; Gelb & Bertschinger 1994; Eisenstein & Hut 1998) algorithm (described in Section 2), which are also at the centre of our method. What these methods have in common is that they are essentially geometrical and do not use the entire phase-space information, and hence they need post-processing to test for bound structures. In this paper, we discuss a fast approximate method to remove unbound particles from haloes.

Algorithms that have been used for finding bound structure include SKID (Stadel et al. 1997) and hierarchical adaptations of it, the Bound Density Maxima (BDM) scheme (BDM; Klypin et al. 1999), and SUBFIND (Springel et al. 2001). Quite recently, other methods have been introduced by Gill, Knebe & Gibson (2004), Kim & Park (2004) and Neyrinck, Gnedin & Hamilton (2005). SKID essentially uses the DENMAX algorithm to identify structures and then calculates bound structures by iteratively removing the unbound particle with the largest total energy until all particles are bound. The hierarchical scheme (Ghigna et al. 2000) uses SKID at three different smoothing lengths. The BDM scheme places spheres of a certain scale $r_{\rm sp}$ in the simulation box and then displaces the spheres to the centre of mass (CoM) of the particles inside the sphere. This process is iterated and eventually all maxima within a sphere of size $r_{\rm sp}$ are found. The unbinding is then done by calculating the escape velocity of the halo from the maximal circular velocity; all particles with velocities larger than the escape velocity are removed. For the calculation of the escape velocity, a Navarro–Frenk–White (NFW) density profile is assumed (Navarro, Frenk & White 1995). Recently BDM has been used to identify a vast number of haloes in a large detailed simulation (Kravtsov et al. 2004). The SUBFIND algorithm uses the FOF algorithm to find cluster-sized haloes and then looks for saddle points in the density field to identify subhaloes. Again, the particles of a subhalo are then examined to determine if they are bound. Recently, 11 resimulated clusters have been analysed in great detail with this method (De Lucia et al. 2004).

Most of the work quoted above used the resampling technique and consequently only analysed a small number of 'typical haloes' to high accuracy. Here, we take a complementary approach, by using simulations of volumes containing many target haloes. While sacrificing resolution (as compared with the resampling technique), we gain in sample size by a large factor, with thousands of haloes in our largest runs. In this paper, we will study two quite different simulations. One contains 256^3 particles in a volume $20\,h^{-1}$ Mpc on a side; the haloes from this run have masses typical of large galaxies. This run is discussed in more detail in Bode, Ostriker & Turok (2001); it was evolved with a P^3M code and halted at redshift $z = 1$. The second simulation is of 1024^3 particles with box size $320\,h^{-1}$ Mpc, containing many galaxy cluster sized haloes. This was evolved to $z = 0.05$ using the Tree-Particle-Mesh (TPM) algorithm (Bode & Ostriker 2003). The simulation parameters can be found in Table 1. One difference between the two codes used is that P^3M uses Plummer softening and TPM uses spline softening.

We will define a subhalo at any level of the hierarchy in the following fashion. In the CoM frame defined by the object in question, we take all particles as members that are gravitationally bound ($E < 0$). Thus, if a small smoothing length has been used to identify subclumps, we check if groups of these are bound to one another and if additional 'free' particles are bound to the assemblage. Conversely, if a larger smoothing length has been used to identify objects, we subsequently analyse these with greater refinement to ascertain subcomponents that in their own frames are self-gravitating. Thus, we produce a catalogue that provides labels for a hierarchy of bound objects, where the catalogue is, to a large extent, independent of the geometrical tools used to parse the entire object. We then make an independent catalogue of the hierarchy, where at each level we require all components to be gravitationally bound to the object to which they are attached.

The purpose of this paper is to clearly define the method and attempt to carefully specify the algorithms that define and identify substructure and to explain how seemingly minor variations in procedure can produce large changes in the final result.

2 THE METHOD

Before entering into the details of the method, we present a schematic overview of the substructure finding algorithm that we

Table 1. The top two rows give the parameters for the two simulations analysed in this paper, cold dark matter (CDM) universes with a cosmological constant. The last row gives the parameters for the resimulated cluster by Ghigna et al. (1998).

Model	z	$\Omega_{\rm b}$	$\Omega_{\rm c}$	Ω_Λ	$H_0\,(\frac{\rm km\,s^{-1}}{\rm Mpc})$	σ_8	n	N	$L\,(h^{-1}\,{\rm Mpc})$	$m_{\rm p}\,(h^{-1}\,{\rm M}_\odot)$	$\epsilon\,(h^{-1}\,{\rm kpc})$
ΛCDM	1.0	0.04	0.26	0.70	67	0.900	1.0	256^3	20	3.97×10^7	1.2
ΛCDM	0.05	0.04	0.26	0.70	70	0.975	1.0	1024^3	320	2.54×10^9	3.2
VIRGO			1.0	0.0	50	0.7		1314 161		8.6×10^8	5,10

will employ. We first apply the FOF method, which groups together large structures in a speedy way, on the entire simulation volume. At the core of our approach is the geometrically based DENMAX routine by Bertschinger & Gelb (1991), which moves particles up density gradients and identifies groups as all particles reaching the same density maximum. We run DENMAX with high resolution on each FOF halo. We then build a family tree and identify, with an iterative approximation scheme, energetically bound particles within the structures.

In this way, we create, hierarchically: (i) gravitationally bound objects ('mothers'); (ii) those substructures that lie within a given bound object ('daughters') and are themselves gravitationally bound; and (iii) further sublevels.

2.1 Creating the family trees

The first step to identify large groups in the simulation is the application of the FOF routine. We choose as a linking length $R_{\rm link} = 0.2\bar{n}^{-1/3}$, where $\bar{n}^{-1/3}$ is the mean interpartical separation. This ensures that we find clusters and also trace them out to the virial radius. Furthermore, this choice will select groups of particles with overdensities close to the value predicted by the spherical collapse model. With this linking length and a minimum number of 10 particles required to be identified as a group, FOF finds a large number of low-mass haloes and a decreasing number of more massive objects.

We estimate the density at each point of the simulation by measuring the weighted volume over the 16 nearest neighbours of each particle in the simulation (using the SMOOTH code; see http://www-hpcc.astro.washington.edu/tools). This enables us to estimate the position of the density peak of a halo. Also, the smallest rectangular box enclosing each halo is found. Next, we order the groups according to their mass and deploy a bottom-up scheme for identifying which groups and particles are within more massive structures. For each halo in turn (starting with the least massive), the remainder of the list is searched to see if the density peak is within a box containing a more massive structure; if such a box is found then the halo is associated with the more massive structure. If a structure already has associated substructures, they will also belong to the bigger structure. If the density peak is not within any other box, the search is repeated to check if there is an overlap of the minimum size boxes and, if an overlap is found, the halo is associated with the more massive structure. In this way, each structure will either belong uniquely to a combined group or be an isolated structure. We then calculate the minimal size box of each combined group or isolated structure and read in all particles inside this box, as long as they have not already been identified as belonging to another structure. Note that, in this way, each particle that has been associated with a structure by FOF belongs *uniquely* to a family. However, a small number of particles that have not been associated to a structure by FOF (either by being isolated or belonging to a group with less than 10 particles) might belong to more than one family; these particles are usually at the margin of the family and are not significant for further analysis.

In Fig. 1, we show the projection of the simulation with 256^3 particles in a box of length $L = 20\,h^{-1}$ Mpc at a redshift $z = 1$ and a mass resolution of $\approx 4 \times 10^7\,h^{-1}\,{\rm M}_\odot$. The boxes are the minimum size boxes of the five most massive structures in the simulation.

This rough analysis of structure enables one already to estimate the mass distribution in the simulation. In Fig. 2, we show the mass distribution of families for the 1024^3 simulation described in Table 1. The dominance of low-mass objects is clear. Also, we show a fit to the slope of the distribution to a generalized Schechter function (Press & Schechter 1974; Schechter 1976) $\mathrm{d}n_{\rm h}/\mathrm{d}ln(M_{\rm h}) = N_*(M_{\rm h}/M_*)^{-\alpha}\,\exp(-M_{\rm h}/M_*)$ and obtain $\alpha \approx 0.9$. The fit was performed with a non-linear least-squares Marquadt–Levenberg algorithm. Note that, at this stage, we plot the mass function of the families, which makes it harder to compare with the standard Press–Schechter prescription, which assumes virial masses and does not take into account the linking of overlapping structures; however, our findings are consistent with previous work (Ghigna et al. 2000). The dashed line in Fig. 2 shows the distribution measured by Evrard et al. (2002), which establishes that this rough catalogue agrees well with standard expectations.

After this first step, we have identified large structures in the simulation and assigned all particles that potentially belong to these structures. In the next step, this will enable us to refine the analysis within a single family.

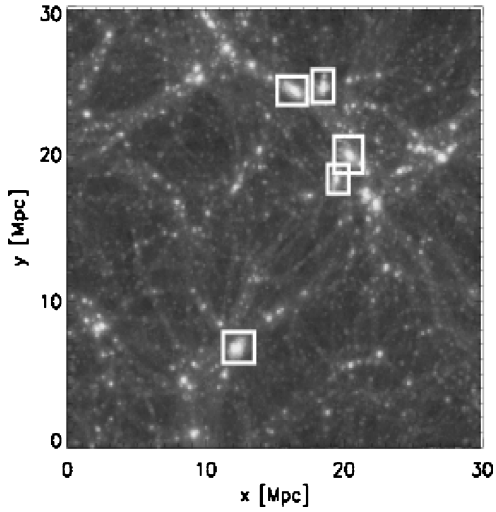

Figure 1. Projection of a simulation with 256^3 particles. The boxes are the minimum size boxes of the five most massive haloes found with FOF with a linking length of $R_{\rm link} = 1/(5\sqrt[3]{\bar{n}})$. Note that the apparent overlap of the boxes is just a projection effect.

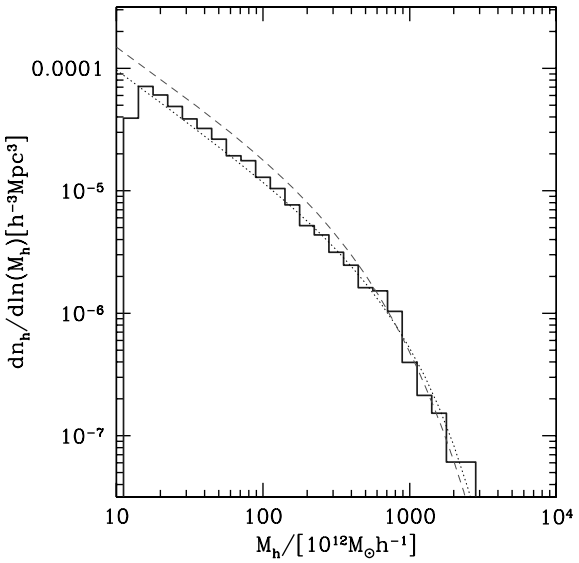

Figure 2. Mass function of clusters with more than 5000 particles for the 1024^3 simulation (right panel). The dotted line is the Schechter function with a slope of $\alpha = 0.9$ and an exponential cut-off at a scale $M_* = 8.0 \times 10^{14}\,h^{-1}\,{\rm M}_\odot$. For comparison, we also plot the mass function from Evrard et al. (2002; dashed line).

2.2 Identification of substructure and bound particles in haloes

We are now in the position to study a single family in more detail. We first perform an identification of groups within one family using the DENMAX algorithm (Bertschinger & Gelb 1991; Gelb & Bertschinger 1994). DENMAX first interpolates the density field ρ by applying a Gaussian kernel with a given smoothing length $R_{\rm smooth}$ to the particle positions. The particles are then shifted along the density gradient via the fluid equation

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\tau} = \boldsymbol{\nabla}\frac{\delta\rho}{\rho}. \tag{1}$$

Each particle moves towards a density maximum where it comes to rest, or more probably oscillates around the peak. The groups are then identified by using the FOF scheme on the shifted particles, with a linking length comparable to $R_{\rm smooth}$. We use a much smaller smoothing length $R_{\rm smooth}$ than the linking length $R_{\rm link}$ in the FOF scheme used previously for finding the rough structures. We take

$$R_{\rm smooth} = f_{\rm sub}\epsilon, \tag{2}$$

where ϵ is the softening length of the simulation and $f_{\rm sub}$ is a free parameter in our analysis, which we typically choose to be $f_{\rm sub} = 5$. This choice ensures that we identify the smallest structures that are still above the resolution threshold of the simulation (Ghigna et al. 2000). We also set the threshold for the minimum number of particles in a group to 10. In this way, we obtain a list of groups within the single family.

After the refined DENMAX step, there are still particles that are not assigned to any group with more than 10 particles. For each such particle, we locate the nearest neighbour structures and calculate the distance δr to their density peak positions. We also calculate the distance to the peak of the most massive group, which we call the *mother halo*. We then calculate $m/\delta r^2$, where m is the mass of the neighbour, and assign the particle to the group (or the mother) where this quantity is maximal. We note that any misassignments made at this stage will be rectified at a later stage in the analysis and the purpose of this simple criterion is to minimize the necessary amount of reassignment.

As an example, we show in Fig. 3 the five most massive substructures identified in the most massive mother halo of the 256^3 simulation, which has initially $\approx$321 000 particles, or a mass of $1.3 \times 10^{13}\,h^{-1}\,\mathrm{M}_\odot$. The masses of all the substructures vary between $1.2 \times 10^{9}\,h^{-1}\,\mathrm{M}_\odot$ and $2.3 \times 10^{11}\,h^{-1}\,\mathrm{M}_\odot$, where we assume we can reliably identify a substructure if it comprises at least 30 particles.

The next step is the build-up of the family tree within this family. In order to obtain the family tree, we calculate the minimum size box that contains each identified substructure. Then, as before, we apply a bottom-up scheme starting with the lowest mass halo and determine if its density peak is within the minimal box enclosing a more massive structure. The structure with the lowest mass that contains the halo is identified to be the *mother* of this halo, while the halo becomes the *daughter* and hence a substructure of the mother. If the density peak is not within any other halo, we check if the minimal box is overlapping with any other box. In this case, we take the lowest mass overlap halo as the mother. We then move to the next more massive halo and repeat the procedure. Once we have identified the mother, all the substructures of the daughter will also become daughters of the mother. In this way, we obtain a unique mother for each halo and, for each mother, a list of daughters that contains all substructures of the hierarchy. We should actually talk of daughters, granddaughters, great-granddaughters and so on, but there is no need to distinguish daughters and granddaughters from a mother's point of view, as long as each daughter knows who her mother is, which is ensured by our procedure. In other words, each mother knows about the whole younger generation, but only her mother from among her ancestors. 'Isolated' substructures will have the original mother halo as a mother. In Fig. 4, we show schematically the build-up of a family tree.

We further introduce a threshold particle number $N_{\rm t}$. Structures with fewer particles than $N_{\rm t}$ are dissolved into their associated mothers. Typically we choose $N_{\rm t} = 30$, discarding smaller groups found earlier. In Ghigna et al. (2000) a threshold of $N_{\rm t} = 16$ has been used for using haloes as tracers, but $N_{\rm t} = 32$ for the reliable analysis of

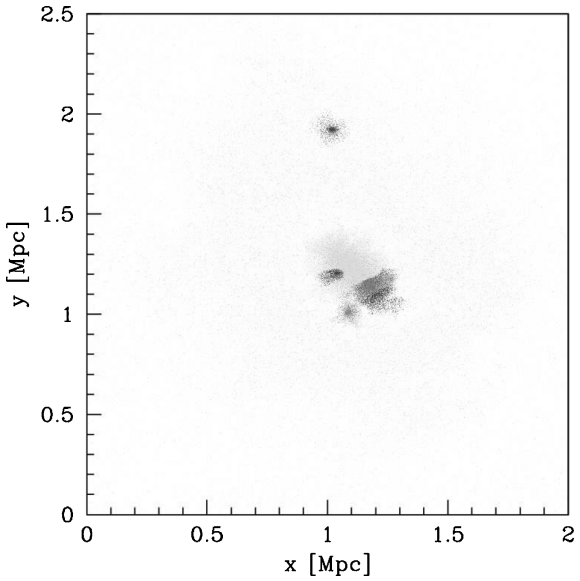

Figure 3. Projection of all particles in the mother halo (background) and the five most massive substructures found by the refined DENMAX run. The cluster has a total mass of $1.3 \times 10^{13}\,h^{-1}\,\mathrm{M}_\odot$ and the substructures range between $1.2 \times 10^{9}\,h^{-1}\,\mathrm{M}_\odot$ and $2.3 \times 10^{11}\,h^{-1}\,\mathrm{M}_\odot$ for this simulation.

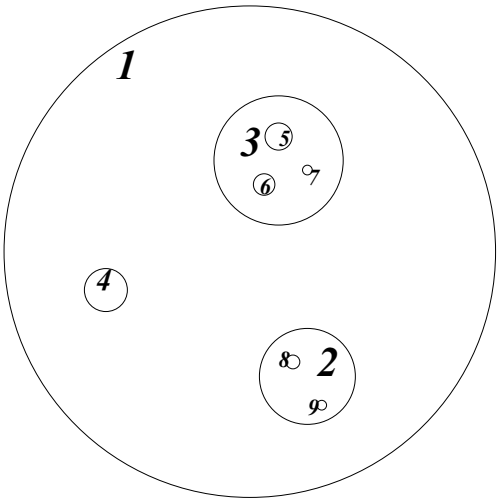

Figure 4. Schematic description of the family tree. Halo number 1 counts all subhaloes 2 to 9 as her daughters. Halo 1 is the mother of subhaloes 2, 3 and 4. Subhaloes 5, 6 and 7 are also the daughters of subhalo 3 and see 3 as their mother; 7 and 8 are in a similar relation to subhalo 2. Number 4 is an isolated subhalo.

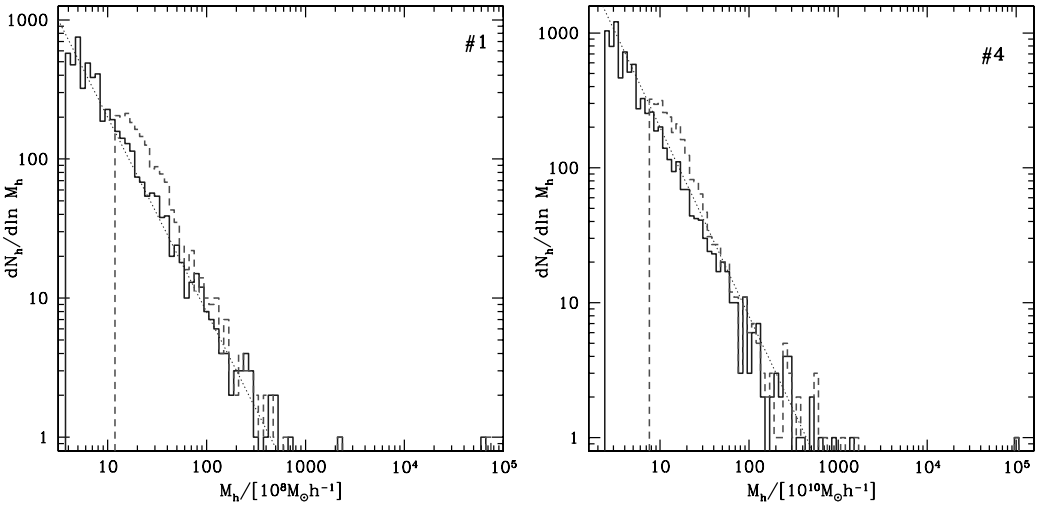

Figure 5. Left panel: mass function of the substructures of the most massive halo of the 256^3 simulation, before (solid line) and after (dashed line) structures with less than $N_t = 30$ have been dissolved into their mothers. The mother halo has ≈160 000 particles or a mass of $6.5 \times 10^{12}\, h^{-1}\, M_\odot$. The dotted line is from a power law $d\,N_h/d\ln(M_h) \propto M_h^{-1.4}$. Right panel: the same plot for the fourth most massive halo of the 1024^3 simulation with ≈392 000 particles and a mass of $1.1 \times 10^{15}\, h^{-1}\, M_\odot$ in the mother halo.

properties of haloes. However, Diemand, Moore & Stadel (2004) find that the results are only stable for $N_t = 100$. This is necessary to resolve a complete sample of subhaloes. In Fig. 5, we show the distribution of the masses of substructures in the most massive halo of the 256^3 run and in the fourth most massive halo from the 1024^3 simulation. In the following, we call these haloes cluster #1 (left panel) and cluster #4 (right panel). The entire structure, including the mother halo and all daughters, has a mass of $1.3 \times 10^{13}\, h^{-1}\, M_\odot$ for cluster #1 and $1.6 \times 10^{15}\, h^{-1}\, M_\odot$ for cluster #4. We clearly see the large number of small structures (solid lines); the original distribution follows roughly a scaling of $dN_h/dln(M_h) \propto M_h^{-1.4}$ (dotted lines). This becomes even steeper if the haloes with less than $N_t = 30$ particles are dissolved into their mothers (dashed lines). The DENMAX routine had originally recovered about 5100 substructures that have more than 10 particles within cluster #1 and 7760 haloes within cluster #4. About 20–25 per cent of the total mass of the structure is in haloes with less than 30 particles as identified by DENMAX. After haloes with less than 30 particles have been dissolved into their mothers, we are left with about 1790 substructures in cluster #1 and 2590 in cluster #4. During this procedure, the original mother halo gained $5.7 \times 10^{11}\, h^{-1}\, M_\odot$ for cluster #1 and $6.4 \times 10^{13}\, h^{-1}\, M_\odot$ for halo #4. The rest of the mass is distributed among the lower mass haloes, as seen in the dashed histograms in Fig. 5.

As mentioned in the introduction, DENMAX itself has been applied in a hierarchical way, either as part of SKID by applying three different smoothing lengths $l_{link} = 1.5, 5, 10 l_{soft}$ (Ghigna et al. 2000), or by using it on larger scales with $R_{smooth} = 0.2\bar{n}^{-1/3}$ and re-analysing each halo with $R_{smooth} = 0.1\bar{n}^{-1/3}$ (Neyrinck, Hamilton & Gnedin 2004). The reason for this is that, in general, there is no single smoothing length that is suited to find structures over a large mass range in the simulation. If the smoothing length is too large then small structures are not resolved and if it is too small then large structures are broken up. We choose a small smoothing length and recombine larger objects using the family tree hierarchy.

We have now a clearly defined, geometrically based picture of substructures, which we can proceed to analyse in a more physical fashion so that unbound particles are culled out. In some situations, the DENMAX procedure may err in assigning some particles to substructures. Imagine a particle that is dynamically a part of the mother halo: the DENMAX algorithm will move this particle towards the cluster centre, but if a significant substructure just happens to intervene, the particle will reach this local maximum and stop. Thus, there will be particles extending in a radial wedge outside of any bound structure arbitrarily attached to it, even if they are gravitationally not bound to it. To correct for such unphysical identifications, we need now a post-identification dynamical treatment of the haloes.

2.2.1 Velocity outliers

It can be shown (Binney & Tremaine 1987) that the rms escape velocity from a finite, bound self-gravitating system is related to the rms velocity by $\langle v_{esc}^2 \rangle = 4 v_{rms}^2$. Thus, particles having a velocity greater than $\sqrt{2\langle v_{esc}^2 \rangle} = \sqrt{8} v_{rms}$ are very unlikely to be bound to the structure. One way to calculate the escape velocity is by measuring the maximum value of the circular velocity $v_{circ}(r) = \sqrt{GM(r)/r}$; by assuming an NFW profile, this can be related to the escape velocity (Klypin et al. 1999). However, this method relies on the NFW profile, which we do *not* want to assume at this stage.

To remove unbound particles from a substructure, we will instead proceed with a first approximation by calculating the typical rms velocity and removing particles that are statistical outliers. However, we cannot calculate the velocity dispersion until we know the true CoM velocity; therefore, because we have not removed unbound particles from the structure, we must proceed iteratively, beginning with an approximation for the CoM. We choose the density peak of the substructure (not including its daughters) as a first approximation to the CoM. In order to obtain the CoM velocity, we calculate the median of the velocity of the $N_v = 100$ nearest neighbours to the density peak within the structure. If the number of particles is less than 30, we take half the particles of the structure. In order to obtain a valid answer, we must pay attention to binaries, which could bias the result to large velocities. Hence, we identify binaries by searching the whole simulation for bound pairs. We so far have not found bound pairs of particles in all the simulations we studied, which also provides evidence that the simulation is not over-resolved. If we did find a bound pair, the two particles would be replaced by a

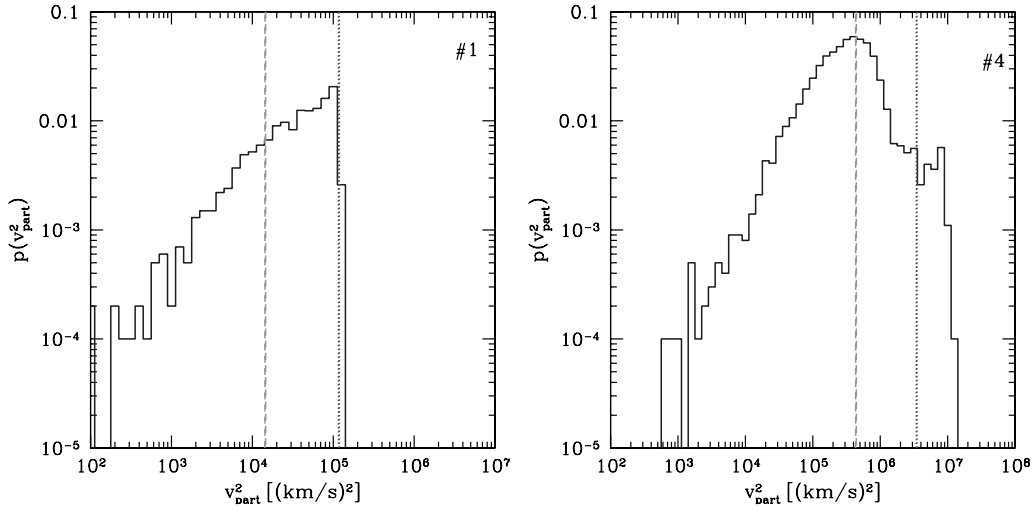

Figure 6. Distribution of the particle velocities in the most massive substructure, the velocity threshold $8v_{\rm rms}^2$ (dotted) and the rms velocity (dashed). All the particles above the threshold are moved to the associated mother. Cluster #1 is on the left and #4 is on the right.

single particle with twice the mass and the CoM position and CoM velocity of the pair. This ensures that we do not encounter velocity biases due to binaries. We then can proceed to calculate the rms velocity $v_{\rm rms}^2 = \langle(\boldsymbol{v}_{\rm part} - \boldsymbol{v}_{\rm cm})^2\rangle$ for the $N_{\rm v} = 100$ particles around the density peak. All particles in the substructure that have a velocit

$$\boldsymbol{v}_{\rm part}^2 > f_{\rm cut} v_{\rm rms}^2 \tag{3}$$

are then removed and added to their associated mother structure. We iterate this process until the mass change of the substructure is less than 5 per cent. We perform this velocity cut at two levels: first, we use $f_{\rm cut} = 8$ as noted earlier and as mentioned above $N_{\rm v} = 100$ particles for the CoM velocity and rms velocity calculation. Then, choosing a tighter limit with $f_{\rm cut} = 6$, we find the CoM mean velocity and velocity dispersion of the inner half of the particles and repeat the process.

In Fig. 6, we show the velocity distributions of the particles in the most massive substructure (solid line) in clusters #1 and #4. We also show the threshold rms velocity (dotted line) during the first step of the iteration. All particles above this threshold are moved to the associated mother. The mass of the mother halo at the end of this procedure increases by just $4.2 \times 10^{11}\,h^{-1}\,{\rm M}_\odot$ for cluster #1 and $2.0 \times 10^{13}\,h^{-1}\,{\rm M}_\odot$ for cluster #4, where most of the change occurs during the first cut-off scheme. After the removal of the velocity outliers, we again dissolve haloes with less than $N_{\rm t} = 30$ particles into their mothers.

At the end of this step, we recalculate the CoM and $v_{\rm rms}$ for the subhalo and then we move daughters that have a faster CoM velocity than $\sqrt{6}v_{\rm rms}$ to the associated mother of the substructure under consideration.

2.2.2 *Tree calculation of potentials and bound particles*

We now reach the step where we can remove particles that have a total energy larger than zero in the CoM frame of the structure to which they belong. We will check within each substructure which particles are bound to it. We calculate the CoM of a substructure including *all* its daughters and compute the potential ϕ of the particles within the substructure. The potential calculation is done using an adaption of a tree code by Hernquist (1987). Note that we switch to an exact direct summation of the potential energy if there are less than 100 particles in the system. The total energy of a particle is then

$$E_{\rm tot} = m\phi + \frac{1}{2}m(\boldsymbol{v}_{\rm part} - \boldsymbol{v}_{\rm cm})^2, \tag{4}$$

where m is the mass of a particle, ϕ the potential from all the other masses within the substructure and $\boldsymbol{v}_{\rm cm}$ the CoM velocity of the substructure. We calculate $E_{\rm tot}$ for each particle and then remove the third of the unbound particles with the highest energies, moving them to their associated mother structure. Note that we choose only a third of the particles because otherwise particles are removed too quickly without taking into account that the CoM velocity and hence the kinetic energy are changing with each removed particle. Ideally one should remove only one particle at a time, as is done in SKID (Stadel et al. 1997), but this is too time consuming for hundreds of haloes with over 10^5 particles. We tested different fractions and observed that one-third was the largest number that results in a stable result. We then recalculate the CoM and iterate this step until there is no change in the mass of the system.

In Fig. 7, we show the mass distribution before and after unbound particles have been moved to the mother structures. Note that all daughters with less than $N_{\rm t} = 30$ particles have been dissolved into their mothers. There are many unbound particles in the substructures that returned to the original mother. The mother in halo #1 now has a mass of $1.2 \times 10^{13}\,h^{-1}\,{\rm M}_\odot$ which corresponds to ≈293 000 particles; there are now only 134 daughters with a total mass of $1.1 \times 10^{12}\,h^{-1}\,{\rm M}_\odot$. The mother of halo #4 has a mass of $1.5 \times 10^{15}\,h^{-1}\,{\rm M}_\odot$ or 582 000 particles, with $1.2 \times 10^{14}\,h^{-1}\,{\rm M}_\odot$ remaining in 106 daughters.

Before we proceed with the next step, we will remove any daughter that is not bound to its mother. We approximate the potential energy for the daughters by

$$E_{\rm pot}^{\rm d} = -G\frac{m_{\rm d}M(r_{\rm d})}{r_{\rm d}} - G\int_{r_{\rm d}}^{\infty}\frac{m_{\rm d}}{r}\,{\rm d}M(r), \tag{5}$$

where $m_{\rm d}$ is the daughter mass, $r_{\rm d}$ is the distance of the CoM of the daughter to the CoM of its mother and $M(r)$ is the total mass of the mother within radius r including all other daughters. We then can calculate the kinetic energy of the daughter with respect to the CoM its mother. If the daughter is *not* bound to her mother, we move her to the mother of the mother.

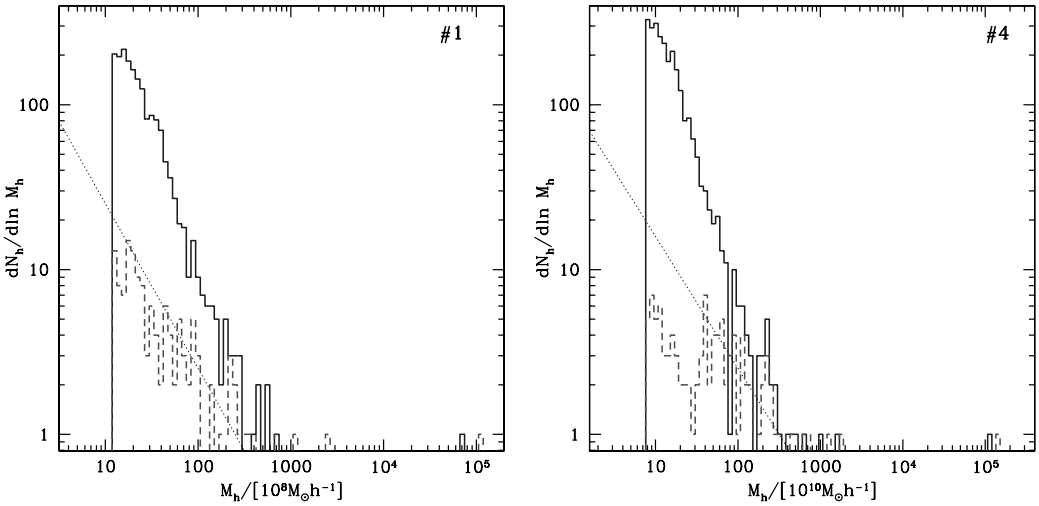

Figure 7. Mass function of substructure in the test clusters #1 (left panel) and #4 (right panel) after particles with high velocities have been moved to the associated mothers (solid line) and after unbound particles have been moved to the mother structure (dashed line). Note that the power laws are now $\mathrm{d}N_\mathrm{h}/\mathrm{d}ln(M_\mathrm{h}) \propto M_\mathrm{h}^{-1.0}$ for #1 and $\mathrm{d}\,N_\mathrm{h}/\mathrm{d}\,ln(M_\mathrm{h}) \propto M_\mathrm{h}^{-0.8}$ for #4.

2.2.3 Search for hyperstructures

In order to obtain a stable algorithm with respect to the smoothing length for the refined DENMAX procedure, we need, as noted above, to look for 'hyperstructures', groups of substructures that are gravitationally collectively bound to one another.

This problem has been addressed previously by combining the SKID algorithm with an adaptive FOF analysis (Diemand et al. 2004); we will take a different approach here. In order to do this, we investigate primary substructures, ie. structures that are *direct* daughters of the largest structure, which is the mother structure. For each such primary substructure, we calculate the distance δr_i to each other primary substructure with mass m_i and examine the one with the maximal $m_i/\delta r_i^3$ as follows; note that the masses include all daughters of the primary substructures. If these two structures are bound with respect to their common CoM, they form a hyperstructure, and the less massive of the two becomes a daughter of the more massive structure. We than recalculate the CoM and the maximum extension box of the new hyperstructure and check each particle of the mother within this box. If it is bound to the hyperstructure, we then move it from the mother to this hyperstructure. We will iterate this step three times. Note that, for both haloes, the mass in the mother structure does not change significantly during this step.

In this fashion, bound objects, whose identity is independent of the geometrical tool used to analyse substructure, are assembled.

2.2.4 Final steps: daughters and particles unbound to the entire family

The next step we perform is to remove daughters that are not bound to the biggest structure, the mother halo. Then, finally, we remove particles that are not bound to the family tree at all. For both haloes, none of the daughters is unbound and the number of unbound particles is negligible.

In Fig. 8, we show the mass distribution from the refined DENMAX run (solid line) and after the unbinding steps (dashed line). There

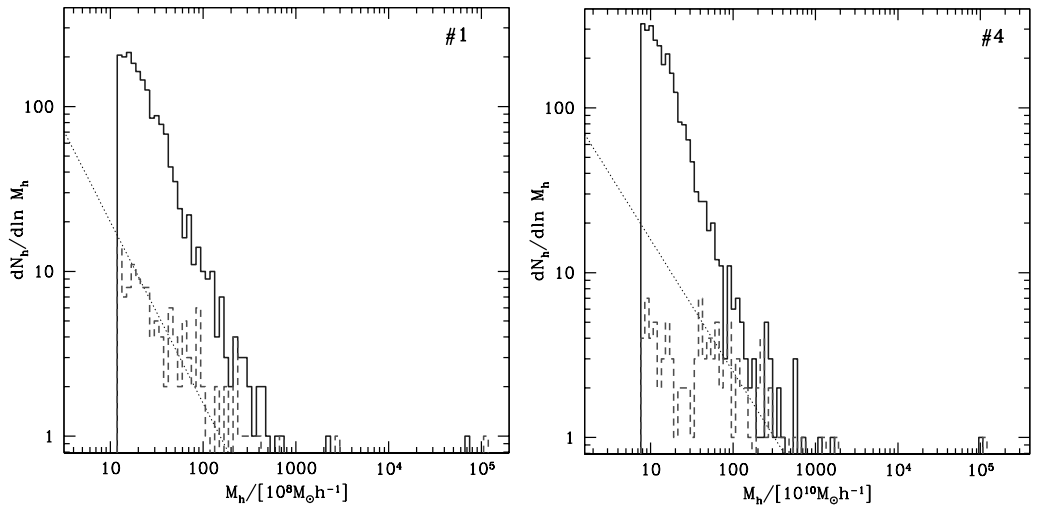

Figure 8. Mass function of substructures in the halo after the refined DENMAX run and the mass cut (solid) and at the end of the the iterative scheme (dashed). The dotted lines show the power laws $\mathrm{d}N_\mathrm{h}/\mathrm{d}ln(M_\mathrm{h}) \propto M_\mathrm{h}^{-1.1}$ (left panel) and $M_\mathrm{h}^{-0.8}$ (right panel).

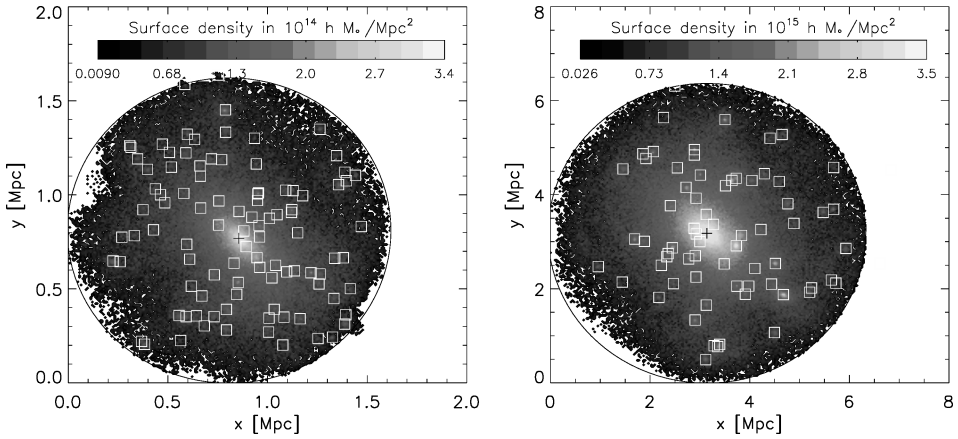

Figure 9. Particles and daughters in halo # 1 (left panel) and halo # 4 (right panel) after the clean-up procedure and removal of particles and subhaloes outside the virial radius. The dark cross marks the centre of the halo and the boxes the identified and bound daughters. The colour corresponds to the surface density as indicated by the colour bar. Cluster #1 has 96 subhaloes and #4 has 65.

remains 8.8 per cent of the mass in substructures for #1 (left panel) and 7.4 per cent for halo #4 (right panel). For cluster #1, $\mathrm{d}N_\mathrm{h}/\mathrm{d}ln(M_\mathrm{h}) \propto M_\mathrm{h}^{-1.1}$ for small mass haloes; the distribution in cluster #4 is roughly approximated by $\mathrm{d}N_\mathrm{h}/\mathrm{d}ln(M_\mathrm{h}) \propto M_\mathrm{h}^{-0.8}$. Note these are only rough power laws.

2.2.5 Truncation of halo at the 'virial radius' and identification of companions

Now we check if we have artificially linked together separate structures that are only weakly coupled together gravitationally, and if we have artificially included distant infalling matter. For a ΛCDM cosmology, it is conventional to define the virial mass M_vir and radius R_vir with

$$M_\mathrm{vir} = \frac{4}{3}\pi R_\mathrm{vir}^3 \Delta_\mathrm{c}(z)\rho_\mathrm{c}(z)\,, \tag{6}$$

where ρ_c is the critical density of the universe and the mean overdensity $\Delta_\mathrm{c} = 178\,\Omega_\mathrm{m}(z)^{0.45}$. Thus, we make a rank ordered list of our mother haloes and, in each one, we start at the density maximum and proceed outwards until we reach the virial radius, within which the mean overdensity is Δ_c. We truncate the halo at this point, removing all particles from outside the virial radius of the halo and identifying daughters with centres outside this radius as separate companion structures. In this step, the mass of the mother halo in cluster #1 stays almost constant at $1.1 \times 10^{13}\,h^{-1}\,\mathrm{M}_\odot$ while 96 subhaloes with a total mass of $9.4 \times 10^{11}\,h^{-1}\,\mathrm{M}_\odot$ remain. In cluster #4, the mother mass is reduced to $1.2 \times 10^{15}\,h^{-1}\,\mathrm{M}_\odot$ with 65 remaining subhaloes of a mass of $8.2 \times 10^{13}\,h^{-1}\,\mathrm{M}_\odot$.

In Fig. 9, we show the projected density inside the virial radius of cluster #1 (left panel) and cluster #4 (right panel), with the daughters marked. Note that cluster #1 is at a redshift of $z = 1$ while cluster #4 is at redshift $z = 0.05$. Halo #1 has a virial radius of $r_\mathrm{vir} \approx 0.81$ Mpc and halo #4 a radius of $r_\mathrm{vir} \approx 3.0$ Mpc. Note that for halo #4 the substructure is much more centred around the core than for halo #1.

3 DEPENDENCE ON PARAMETERS AND TEST OF STABILITY

In this section, we discuss the stability of our method, as different possible choices of the parameters and procedures may affect the results.

3.1 Group finding

We will first vary the linking length in our rough initial FOF analysis in order to establish how sensitive we are to this parameter choice. We perform an analysis with a linking length of $R_\mathrm{link} = 0.167\bar{n}^{-1/3}$: this could potentially lead to a larger fragmentation of initial haloes and families and hence potentially change our results. We find that the results of this run are almost identical to results obtained with the original linking length. For halo #4, we have after the initial fine DENMAX run 35 per cent of the mass in substructures (compared to 34 per cent in the original run), which is lowered to 9 per cent (8 per cent) after we test for bound particles; after the final virial cut, there was 8 per cent of the halo mass in substructures, while the original run resulted in slightly lower than 8 per cent. This is due to the fact that our family tree procedure followed by a refined DENMAX run produces almost the same large structures.

We did a further consistency check where instead of FOF we used a rough DENMAX run with a smoothing length of $R_\mathrm{smooth} = \frac{1}{5}\bar{n}^{-1/3}$ to identify the initial halo list. The results were essentially the same as in the original runs. Hence, we conclude that our method is stable with respect to sensible changes in the initial halo finding algorithm to within ±1 per cent, which is well below the statistical fluctuation of the sample.

The next halo finding step we perform is the refined DENMAX run. We crucially chose in this step the smallest sensible smoothing length and then built up the halo hierarchy by our family tree algorithm. Making the smoothing length smaller than $R_\mathrm{smooth} = 5\epsilon$ would enter the regime dominated by uncertainties in the force softening, so we do not extend a stability test in this direction.

Instead, we repeated the analysis with a smoothing length of $R_\mathrm{smooth} = 10\epsilon$. Due to the larger smoothing length, we find 56 per cent of the mass in substructures after the initial DENMAX step for halo #4; however, when we test for particles that are actually bound to these structures, we obtain already 9 per cent of the mass in substructures. After the inclusion of hyperstructures and the density cut, this drops to 7 per cent, which is in excellent agreement compared to the run with a smoothing length $R_\mathrm{smooth} = 5\epsilon$ (8 per cent).

We hence conclude that we have a reasonably stable criterion if the smoothing length is chosen within a reasonable range. Of course, as the smoothing length is made larger, we will miss more and more structures.

3.2 Removal of unbound particles

The first step of removing unbound particles is performed by removing velocity outliers in a gentle way. Because we do this already in two steps with first a gentler and then a harder cut-off at $8v_{\rm rms}^2$ and $6v_{\rm rms}^2$ we established that most of the cut is happening during the first iteration step. However, the velocity cut does not change the mass fraction significantly. Final results do not depend on the specific numbers $(8, 6) \times v_{\rm rms}^2$, as long as we approach the final cut gradually. Furthermore, we note that this cut was mainly done to avoid an unphysical bias towards large CoM velocities, which is important for the calculation of the kinetic energies with respect to the CoM.

3.3 Virial cut

Because the definition of a mass or size of a halo is to some extent arbitrary (see for example: Jenkins et al. 2001; Evrard et al. 2002; White 2002), we will investigate how this definition influences our results. We chose initially the virial mass and radius corresponding in a ΛCDM cosmology to the overdensity $\Delta_{\rm c}(z) = 178\,\Omega_{\rm m}(z)^{0.45}$. We saw already in the comparison of the 256^3 and 1024^3 simulations that, after this density cut, the fraction of mass in substructures can be quite different. However, the simulation with 256^3 was also at a redshift of $z = 1$, compared with $z = 0$ for the 1024^3 simulation. Hence, we performed an analysis of the 1024^3 run where we chose the cut-off overdensity to be $\Delta_{\rm c} = 200$ in agreement with another commonly used definition. With this cut-off, the final mass fraction in subhaloes only decreases from 8 to 7 per cent for halo #4.

4 APPLICATION TO A DIFFERENT SIMULATION

In order to test our algorithm, we applied it to a simulation provided by Diemand et al. (2004). We chose their cluster *D*6, where the simulation was done with a smoothing length of $\epsilon = 3.6$ kpc, which is comparable to our runs. Diemand et al. (2004) state that their halo finding is complete for haloes with more than 100 particles and they find about 5 per cent of the mass in substructures. With our scheme, we also obtain 5 per cent of the mass in substructures, which is excellent agreement given the difference of the analysis methods. Diemand et al. (2004) use a hierarchical version of the SKID algorithm, which is based on DENMAX. They perform the unbinding iteratively and exactly with no approximation like the one discussed in Section 2.2.

In order to get a further insight into the statistics of substructures, we compare the cumulative mass functions of halo *D*6 from our analysis and the analysis by Diemand et al. (2004). In Fig. 10, we show the cumulative mass function for substructures in the test halo *D*6. The solid line is from our analysis and the dashed line from (Diemand et al. 2004). They look both very similar and scale very closely to M^{-1} until a cut-off at less than 1000th of the total cluster mass. Our analysis results in slightly more substructures than the one of (Diemand et al. 2004). We find 272 substructures, while they find 241. This is actually strikingly similar given that the difference of the presented algorithms and the overall mass fraction in substructures for both analyses is at the 5 per cent level within 1 per cent uncertainty.

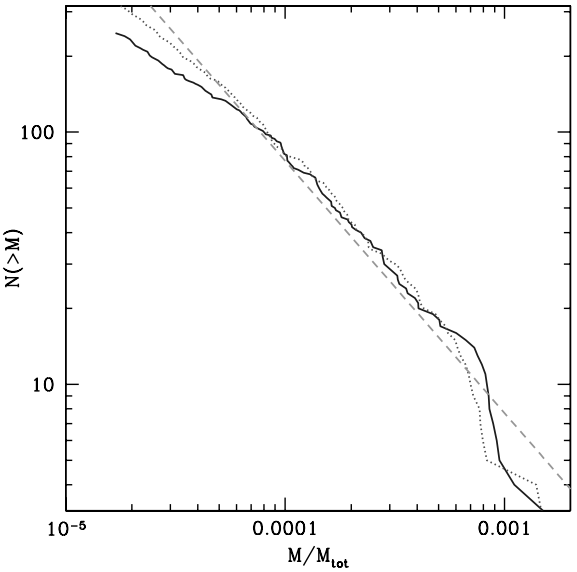

Figure 10. Cumulative mass function for substructures with more than 30 particles in halo *D*6 as presented by Diemand et al. (2004). The solid line is from the algorithm presented here, the dotted line from (Diemand et al. 2004) and the dashed line has a slope of M^{-1}.

5 CONCLUSION

In this paper, we have established a fast and stable algorithm to identify vast numbers of substructures in large *N*-body simulations in a speedy fashion. For example, to analyse the most massive halo of about 1.5 million particles takes about 8 h on a Sun Blade 2000 (Sun Microsystems, Inc., Santa Clara, CA, USA) on a single 900-MHz processor with 3 Gbyte of RAM. We established an approximate method to identify and remove unbound particles from subhaloes, which allows for the efficient calculation of bound structures. We analysed three simulations, two done by the TPM code developed by Bode et al. (2000) and one by Diemand et al. (2004). For all three, we find similar mass fractions of about 5–8 per cent. (Diemand et al. 2004) find about 5 per cent of the mass in substructures, which is identical to our findings.

The fraction of substructure in a CDM cluster is to some extent a question of definition. For instance, if one is interested in strong lensing, which tests the distribution of matter, the question of bound or unbound structures is irrelevant. However, if substructures are the places where galaxies form, a full dynamical treatment is relevant and requires the inclusion of *all* forces, including the ones from internal and external potential.

To conclude, we emphasize that the presented algorithm is stable and fast and ready to be employed for large cosmological data sets as well as detailed simulations of clusters of galaxies.

ACKNOWLEDGMENTS

We thank A. Amara, J. Diemand, G. Efstathiou, S. Kazantzidis, A. Kravtsov, B. Moore and T. Naab for useful discussions, B. Moore and J. Diemand for the provision of the external simulations and the referee for valuable suggestions. The parallel computations were done in part at the UK National Cosmology Supercomputer Centre funded by PPARC, HEFCE and Silicon Graphics/Cray Research. This research was supported by the National Computational Science Alliance under NSF Cooperative Agreement ASC97-40300, PACI Subaward 766; also by NASA/GSFC (NAG5-9284). Computer time was also provided by NCSA and the Pittsburgh Supercomputing Center.

REFERENCES

Bertschinger E., Gelb J., 1991, Comp. Phys., 5, 164
Binney J., Tremaine S., 1987, Galactic Dynamics. Princeton Univ. Press Princeton
Bode P., Ostriker J. P., 2003, ApJS, 145, 1B
Bode P., Ostriker J. P., Xu G., 2000, ApJS, 128, 561
Bode P., Ostriker J., Turok N., 2001, ApJ, 556, 93
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
De Lucia G., Kauffmann G., Springel V., White S. D. M., Lanzoni B., Stoehr F., Tormen G., Yoshida N., 2004, MNRAS, 348, 333
Diemand J., Moore B., Stadel J., 2004, MNRAS, 352, 535
Eisenstein D., Hut P., 1998, ApJ, 498, 137
Evrard A. E. et al., 2002, ApJ, 573, 7
Gelb J., Bertschinger E., 1994, ApJ, 436, 467
Ghigna S., Moore B., Governato F., Lake G., Quinn T., Stadel J., 1998, MNRAS, 300, 146
Ghigna S., Moore B., Governato F., Lake G., Quinn T., Stadel J., 2000, ApJ, 544, 616
Gill S. P. D., Knebe A., Gibson B. K., 2004, MNRAS, 351, 399
Gnedin O. Y., Hernquist L., Ostriker J. P., 1999, ApJ, 514, 109
Hernquist L., 1987, ApJS, 64, 715
Huchra J., Geller M., 1982, ApJ, 257, 423
Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, MNRAS, 321, 372
van Kampen E., 1995, MNRAS, 273, 295
Kim J., Park C., 2004, preprint (astro-ph/0401386)
Klypin A., Gottlöber S., Kravtsov A. V., Khokhlov A. M., 1999, ApJ, 516, 530
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Lacey C., Cole S., 1994, MNRAS, 271, 676
Moore B., Katz N., Lake G., 1996, ApJ, 457, 455
Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, ApJ, 524, L19
Navarro J., Frenk C., White S. D. M., 1995, MNRAS, 275, 720
Neyrinck M. C., Hamilton A. J. S., Gnedin N. Y., 2004, MNRAS, 348, 1
Neyrinck M. C., Gnedin N. Y., Hamilton A. J. S., 2005, MNRAS, 356, 1222
Okamoto T., Habe A., 1999, ApJ, 516, 591
Press W., Schechter P., 1974, ApJ, 187, 452
Schechter P., 1976, ApJ, 203, 297
Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
Stadel J., Katz N., Weinberg D. H., Hernquist L., 1997, online reference only (www-hpcc.astro.washington.edu/tools/skid.html)
Summers F. J., Davis M., Evrard A., 1995, ApJ, 454, 1
Warren M., Quinn P., Salmon J., Zurek W., 1992, ApJ, 399, 405
Weinberg D., Hernquist L., Katz N., 1997, ApJ, 477, 8
White S. D. M., 1976, MNRAS, 177, 717
White M., 2002, ApJS, 143, 241

This paper has been typeset from a TEX/LATEX file prepared by the author.