RAA 2020 Vol. **20** No. **4**, 46(12pp) doi: 10.1088/1674–4527/20/4/46
© 2020 National Astronomical Observatories, CAS and IOP Publishing Ltd.
http://www.raa-journal.org *http://iopscience.iop.org/raa*

Research in Astronomy and Astrophysics

HIKER: a halo-finding method based on kernel-shift algorithm

Shuang-Peng Sun[1,2], Shi-Hong Liao[1], Qi Guo[1,2], Qiao Wang[1,2] and Liang Gao[1,2,3]

[1] Key Laboratory of Computational Astrophysics, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; *sunshp@nao.cas.cn*
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Institute for Computational Cosmology, Department of Physics, Durham University, Science Laboratories, South Road, Durham DH1 3LE, England

Received 2019 September 29; accepted 2019 October 31

Abstract We introduce a new halo/subhalo finder, HIKER (a Halo fInder based on KERnel-shift algorithm), which takes advantage of a machine learning method – the mean-shift algorithm combined with the Plummer kernel function, to effectively locate density peaks corresponding to halos/subhalos in density field. Based on these density peaks, dark matter halos are identified as spherical overdensity structures, and subhalos are bound substructures with boundaries at their tidal radius. By testing HIKER code with mock halos, we show that HIKER performs excellently in recovering input halo properties. In particular, HIKER has higher accuracy in locating halo/subhalo centres than most halo finders. With cosmological simulations, we further show that HIKER reproduces the abundance of dark matter halos and subhalos quite accurately, and the HIKER halo/subhalo mass functions and $V_{\rm max}$ functions are in good agreement with two widely used halo finders, SUBFIND and AHF.

Key words: methods: *N*-body simulations — galaxies: halos — galaxies: evolution — cosmology: theory — dark matter

1 INTRODUCTION

Cosmological *N*-body simulations are one of the most crucial methods to study the structure formation and evolution of the universe (see e.g., Frenk & White 2012; Kuhlen et al. 2012, for reviews). To compare simulations with observations so that cosmological models can be constrained,it is essential to identify gravitationally bound structures (e.g., dark matter halos and subhalos) from simulation data. In particular, modern cosmological simulations with large simulated volume and high numerical resolution resolve plenty of structures and substructures spanning a wide range of masses/sizes. How to identify dark matter halos/subhalos efficiently and robustly from these large simulations is of great importance. In the past few decades, many codes/methods have been developed to identify halos/subhalos from simulation snapshots (i.e., halo finders); see Knebe et al. (2011) and Onions et al. (2012) for reviews.

Usually, the algorithms to identify halos can be classified into two categories, spherical overdensity (SO, Press & Schechter 1974; Lacey & Cole 1994) and friends-of-friends (FOF, Davis et al. 1985) method. The SO halo finder first locates peaks in the density field, then a sphere centred on each peak is grown out until the mean matter density within this sphere reaches a given threshold. Usually this threshold is expressed as an overdensity parameter, Δ, measured with respect to the mean matter density or the cosmic critical density. The final sphere defines the boundary of a halo. The FOF halo finder groups simulation particles whose pairwise distances are smaller than a given threshold, which is usually expressed as the linking length parameter, b, measured with respect to the mean interparticle distance (More et al. 2011). The centre of an FOF halo is usually defined as the position of the most bound particle. In contrast, the identification of subhalos is more complicated than that of halos. In recent years, many subhalo finders have been developed (see Onions et al. 2012, for examples), and there is less agreement on the definition of a subhalo among different methods. For instance, based on the FOF group finding results, SUBFIND (Springel et al. 2001) defines locally over-dense and self-bound substructures as subhalos, and uses saddle point to determine the boundary of a subhalo; AHF (Amiga Halo Finder, Knollmann & Knebe 2009) constructs adaptive mesh refinement on the density field in the simulated box size and finds subhalos based on the hierarchy of the refined grids; HBT (Hierachical Bound-Tracing, Han et al. 2012, 2018) identifies subhalos by tracking the merging hierarchy of halos starting from the earliest given snapshot.

Recently, machine learning algorithms have been widely adopted in the astronomical community, and they are becoming more and more useful as the astronomical data size grows more rapidly. For example, Hui et al. (2018) use support vector machine (SVM) to classify galaxies to study the relation with large-scale structures; Aragon-Calvo (2019) uses deep convolutional neural network (CNN) to perform large-scale structures classification; He et al. (2019) uses deep neural networks to predict the formation of non-linear large-scale structures. We refer the reader to Baron (2019) for a recent review of various applications of machine learning methods in astronomy. Unsupervised machine learning algorithms are widely used in clustering analysis, outlier detection, dimensionality reduction, etc. In particular, the concept of clustering analysis for data points, which tries to group objects so that the objects in the same group (or cluster) are more similar to each other than objects in other groups (or clusters), is fairly similar to halo finding in cosmological simulations. This motivates us to develop a new halo finder code based on clustering analysis algorithms.

The mean-shift algorithm (Fukunaga & Hostetler 1975; Cheng 1995; Comaniciu & Meer 2002) is a non-parametric cluster finding procedure, and it is popular in computer vision and image processing applications. It starts from an initial position, and then iteratively shifts to the local maximum (i.e., mode) of a density field with a shift vector computed from the average property of data points in the neighborhood. The mean-shift algorithm does not assume the shape of the distribution nor the number of clusters, and it is quite efficient to locate the local maxima of a density distribution. These advantages make it an attractive candidate method to identify halos as well as subhalos from cosmological simulations. In particular, as we show in this paper, the mean-shift algorithm equipping with a physically motivated kernel function, the Plummer kernel, has another advantage in accurately identifying substructures from simulations. To stress the feature of our algorithm (i.e., the mean-shift algorithm with the Plummer kernel function), we use the name of "kernel-shift" in this paper.

In this work, we introduce a new halo finder, HIKER (a Halo-fInding method based on KERnel-shift algorithm) to simultaneously identify dark matter halos and subhalos in cosmological N-body simulations. The organization of our paper is as follows. Section 2 introduces the kernel-shift algorithm and other details of our halo finding procedures. We then apply HIKER to mock halo data, as well as cosmological N-body simulations. We present our test results on field halos and subhalos in Section 3. Section 4 presents our conclusions and discussions.

2 HIKER ALGORITHM

The general procedures of the HIKER algorithm can be described as follows: (i) Identifying density peaks with the kernel-shift algorithm. (ii) For each density peak, a sphere is grown around its centre until the enclosed mean density is 200 times the critical density. Halos and subhalos are determined according to their geometrical relations. (iii) For each subhalo, we determine its tidal radius and apply for an unbinding procedure to exclude unbound particles.

2.1 Mean-shift Algorithm

The mean-shift algorithm is one of the clustering algorithms in the unsupervised machine learning category, which is very efficient and robust to locate density peaks of a density field. The algorithm was first presented in Fukunaga & Hostetler (1975) and later generalized by introducing kernel functions in Cheng (1995).

To locate the local density maxima of a set of particles, the mean-shift algorithm starts from an initial guessing position, moves iteratively according to mean-shift vectors computed from the neighboring particles, until the convergence is reached. The key concept of the algorithm is the mean-shift vector. In the original mean-shift algorithm proposed by Fukunaga & Hostetler (1975), for a position $\boldsymbol{x}$, the mean-shift vector is computed as

$$\boldsymbol{M}_b(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{x})\,, \tag{1}$$

where $\boldsymbol{x}_i$ is the coordinate of the i-th particle whose distance to position $\boldsymbol{x}$ is less than the bandwidth radius, b, and N is the total number of such particles. Note that the bandwidth radius here is the single input parameter in the mean-shift algorithm. Physically, if a given set of particles have equal masses, the mean-shift vector is defined as the difference between position $\boldsymbol{x}$ and the centre-of-mass position of the particles within the bandwidth radius.

Cheng (1995) later generalized the mean-shift vector by introducing a Kernel function, $K(y;b)$,

$$\boldsymbol{M}_b^K(\boldsymbol{x}) = \frac{\sum_{i=1}^{N} K(|\boldsymbol{x}_i - \boldsymbol{x}|\,;b)\,(\boldsymbol{x}_i - \boldsymbol{x})}{\sum_{i=1}^{N} K(|\boldsymbol{x}_i - \boldsymbol{x}|\,;b)}, \tag{2}$$

where $|\boldsymbol{x}_i - \boldsymbol{x}|$ is the distance from $\boldsymbol{x}$ to $\boldsymbol{x}_i$, and N here denotes the total number of particles which have their $|\boldsymbol{x}_i - \boldsymbol{x}|$ smaller than an input radius r_K. Under such generalizations, the original mean-shift vector, $\boldsymbol{M}_b(\boldsymbol{x})$, can be regarded as $\boldsymbol{M}_b^K(\boldsymbol{x})$ computed with a flat kernel, i.e., $K(y;b) = 1$ if $y \leq b$, and 0 otherwise. For each position, it can be moved to a new position with higher density according to the mean-shift vector; i.e., $\boldsymbol{x}_{\rm new} = \boldsymbol{x} + \boldsymbol{M}_b^K(\boldsymbol{x})$. Iteratively repeating the procedure leads to convergence once it reaches a local density maximum. For the detailed mathematical proof of the convergence for the mean-shift algorithm, see Cheng (1995).

In HIKER, we intend to adopt a non-flat kernel function, which gives more weight in the inner region rather than using equal weight everywhere. Such a non-flat kernel function assists in locating halo centres more robustly, and this is particularly important in identifying subhalos. Comparing to field halos in cosmological simulations, subhalos usually reside in more complicated density environments. It is more robust to locate subhalo centre when giving more weight to the central region. In contrast, using a flat kernel fails to identify some subhalos; see Appendix A for a quantitative comparison between using flat and non-flat kernel functions to identify subhalos.

To construct a non-flat kernel function, we start from the spherically symmetric Plummer density profile, which was first introduced by Plummer (1911) to describe the stellar distribution of globular clusters,

$$\rho(r) = \frac{3M}{4\pi r_s^3} \frac{1}{(1 + r^2/r_s^2)^{5/2}} \,, \tag{3}$$

where M is the total mass and r_s is the scale radius. The corresponding Plummer potential is

$$\phi(r) = -\frac{GM}{(r^2 + r_s^2)^{1/2}} \,, \tag{4}$$

where G is the gravitation constant, and the corresponding force is

$$\boldsymbol{F}(\boldsymbol{r}) = -\frac{GM\boldsymbol{r}}{(r^2 + r_s^2)^{3/2}} \,. \tag{5}$$

Our Plummer kernel function is constructed as the following form,

$$K_p(r; r_s) = \frac{C}{(r^2 + r_s^2)^{3/2}} \,, \tag{6}$$

where C is a normalization constant, while r and r_s are respectively corresponding to $|\boldsymbol{x}_i - \boldsymbol{x}|$ and b in Equation (2). In this work, we set $r_s = 3\epsilon$ with ϵ being the gravitational softening length from simulations. If a particle is infinitely close to a Plummer potential minimum, then the Plummer style mean-shift vector

$$\boldsymbol{M}_b^{K_p} \propto \sum_{i=1}^{N} \frac{\boldsymbol{r}_i}{(r_i^2 + r_s^2)^{3/2}} \tag{7}$$

is proportional to the sum of Plummer force. With this kernel function, shifting from an initial seed in iterations approximately follows the force exerted by the corresponding Plummer potential. In other words, the iterative procedure mimics the process of a particle falling into a Plummer potential well. Note, BDM halo finder (Klypin & Holtzman 1997; Riebe et al. 2013) adopts a scheme essentially identical to mean-shift algorithm with a flat kernel to locate density peaks, while it was called “sphere jittering” mechanism in the paper.

2.2 Seeds Function

A seeds function is necessary to generate the initial positions (i.e., seeds) to start the kernel-shift algorithm. The simplest scheme for seeds generation is to randomly choose positions of a certain fraction of simulation particles. However, some small targets in low-density regions may be missed accidentally in such a scheme, and thus additional seeds should be placed in these low-density regions (Klypin & Holtzman 1997). Also note that the number of seeds has direct impact on the amount of computational expense; i.e. more seeds causes more subsequent calculations (or computation time), whilst they do not affect the final results on the number of peaks. For an optimal seeds function, we expect that every candidate halo/subhalo should contain at least one seed in order to avoid being missed, but at the same time, the total number of seeds should be as few as possible in order to reduce computational cost.

In HIKER, we develop a novel approach to generate seeds as detailed below. First, for each particle in the simulation, we search its 20 nearest neighbor particles using k-d trees (Bentley 1975), and compute its distance to the 20th nearest particle, d_{20}, which is an approximate proxy of its local density. Second, we use a predefined distance threshold, r_c, which is similar to the linking length parameter in an FOF algorithm but is slightly larger to be more conservative, to exclude those particles whose local densities are too low to be associated with any halo. We set r_c as 0.24 times the mean inter-particle separation in our code. Third, we loop over the remaining particles and adopt the particles whose d_{20} are the smallest among the d_{20} of their 20 nearest neighbor particles (i.e., those particles which correspond to the local density maxima) as candidate seeds. In the final step, note that the candidate peaks which are close to each other are still possible to be moved to the same halo centre during the kernel-shift iteration. Therefore, to further reduce and computational cost, we loop over the candidate seeds, and for each candidate seed, we search its neighbouring candidate seeds within a spherical region with a radius defined as 3 times of the scale radius of the Plummer kernel function (see Sect. 2.1 for details). A candidate seed peak will be determined as a true seed if its local density is still the maximum among its neighbors. Note that in this procedure, the seeds we identified fairly represent the realistic density peaks. Starting from each seed, we select all particles inside a radius of $r_K = 10 \times \epsilon$ and apply for mean-shift procedure to derive all density peaks. Note that more than one seeds may converge to a single peak with mean shift algorithm.

2.3 Halo/Subhalo Identification

Based on density peaks, we follow an approach of SO halo finder to identify field halo (Lacey & Cole 1994). The density peaks are sorted in descending order according to their local densities which are estimated from the distance to the 20th nearest neighbour. We first consider each peak as the centre of a field halo, while many of them are indeed centres of subhalos. Starting from the density peak with the highest local density, then for each peak, a sphere is grown around centre until the enclosed mean density is equal to 200 times the critical density. Then we further distinguish halo and subhalo using simple geometrical relation, namely if a halo is contained in the other larger halo, the smaller halo is labeled as a subhalo. Once halos/subhalos are distinguished, we need to define subhalo boundary because only subhalo halo centres are known at the stage. There are various definitions of subhalo boundary in the literature. For example, SUBFIND uses saddle points, HBT (Han et al. 2012, 2018) groups the bound particles of its progenitor as a subhalo. In HIKER we use a physically motivated value – tidal radius to define the boundary of a subhalo. In practise, starting from each subhalo centre, a sphere is grown until its tidal radius reaches. We follow the definition of tidal radius in Tormen et al. (1998); i.e.,

$$r_t = R\left[\frac{m}{(2 - \partial \ln M/\partial \ln R)M(R)}\right]^{1/3}, \qquad (8)$$

where m, R and M are subhalo mass, the distance from subhalo centre to host halo centre, and the enclosed mass within R for the host halo respectively. This definition is under the assumption that the subhalo mass and the tidal radius are far less than the values of its host halo.

A halo/subhalo is usually regarded as a gravitationally bound structure, and thus it is often for some halo finders to remove unbound particles from the candidate halos/subhalos (i.e., unbinding procedure). In HIKER, we carry out the unbinding procedure following that of the AHF halo finder except of determining the bulk velocity for a potential halo/subhalo. For that we use our Plummer kernel to weight particle velocities according to their distances to the centre, instead of using a small fraction of particles within the central region to get the mean velocity as the bulk velocity. For other details of the unbinding procedure, we refer the reader to Knollmann & Knebe (2009).

3 TEST RESULTS

In this section, we test our halo finder in three aspects using three different types of simulation data, i.e., we use (i) a set of mock halos to test the accuracy of halo properties given by HIKER (Sect. 3.1), (ii) a suite of large-scale full-box cosmological simulations to test the ability in identification of field halos (Sect. 3.2), (iii) the data from the Aquarius project (Springel et al. 2008) to test the ability to identify subhalos (Sect. 3.3). Note that these data have been used in the halo/subhalo finder comparison Knebe et al. (2011) and Onions et al. (2012), and thus almost all our results can be directly compared to those of two studies. For the tests on mock halos, we will compare the HIKER results with those from all halo finders discussed in Knebe et al. (2011), while for the other two tests, to be concise, we mainly compare the HIKER results with two widely used halo finders, SUBFIND and AHF.

3.1 Mock Halos

Mock halos, whose properties are known by construction, are introduced in Knebe et al. (2011) to examine the accuracy of different halo finders in recovering halo/subhalo properties. A set of mock halo data has been prepared with a given NFW density profile. With such halo model, three setups have been generated: (i) an isolated host halo only; (ii) an isolated host halo $+$ a subhalo at $0.5R_{100}^{\rm host}$; (iii) an isolated host halo $+$ a subhalo at $0.5R_{100}^{\rm host}$ $+$ a subsubhalo at $(0.5R_{100}^{\rm host} + 0.5R_{100}^{\rm subhalo})$. Here, $R_{100}^{\rm host}$ ($R_{100}^{\rm subhalo}$) is the radius of the host halo (subhalo) within which the mean density is 100 times the critical density. The properties of mock halos are summarized in Table 1; see Knebe et al. (2011) for further details of these mock data.

In the following, we will use these three setups of mock halos to test the accuracy of HIKER in recovering halo/subhalo properties.

Following Knebe et al. (2011), we first quantify the accuracy of HIKER in recovering halo/subhalo centres by computing the distance offset between the centre returned by HIKER and the actual one. Results are shown with red symbols in Figure 1, in which the results of other 17 halo finders[1] from Knebe et al. (2011) have also been shown with black symbols for easy comparison. We use squares, diamonds, and triangles to plot the centre offsets of host halos, subhalos, and subsubhalos respectively. The symbol

[1] These halo finders are AHF (Knollmann & Knebe 2009), ASOHF (Planelles & Quilis 2010), BDM (Klypin & Holtzman 1997), pSO (Sutter & Ricker 2010), LANL (Habib et al. 2009), SUBFIND (Springel et al. 2001), FOF, pFOF (Courtin et al. 2011; Rasera et al. 2010), NTROPYFOF (Gardner et al. 2007a,b), VOBOZ (Neyrinck et al. 2005), SKID (Stadel 2001), ADAPTAHOP (Aubert et al. 2004; Tweed et al. 2009), HOT_3D, HOT_6D (Ascasibar & Binney 2005; Ascasibar 2010), HSF (Maciejewski et al. 2009), 6DFOF (Diemand et al. 2006) and ROCKSTAR (Behroozi et al. 2013).

Table 1 The (Sub)Halo Properties of the NFW Mock Data

Profile	Type	N_{200}	$M_{200}\,[h^{-1}M_\odot]$	$R_{200}\,[h^{-1}\mathrm{kpc}]$	$R_\mathrm{s}\,[h^{-1}\mathrm{kpc}]$	$V_\mathrm{max}\,[\mathrm{km\,s^{-1}}]$
NFW	host	760 892	7.61×10^{13}	689.1	189.5	715
	sub	8066	8.07×10^{11}	151.4	17.0	182
	subsub	84	8.42×10^{9}	33.1	2.6	43

N_{200} and M_{200} are the particle number and virial mass inside the virial radius, R_{200}. R_s is the scale radius of the corresponding profile, and V_max is the maximum of the rotation curve.

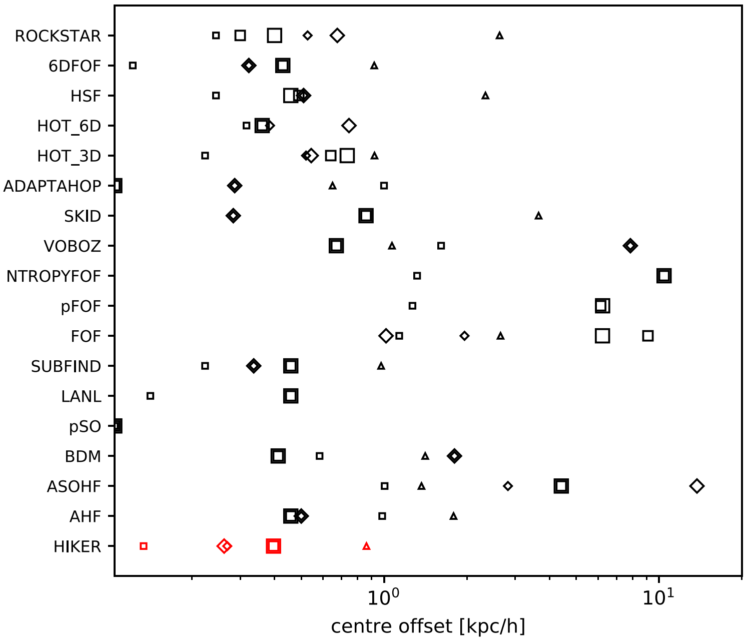

Fig. 1 Centre offset between the actual centre and the value recovered by HIKER. The results of different halo finders from Knebe et al. (2011) are plotted with *black color* at different y-coordinates, and the corresponding names are labeled on the y-axis. The HIKER results are highlighted with *red colors*. Host halos, subhalos, and subsubhalos are marked with *squares*, *diamonds*, and *triangles*, respectively. The symbol size distinguishes the results of different setups, with larger symbols representing setups containing more substructures. Specifically, taking the ROCKSTAR results as an example, the three *squares* from left to right show the results of host halos from setup (i), (ii), and (iii) respectively, the two *diamonds* from left to right show the subhalos from setup (ii) and (iii) respectively, and the *triangle* shows the subsubhalo from setup (iii).

size distinguishes different mocks, with larger symbol corresponding to the mocks with more subhalos.

For NFW mocks, HIKER recovers the input halo properties fairly well. Especially for the isolated host halo of setup (i), the halo centre recovered by HIKER only deviates $0.13\,h^{-1}\mathrm{kpc}$ (i.e., $\sim 2\times10^{-4}\,R_{200}$) from the actual centre. For host halos containing substructures, due to the asymmetry in the density field caused by nesting substructure, the deviations become slightly larger ($\sim 0.4\,h^{-1}\mathrm{kpc}$, or $\sim 6\times10^{-4}\,R_{200}$), but they are still smaller than the centre offsets of most halo finders. For subhalos and subsubhalos, HIKER also recovers their centres more accurately than many other halo finders.

We also investigate how accurately HIKER can recover some other halo properties, including bulk velocities, virial masses, and the maximum circular velocities for both halos and subhalos (i.e., $x = V_\mathrm{bulk}, M_{200}$, and V_max), and the results are presented in Figures 2, 3 and 4 respectively. Note that the accuracy is defined as the fractional difference,

$$\frac{\Delta x}{x_\mathrm{model}} \equiv \frac{x_\mathrm{code}-x_\mathrm{model}}{x_\mathrm{model}}, \tag{9}$$

where x_code and x_model stand for the halo properties computed by a halo finder and the input mock properties respectively. The layout and symbols in these figures are similar as those in Figure 1.

From Figures 2, 3 and 4, it is easy to find that HIKER recovers the aforementioned properties quite successfully both for field halos and subhalos (i.e., the fractional differences are less than 1% for V_bulk and V_max, and less than 6% for M_{200}), and the HIKER results are usually better than (or comparable to) those of the other halo finders. For subsubhalos, the HIKER recoveries of M_{200} (i.e., $\sim$15%) and V_max (i.e., $\sim$3%) are not so good as the cases for field halos and subhalos, but it is still better than many other halo finders.

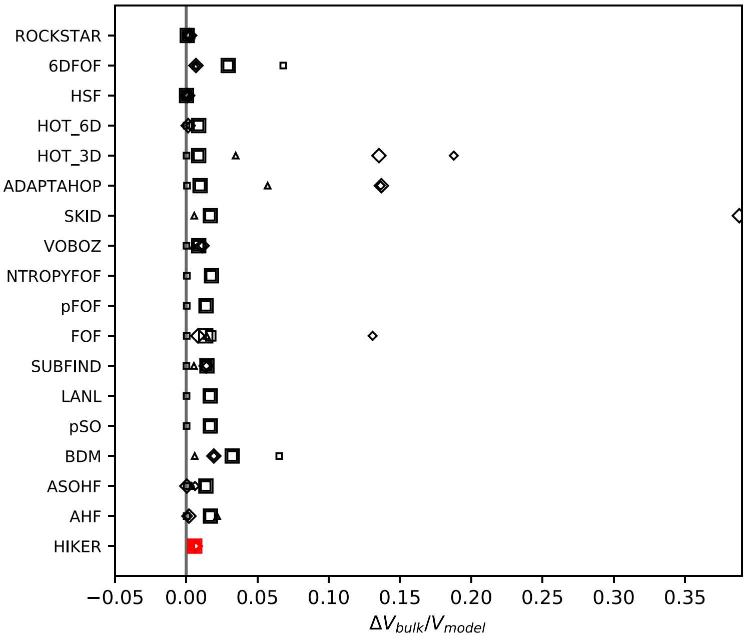

Fig. 2 Fractional differences between the input halo bulk velocity in mock data and the recovered bulk velocities from different halo finders. The *grey vertical lines* in both panels indicate no difference from the model analytical value. The layout and the symbols are in accordance with Fig. 1.

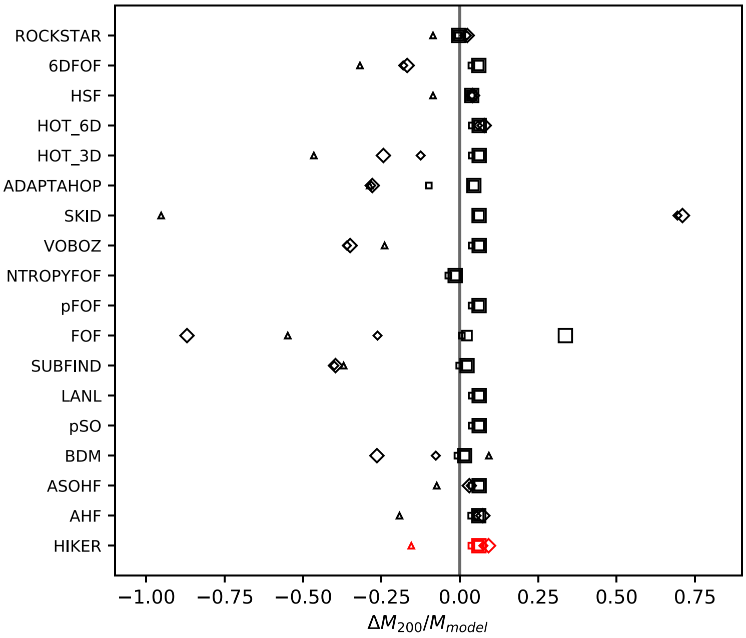

Fig. 3 Similar to Fig. 2, but for the virial mass.

Further more, as mentioned earlier, both of BDM and HIKER are based on the mean-shift algorithm but with different kernel functions. Comparing the results between BDM and HIKER in Figures 1-4, HIKER recovers the halo properties better than BDM which uses a flat kernel function. This is a consequence of a non-flat kernel adopted in HIKER.

Besides the tests of recovering the properties of mock halos, we also reproduce the dynamic evolution of a subhalo falling into a host halo with the same data used in Knebe et al. (2011). The infall process is designed in the following way. An NFW model subhalo (see the second type of Table 1 for detailed properties) is set up at a distance of $D = 3 \times R_{100}^{\rm host}$ with an initial velocity toward

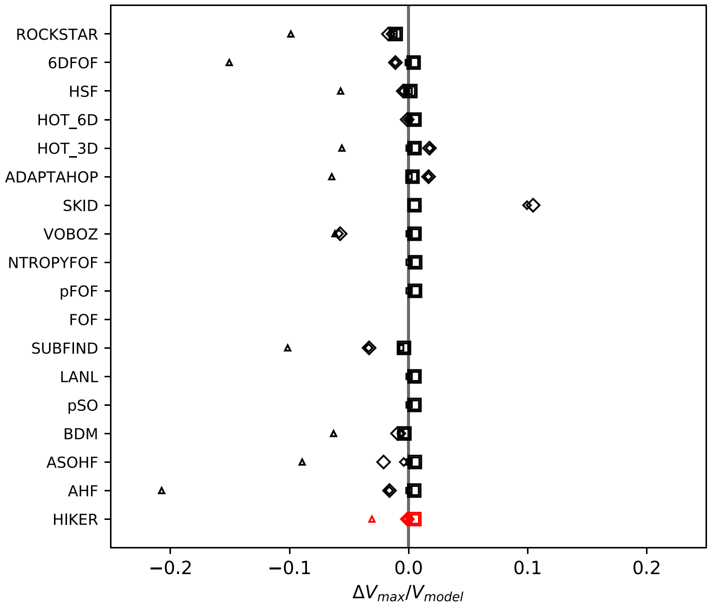

Fig. 4 Similar to Fig. 2, but for the maximum circular velocity.

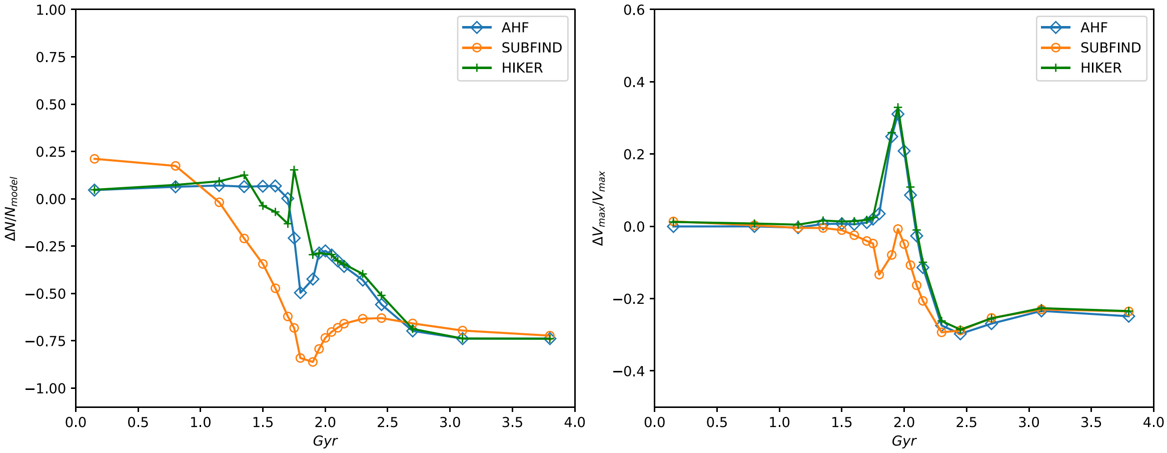

Fig. 5 Evolution of the number of particles (*left*) and the maximum circular velocity (*right*) of a subhalo falling into its host. The *green curve* shows results reproduced by HIKER. As comparison, the *blue* and *orange curves* show results of AHF and SUBFIND Knebe et al. (2011), respectively.

the host halo centre. The host halo is more massive than this subhalo by two orders of magnitude. After approximately 1.8 Gyr it will reach the host halo centre and pass through. In Figure 5 we present the evolution of the number of particles (left panel) and maximum circular velocity (right panel) recovered by HIKER as well as the original AHF and SUBFIND results from Knebe et al. (2011). Both for particle number and $V_{\rm max}$, the trends of HIKER are quite similar with those of AHF. The total number of particles decline generally when passing through the central region of host halo due to strong stripping. At the snapshots when the subhalo is very close to the its host centre, there are a little rise in the particle number curve and a sharp rise in the $V_{\rm max}$ curve. These sudden changes may be due to our unbinding procedure, which could be largely affected since the assumption which requires the (sub)halo to be spherical break down in the situation. We refer the reader to Knebe et al. (2011) for results from more halo finders.

From the discussions in this subsection, we conclude that HIKER is quite successful in recovering halo properties.

3.2 Field Halos

In this subsection, we use a suite of large-volume cosmological simulations to test the accuracy of HIKER in identifying field halos. The simulation data comes from

Table 2 Details of the Large-volume Cosmological Simulations

Names	$N_{p,\rm dm}$	$m_{\rm dm}[h^{-1}M_{\odot}]$	$\epsilon[h^{-1}\rm kpc]$
MAD-Halo-256	256^3	5.40 $\times 10^{11}$	50
MAD-Halo-512	512^3	6.59 $\times 10^{10}$	25
MAD-Halo-1024	1024^3	8.24 $\times 10^{9}$	15

$N_{p,\rm dm}$ gives the total number of dark matter particles in each simulation, $m_{\rm dm}$ is the original uncorrected mass for dark matter particles, and ϵ is the comoving softening length.

the 'Haloes gone MAD' halo-finder comparison project (Knebe et al. 2011), and it consists of three simulations with different mass resolutions; i.e., containing 256^3, 512^3, and 1024^3 dark matter particles respectively. These simulations have simulated the formation and evolution of the large-scale structures with the GADGET2 code (Springel 2005) in a comoving periodic box with a size $500\,h^{-1}\rm Mpc$ on a side. The adopted cosmological parameters are $\Omega_m = 0.3, \Omega_b = 0.045, \Omega_\Lambda = 0.7$, and $h = 0.7$, respectively. The simulation names, dark matter particle masses, and softening lengths are summarized in Table 2. For each simulation, we only use the snapshot at $z = 0$ to perform our tests.

Note that the original simulations contain both dark matter and gas particles. However, following Knebe et al. (2011), we only use dark matter particles to perform identify halos, and the mass of dark matter particles in each simulation (i.e., $m_{\rm dm}$ shown in Table 2) has been scaled by multiplying a factor of $\Omega_m/(\Omega_m - \Omega_b)$ accordingly.

We identify field halos containing at least 20 particles from all three simulations. Their cumulative mass functions and $V_{\rm max}$ functions are presented in Figure 6. Because these three simulations have the same phases in the initial conditions and they only differ in resolutions, we expect that the halo mass functions as well as $V_{\rm max}$ functions should converge in the reliable mass range among different resolution simulations. This is indeed true in Figure 6, indicating that HIKER works successfully in simulations with different mass resolutions.

We then compare the halo mass function and $V_{\rm max}$ function obtained from HIKER in the MAD-Halo-1024 simulation with those from SUBFIND and AHF in Figure 7, here the SUBFIND and AHF results come from Knebe et al. (2011), and we refer the reader to figures 17 and 18 of the paper for more results of other halo finders. We also over-plot the analytical halo mass functions as given by Warren et al. (2006) and Tinker et al. (2008) in the figure for comparison.

We can see that the HIKER mass function agrees with the SUBFIND and AHF very well in all mass range. In the reliable mass range, all three mass functions from halo finders lie between the parameterized mass functions of Warren et al. (2006) and Tinker et al. (2008). For the $V_{\rm max}$ function, HIKER tends to be slightly higher than the other two, and it is more evident in the lower $V_{\rm max}$ end. As there is no such difference in the halo mass functions among these halo finders, this implies that some HIKER halos (especially some with lower masses) tend to have higher $V_{\rm max}$, or equivalently deeper inner potentials, comparing to AHF or SUBFIND results. This possibly comes from the fact that HIKER locate halo centres more accurately, resulting in larger $V_{\rm max}$ in low mass halos.

3.3 Subhalos

We use the Aq-A halo from the Aquarius project (see Springel et al. 2008, for details) to test HIKER in identifying subhalos. The Aq-A halo has been re-simulated with five different resolutions, here we use three of them, i.e., Aq-A-4, Aq-A-3, and Aq-A-2, to perform our tests. Among these three simulations whose details are summarized in Table 3, the Aq-A-2 has the highest mass resolution while the Aq-A-4 has the lowest mass resolution. For all three simulations, we select a cubic region with edge length of $1\,h^{-1}\rm Mpc$ centring the the position of $\boldsymbol{r}_{\rm fiducial} = (57060.4, 52618.6, 48704.8)\,h^{-1}\rm kpc$ which is the fiducial centre defined in Onions et al. (2012) to run our halo finder. Note that within this selected region, the number of low-resolution particles is extremely few (i.e. less than 10), and thus we simply leave out these low-resolution particles when running HIKER. In the following, we mainly compare the HIKER results with the AHF and SUBFIND ones, and the reader can refer to Onions et al. (2012) for results of other halo finders.

In Figure 8, We first compare the subhalos identified with HIKER from the Aq-A-4 data to those identified with AHF and SUBFIND by visualisation. To be in accordance with Onions et al. (2012), we have show the same quadrant region around the fiducial position here. Each identified subhalo is represented with a green circle whose radius scales with its $V_{\rm max}/3$, and the subhalo centre is marked with a red dot. Note that only subhalos with $V_{\rm max} > 10~\rm km~s^{-1}$ are plotted in this figure. Comparing to AHF and SUBFIND, HIKER misses one subhalo in the upper left corner, and it identifies a few more low-mass subhalos at the lower left corner (i.e., the region near the Aq-A halo centre). But in general, the HIKER subhalos agree very well with the AHF and SUBFIND ones in positions and $V_{\rm max}$.

As a quantitative comparison, in Figure 9 we plot the cumulative mass functions and $V_{\rm max}$ functions for subhalos in the Aq-A-4 simulation identified by HIKER, AHF, and SUBFIND. The subhalos used to plot this figure are within a sphere of $250\,h^{-1}\rm kpc$ from the fiducial position and contain at least 20 particles. Overall the HIKER results are in line with those of AHF and SUBFIND.

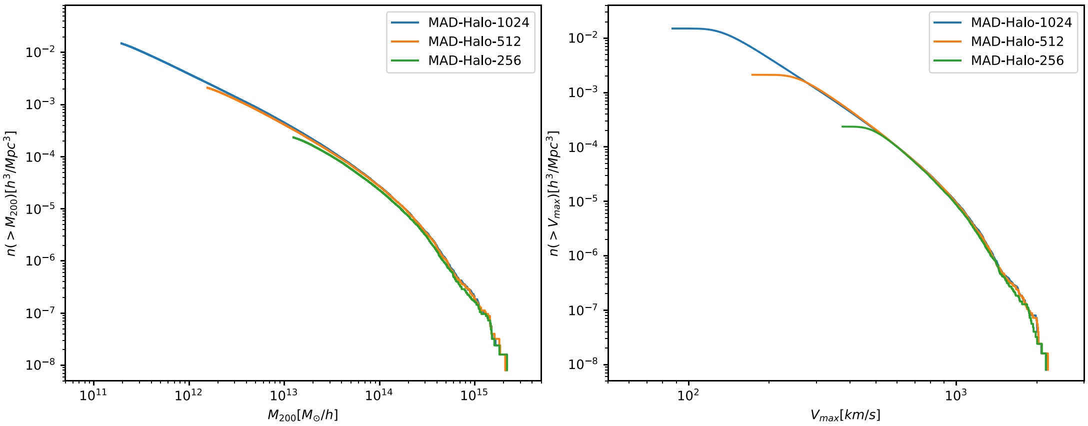

Fig. 6 Cumulative mass functions (*left*) and $V_{\rm max}$ functions (*right*) computed from HIKER field halo catalogues for three simulations with different mass resolutions.

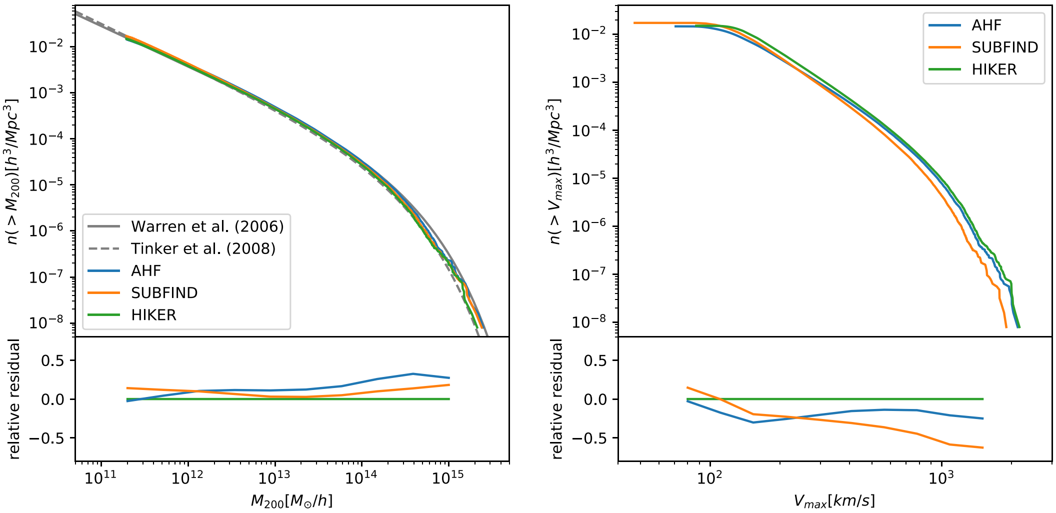

Fig. 7 Cumulative mass functions (*left*) and $V_{\rm max}$ functions (*right*) functions from the MAD-Halo-1024 simulation. The results from AHF, SUBFIND, and HIKER are plotted with *blue*, *orange*, and *green lines*, respectively. In the left panel, for comparisons, we overplot the mass functions from Warren et al. (2006) and Tinker et al. (2008) with *grey solid* and *grey dashed lines* respectively. The lower panels show residuals of mass and $V_{\rm max}$ function relative to HIKER results, respectively.

Table 3 Some Details of the Aq-A Halos Used in This Study

Name	$N_{\rm hres}$	$N_{\rm lres}$	$N_{\rm select}$	$m_{p,{\rm hres}}[h^{-1}M_\odot]$	ϵ[pc]
Aq-A-4	18 535 972	634 793	7 434 975	2.868×10^5	342.5
Aq-A-3	148 285 000	20 035 279	59 347 132	3.585×10^4	120.5
Aq-A-2	531 570 000	75 296 170	212 792 272	1.000×10^4	65.8

$N_{\rm hres}$ ($N_{\rm lres}$) is the number of high-resolution (low-resolution) particles in the simulation, $N_{\rm select}$ is the number of high-resolution particles within our selected region (i.e., a cubic region with edge length of $1\,h^{-1}$Mpc centring $\boldsymbol{r}_{\rm fiducial} = (57060.4, 52618.6, 48704.8)\ h^{-1}$kpc), $m_{p,{\rm hres}}$ is the mass of high-resolution particles, and ϵ is the comoving softening length.

We have also used HIKER to identify subhalo on the level 2 and level 3 Aq-A simulations, and the results are presented in Figure 10. As expected, the HIKER subhalo mass functions and $V_{\rm max}$ functions converge very well in different resolution simulations. These results are consistent with the resolution convergence tests shown in Springel et al. (2008) with SUBFIND. From the discussions above, we conclude that HIKER identifies subhalos with an accuracy comparable to that of the widely used AHF and SUBFIND.

The HIKER code is parallelized with OpenMP. Its performance is quit efficient, for instance, it only takes ~ 1 minute to process MAD-Halo-256 data and ~ 3 minutes to the Aq-A-4 data with 10 Xeon CPU cores (2.4 GHz).

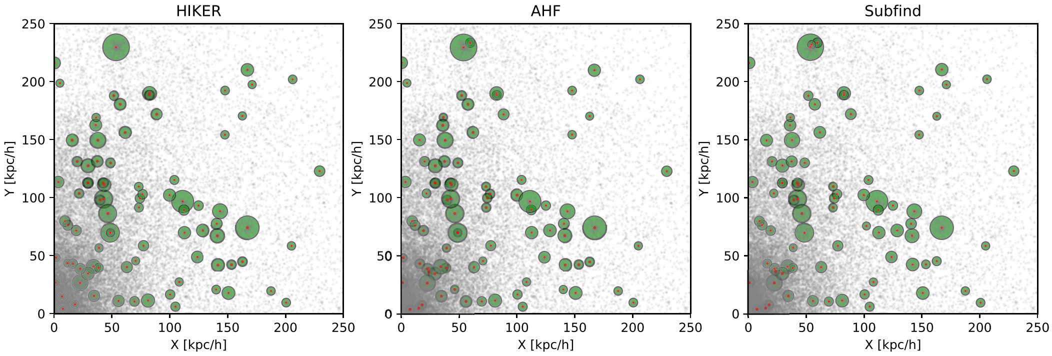

Fig. 8 Visualization of subhalo-finding results from HIKER (*left*), AHF (*middle*) and SUBFIND (*right*) on the Aq-A-4 data. The region shown in each panel is the same quadrant as presented in Onions et al. (2012). The identified subhalos are indicated by *red dots* and *green circles* whose radii are scale with $V_{\rm max}/3$. Only subhalos with $V_{\rm max} > 10\ {\rm km\ s^{-1}}$ are shown here. The grey background shows the dark matter density computed from simulation particles.

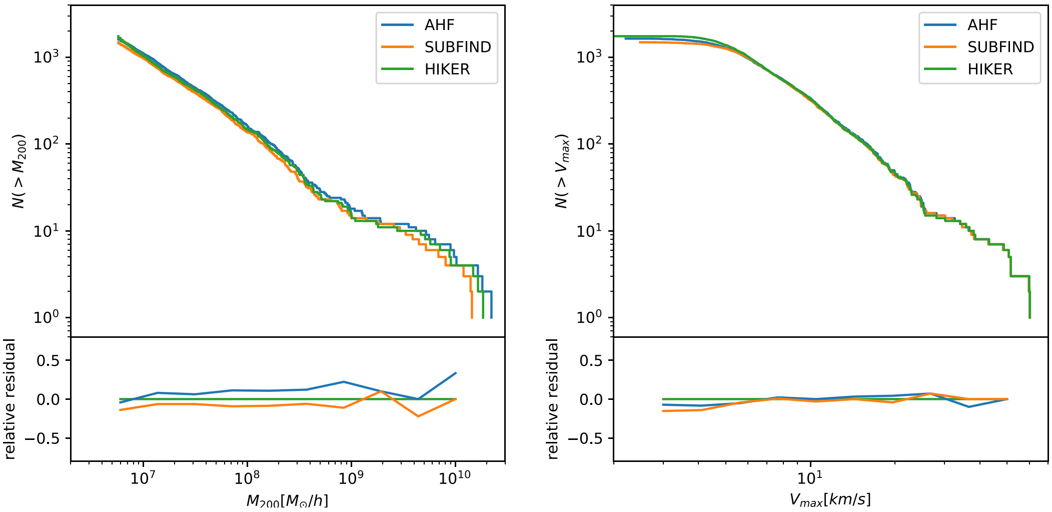

Fig. 9 Cumulative mass functions (*left*) and $V_{\rm max}$ functions (*right*) for subhalos identified from the spherical region with a radius of $250\,h^{-1}{\rm kpc}$ around the fiducial position in the Aq-A-4 data. The results from AHF, SUBFIND, and HIKER are plotted with *blue*, *orange*, and *green lines*, respectively. In the lower panels, both for mass function and $V_{\rm max}$ function, we plot the relative residual results using HIKER result as basis.

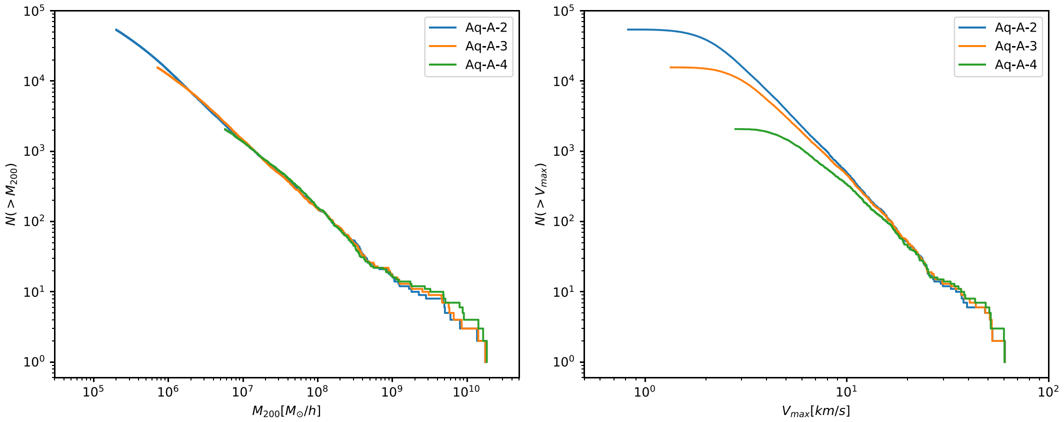

Fig. 10 Cumulative mass functions (*left*) and $V_{\rm max}$ functions (*right*) for subhalos identified from different Aq-A simulations by HIKER. Similar to Fig. 9, these subhalos are from the spherical region with a radius of $250\,h^{-1}{\rm kpc}$ centring the fiducial position. We use the *blue*, *orange*, and *green lines* to plot the results from Aq-A-2, Aq-A-3, and Aq-A-4 simulations, respectively.

For identifying both the field halo and subhalo, HIKER exhibits a time complexity close to (slightly steeper than) a $O(N \log N)$ relation.

4 CONCLUSIONS AND DISCUSSION

In this work, we develop a new spherical overdensity halo/subhalo code–HIKER for cosmological simulations. HIKER employs the mean-shift algorithm combining with a Plummer kernel to efficiently and robustly locate density peaks. Based on density peaks, dark matter halos are further identified as spherical overdensity structures, while subhalos are substructures with boundaries equal to their tidal radius. We use mock halos to test our halo-finding code, and show that HIKER performs excellently in locating halo/subhalo centres and recovering halo properties. In particular, the accuracy of HIKER in recovering halo/subhalo centres is higher than most halo finders. With large-volume and zoom-in cosmological simulations, we further showed that HIKER reproduces the abundance of field halos and subhalos quite accurately, and the HIKER results are in agreement with those of two widely used halo finders, SUBFIND and AHF.

Although we only use HIKER to identify halos/subhalos from dark matter-only simulations in this study, it can be quite straightforward to extend the HIKER algorithm to include particles with different masses (e.g., gas, stars, etc.) by further multiplying the kernel function with different weights for different particle types in Equation (2).

Acknowledgements We thank the referee for a very helpful report. We acknowledge support from the National Key Program for Science and Technology Research and Development (2017YFB0203300). SS is particularly grateful to Prof. Alexander Knebe for providing us data to do our tests as well as a lot of helpful discussions. QG and LG acknowledge support from the National Natural Science Foundation of China (NSFC) (No. 11425312), and two Royal Society Newton Advanced Fellowships, as well as the hospitality of the Institute for Computational Cosmology at Durham University. QG is also supported by NSFC (Grant Nos. 11573033 and 11622325), the "Recruitment Program of Global Youth Experts" of China, and the NAOC grant (Y434011V01).

Appendix A: KERNEL EFFECTS

Kernel functions are a key concept in the HIKER algorithm. To study quantitatively the effects of kernel functions on the identification of field halos, we run HIKER twice, first equipped with a flat kernel and then with a Plummer kernel (with $b = 3\epsilon$), on the MAD-Halo-512 simulation. Similar runs are performed on the Aq-A-4 simulation data to study the effects on subhalo finding. Note that in the unbinding procedures, to estimate halo bulk velocities more reliably, in the runs with flat kernels we only use a certain fraction of particles in the central region as most halo finders do, while in the runs with Plummer kernels we utilize the Plummer kernel to give more weight on the central velocity.

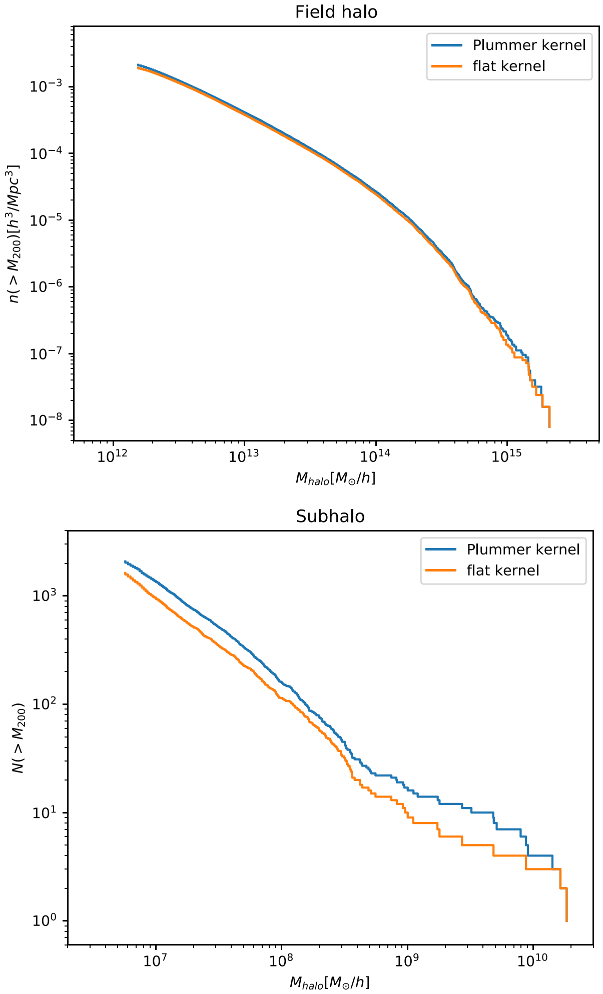

Fig. A.1 Cumulative mass functions for field halos from MAD-Halo-512 simulation (*top*) and for subhalos from Aq-A-4 simulation (*bottom*). In both panels, the blue and orange lines plot the results from HIKER equipped with a Plummer kernel and a flat kernel, respectively.

The cumulative mass functions from these runs are summarized in Fig. A.1. The field halo mass functions are barely affected by kernel functions. However, the subhalo mass functions are very sensitive to kernel functions, i.e. the number of subhalos recovered in the run with a flat kernel is much lower than that in the run with a Plummer kernel. As we have shown in Section 3.3, the HIKER subhalo results agree fairly well with those of AHF and SUBFIND.

The results here point out that introducing a non-flat kernel function can help locate the halo centres in a more robust way, especially in finding subhalos. Usually, a potential subhalo is surrounded by more complex density fields, and this makes it easy for the candidate centre to shift away if the central core is not emphasized. In contrast, field halos are usually isolated, and the density environment around them is much simpler, and a flat kernel will be good enough to capture that.

Because the centre locating method in BDM is equivalent to the mean-shift algorithm with a flat kernel, the results in this appendix also suggest that with a Plummer kernel function, HIKER can significantly improve BDM in identifying subhalos.

References

Aragon-Calvo, M. A. 2019, MNRAS, 484, 5771

Ascasibar, Y. 2010, Computer Physics Communications, 181, 1438

Ascasibar, Y., & Binney, J. 2005, MNRAS, 356, 872

Aubert, D., Pichon, C., & Colombi, S. 2004, MNRAS, 352, 376

Baron, D. 2019, arXiv:1904.07248

Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013, ApJ, 762, 109

Bentley, J. L. 1975, Commun. ACM, 18, 509

Cheng, Y. 1995, IEEE Transactions on Pattern Analysis and Machine Intelligence, 17, 790

Comaniciu, D., & Meer, P. 2002, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 603

Courtin, J., Rasera, Y., Alimi, J. M., et al. 2011, MNRAS, 410, 1911

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

Diemand, J., Kuhlen, M., & Madau, P. 2006, ApJ, 649, 1

Frenk, C. S., & White, S. D. M. 2012, Annalen der Physik, 524, 507

Fukunaga, K., & Hostetler, L. 1975, IEEE Transactions on Information Theory, 21, 32

Gardner, J. P., Connolly, A., & McBride, C. 2007a, Astronomical Society of the Pacific Conference Series, 376, A Framework for Analyzing Massive Astrophysical Datasets on a Distributed Grid, eds. R. A. Shaw, F. Hill, & D. J. Bell, 69

Gardner, J. P., Connolly, A., & McBride, C. 2007b, arXiv:0709.1967

Habib, S., Pope, A., Lukić, Z., et al. 2009, in Journal of Physics Conference Series, 180, 012019

Han, J., Cole, S., Frenk, C. S., et al. 2018, MNRAS, 474, 604

Han, J., Jing, Y. P., Wang, H., & Wang, W. 2012, MNRAS, 427, 2437

He, S., Li, Y., Feng, Y., et al. 2019, Proceedings of the National Academy of Science, 116, 13825

Hui, J., Aragon, M., Cui, X., & Flegal, J. M. 2018, MNRAS, 475, 4494

Klypin, A., & Holtzman, J. 1997, astro-ph/9712217

Knebe, A., Knollmann, S. R., Muldrew, S. I., et al. 2011, MNRAS, 415, 2293

Knollmann, S. R., & Knebe, A. 2009, ApJS, 182, 608

Kuhlen, M., Vogelsberger, M., & Angulo, R. 2012, Physics of the Dark Universe, 1, 50

Lacey, C., & Cole, S. 1994, MNRAS, 271, 676

Maciejewski, M., Colombi, S., Springel, V., Alard, C., & Bouchet, F. R. 2009, MNRAS, 396, 1329

More, S., Kravtsov, A. V., Dalal, N., & Gottlöber, S. 2011, ApJS, 195, 4

Neyrinck, M. C., Gnedin, N. Y., & Hamilton, A. J. S. 2005, MNRAS, 356, 1222

Onions, J., Knebe, A., Pearce, F. R., et al. 2012, MNRAS, 423, 1200

Planelles, S., & Quilis, V. 2010, A&A, 519, A94

Plummer, H. C. 1911, MNRAS, 71, 460

Press, W. H., & Schechter, P. 1974, ApJ, 187, 425

Rasera, Y., Alimi, J. M., Courtin, J., et al. 2010, in American Institute of Physics Conference Series, 1241, eds. J.-M. Alimi, & A. Fuözfa, 1134

Riebe, K., Partl, A. M., Enke, H., et al. 2013, Astronomische Nachrichten, 334, 691

Springel, V. 2005, MNRAS, 364, 1105

Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726

Springel, V., Wang, J., Vogelsberger, M., et al. 2008, MNRAS, 391, 1685

Stadel, J. G. 2001, Cosmological N-body Simulations and Their Analysis, PhD Thesis, University of Washington

Sutter, P. M., & Ricker, P. M. 2010, ApJ, 723, 1308

Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, ApJ, 688, 709

Tormen, G., Diaferio, A., & Syer, D. 1998, MNRAS, 299, 728

Tweed, D., Devriendt, J., Blaizot, J., Colombi, S., & Slyz, A. 2009, A&A, 506, 647

Warren, M. S., Abazajian, K., Holz, D. E., & Teodoro, L. 2006, ApJ, 646, 881