The MultiDark Database: Release of the Bolshoi and MultiDark Cosmological Simulations

Riebe, Kristin[a], Partl, Adrian M.[a], Enke, Harry[a], Forero-Romero, Jaime[a], Gottlöber, Stefan[a], Klypin, Anatoly[b], Lemson, Gerard[c], Prada, Francisco[d], Primack, Joel R.[f], Steinmetz, Matthias[a], Turchaninov, Victor[e]

[a] *Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam (Germany)*
[b] *Astronomy Department, New Mexico State University (NMSU), Las Cruces, NM 88001 (USA)*
[c] *Max Planck Institut für Astrophysik (MPA), Karl-Schwarzschild-Str. 1, 85741 Garching (Germany)*
[d] *Instituto de Astrofisica de Andalucia (CSIC), Glorieta de la Astronomia S/N, 18008 Granada (Spain)*
[e] *Institute of Applied Mathematics, Miusskaya Sq.4, 125047, Moscow (Russia)*
[f] *Department of Physics, University of California, Santa Cruz, CA 95064 (USA)*

Abstract

We present the online *MultiDark Database* – a Virtual Observatory-oriented, relational database for hosting various cosmological simulations. The data is accessible via an SQL (Structured Query Language) query interface, which also allows users to directly pose scientific questions, as shown in a number of examples in this paper. Further examples for the usage of the database are given in its extensive online documentation. The database is based on the same technology as the Millennium Database, a fact that will greatly facilitate the usage of both suites of cosmological simulations. The first release of the *MultiDark Database* hosts two 8.6 billion particle cosmological N-body simulations: the Bolshoi ($250\,h^{-1}$Mpc simulation box, $1\,h^{-1}$kpc resolution) and MultiDark Run1 simulation (MDR1, or BigBolshoi, $1000\,h^{-1}$Mpc simulation box, $7\,h^{-1}$kpc resolution). The extraction methods for halos/subhalos from the raw simulation data, and how this data is structured in the database are explained in this paper. With the first data release, users get full access to halo/subhalo catalogs, various profiles of the halos at redshifts $z = 0 - 15$, and raw dark matter data for one time-step of the Bolshoi and four time-steps of the MultiDark simulation. Later releases will also include galaxy mock catalogs and additional merging trees for both simulations as well as new large volume simulations with high resolution. This project is further proof of the viability to store and present complex data using relational database technology. We encourage other simulators to publish their results in a similar manner.

Keywords: Simulation, Cosmology, Database, Virtual Observatory

Email addresses: kriebe@aip.de (Riebe, Kristin), apartl@aip.de (Partl, Adrian M.), henke@aip.de (Enke, Harry), jforero@aip.de (Forero-Romero, Jaime), sgottloeber@aip.de (Gottlöber, Stefan), aklypin@nmsu.edu (Klypin, Anatoly), lemson@mpa-garching.mpg (Lemson, Gerard), fprada@iaa.es (Prada, Francisco), joel@ucsc.edu (Primack, Joel R.), msteinmetz@aip.de (Steinmetz, Matthias), vturch@utech.ru (Turchaninov, Victor)

1. Introduction

Computer simulations play a very important role in cosmology. The field started in the 1960s and 1970s with $N-$body simulations which had just a few hundred or thousand particles (Aarseth, 1966, 1969; Peebles, 1970; White, 1976; Gott et al., 1979; Efstathiou and Jones, 1979). Thanks to the steady improvement of computer hardware and computational algorithms, we now have large simulations with many billions of particles (Springel et al., 2005; Boylan-Kolchin et al., 2009; Teyssier et al., 2009; Kim et al., 2009; Klypin et al., 2010; Wetzel and White, 2010; Prada et al., 2011; Iliev et al., 2011). These and other large $N-$body simulations are used to address numerous aspects of the evolution of fluctuations and formation of dark matter halos. They provide extraordinary accuracy for such important statistics as the mass function of halos (e.g., Jenkins et al., 2001; Sheth and Tormen, 2002; Warren et al., 2006; Tinker et al., 2008; Klypin et al., 2010), halo concentration (Bullock et al., 2001; Zhao et al., 2003; Neto et al., 2007; Macciò et al., 2008; Prada et al., 2011; Iliev et al., 2011), halo correlation function and biases (Jing, 1998; Kravtsov and Klypin, 1999; Gao et al., 2005; Wechsler et al., 2006; Tinker et al., 2010), statistics and distribution of satellites (Klypin et al., 1999; Moore et al., 1999; Springel et al., 2008; Kuhlen et al., 2008), and dark matter density profiles (Dubinski and Carlberg, 1991; Navarro et al., 1997; Springel et al., 2008; Stadel et al., 2009).

In spite of the fact that cosmological $N-$body simulations give information only on dark matter and do not mimic the evolution of baryons, there are different ways to make theoretical predictions for "galaxies" in these dark-matter-only simulations. For instance, one can use semi-analytical methods to predict properties of galaxies hosted by dark matter halos and subhalos (e.g., Kauffmann et al., 1999; Somerville and Primack, 1999; Croton et al., 2006; Somerville et al., 2011). Another option is to use the Halo-Occupation-Distribution (HOD; Kravtsov et al., 2004; Vale and Ostriker, 2004; Zentner et al., 2005; van den Bosch et al., 2007) to split halos into "galaxies". A third option is Halo-Abundance-Matching (HAM; Kravtsov et al., 2004; Conroy et al., 2006; Trujillo-Gomez et al., 2010) by matching the largest halos (and subhalos) with the brightest galaxies.

However, with the growing size of numerical simulation data some problems evolved. The sheer amount of data in large simulations is difficult to handle. The simulations typically provide so much data that even a compact form of different "catalogs" may become impractical to distribute to all people involved in different research groups, not to mention to provide it to the whole astronomical community, or to release raw simulation data. Accessing data written in different formats puts an extra burden on people who intend to analyze the results of the simulations.

There are different ways of handling the situation. We decided to use a database, which, in combination with its powerful *Structured Query Language* (SQL), allows users to filter the data on the server side, analyze the resulting subsets of the data, and retrieve only their results. Since the amount of data these simulations produce lies nowadays in the multi-terabyte range, the full data set is generally too large to retrieve and manage for most users. Server side filtering is a prerequisite for successful dissemination of such large data sets.

Large observational surveys like the SDSS have pioneered this approach in astronomy and have shown that providing data directly through SQL is a very fruitful approach (skyserver.sdss.org, 2008). The *Millennium Database* (Lemson and Virgo Consortium, 2006) played an important role by making the Millennium simulations (Springel, 2005; Boylan-Kolchin et al., 2009) accessible to numerous users. The *MultiDark Database* uses the same technology, implementation and data structures as the *Millennium Database*, a fact that will greatly facilitate the usage of both databases to study consistency of dark matter halo statistics in simulations

performed using different codes, numerical algorithms, and halo finders.

This paper is structured as follows: Section 2 gives an overview on the database design and describes the methods for accessing the data. In Section 3 we characterize the current simulations in the database. More details on how the data of these simulations are stored in the database are given in Section 4. We complement our presentation of the *MultiDark Database* with a few example science cases in Section 5 and a short summary given in Section 6. Appendixes additionally provide descriptions of the employed halofinders and the merger-tree construction.

2. The MultiDark Database and its design

Databases organize large amounts of structured data for efficient retrieval. The *MultiDark Database* actually is a "relational database". Such relational databases organize the data in collections of tables (originally called "relations"), which in our case store the different objects identified in the simulations and derived data products. For instance, the table containing the main Friends-of-Friends halo catalogue (`FOF`) consists of one record (row) for each FOF group, with its properties mapped to the columns of the table.

The strength of relational databases lies not only in capturing the data itself, but in modeling possible connections between the various datasets. Connections between the various tables are achieved by linking rows of different tables with an unique identifier, which points from a row in one table to another row in the other table. Such "foreign keys" establish links between the various tables in the simulation database. For instance, the table containing the main `FOF` catalogue contains a column with a label *fofId* for each FOF group. This *fofId* is used again in the `FOFParticles` table, which lists the simulation particles that constitute the FOF group (see Section 4.5 for more details).

Another important feature of relational databases is the powerful SQL query language supported by them. As already mentioned above, SQL is used to filter the main data products and retrieve exactly those subsets one is interested in. SQL queries are expressed in terms of the tables and their interrelations in the database[1]. These queries are interpreted by the database engine and compiled into execution plans that access the data in an optimal way to retrieve the results. The language therefore allows users – especially those not intimately familiar with the data format of the simulation – a far more direct path from a science question to an executable expression than a standard scripting or programming language would. I/O, looping, optimization etc. are all handled by the database engine and the user has no need to know about this. Moreover, the abstraction provided by the relational model provides users with a more uniform interface even between different databases than using file based access.

As will be explained in Sections 3 and 4, the *MultiDark Database* now contains two simulations, each in a separate database with several database tables. Users retrieve data from these tables by performing queries using SQL. In our case, Microsoft's SQL dialect T-SQL (MSDN, 2008) is required (as is the case for the Millennium Database and the SDSS SkyServer). This allows users to employ some extensions, e.g. stored procedures to perform often used parts of queries. Furthermore, the Spatial3D library (Lemson et al., 2011), a collection of custom data types, procedures and functions written in C# is provided, which simplifies and substantially speeds up queries for rectangular, spherical, and more general volume shapes.

[1] For an overview and tutorial of this language see for example http://www.w3schools.com/sql/default.asp. See also the demo video on http://www.multidark.org/Help?page=demo/index.

In order to increase the efficiency of queries on large tables, indexes for already known query patterns were generated: for the unique identifiers linking the tables, the mass where applicable, and spatial coordinates. More indexes will be added for further use cases, and for the most common query patterns of the *MultiDark Database* as they emerge.

Interactive data access is provided via the MultiDark website[2], at `http://www.multidark.org`. Users should register using the web form at this page, because registered users obtain full access to all data. However, similar to the *Millennium Database*, unregistered access is enabled to a public mini-version of the MDR1-database, featuring all tables of MDR1 for a subvolume of the MultiDark simulation of about $(100\,h^{-1}\mathrm{Mpc})^3$ in the `miniMDR1` database. This allows the interested user to get an overview of capabilities of the *MultiDark Database* before registering. The miniMDR1 database also serves as a test and development environment for more elaborate queries.

The web application is designed for interactive use. SQL queries are submitted directly via the Query Form and results are either viewed in the browser, plotted with VOPlot[3] or retrieved in various other formats (e.g. CSV-table, VOTable[4]). This interactive interface has its limitations though, since browsers do not react kindly to the task of e.g. rendering some megabytes of ASCII-text. Therefore, registered users can store query results into their private database for further use. As already pointed out, SQL queries can take a long time, and not well formed queries do this often. Therefore, a limit on the query time and the private table space is imposed. Extending private table space or time limits for longer running queries is possible by contacting support.

Scripted access, either for retrieving large results of queries using the UNIX tool `wget`, or for use with graphical tools like `Topcat`, `IDL`, or for doing statistical analyses of radial profiles is provided by another servlet available at `http://wget.multidark.org/MyDB`. Again, the documentation provides many examples and usage hints. A web-page with often used queries is also available.

3. Simulation Data

The Data Release 1 of the *MultiDark Database* contains data of two different cosmological simulations. For each of these simulations separate databases exist, so that further simulations and post-processing results can be incorporated easily. The MultiDark Run1 and Bolshoi simulations are complementary to the Millennium I and II simulations. All four simulations follow the clustering of roughly the same number of dark matter particles (8-10 billion) but within different simulation volumes, using different cosmological parameters and different cosmological codes. In tables 1 and 2 we summarize and compare these parameters.

The Bolshoi simulation (from Russian "grand, great") has a simulation volume of $(250\,h^{-1}\,\mathrm{Mpc})^3$ and contains $2048^3 \approx 8.6 \cdot 10^9$ particles (Klypin et al., 2010). It has been performed in 2009 at the NASA Ames Research center. The underlying cosmological parameters are compatible with the WMAP5 and WMAP7 data, for a discussion see (Klypin et al., 2010). These parameters are listed in Table 1 in comparison to the cosmological parameters used for the Millennium runs. The Bolshoi simulation has been performed using the Adaptive Refinement Tree (ART) code (Kravtsov et al., 1997). The code was parallelized using MPI libraries

[2]The interface was developed within the German Astrophysical Virtual Observatory (GAVO, 2008).

[3]`http://vo.iucaa.ernet.in/~voi/voplot.htm`

[4]`http://www.ivoa.net/Documents/VOTable/`

Table 1: Cosmological parameters of different simulations

Parameter	MDR1/Bolshoi	Millennium	Description
h	0.70	0.73	Hubble parameter
Ω_Λ	0.73	0.75	density parameter for dark energy
Ω_m	0.27	0.25	density parameter for matter (dark matter+baryons)
Ω_b	0.0469	0.045	density parameter for baryonic matter
n	0.95	1.0	slope of the power spectrum
σ_8	0.82	0.9	normalization of the power spectrum

and OpenMP directives (Gottlöber and Klypin, 2008). The simulation is described in detail in (Klypin et al., 2010).

The MultiDark Run1 simulation (MDR1) (Prada et al., 2011) was performed in 2010 at the NASA Ames Research center. This simulation is designed to study galaxy clustering for the SDSS-III/BOSS survey. It contains the same number of particles as the Bolshoi simulation but in a (1 Gpc h^{-1})^3 cube and takes the same cosmological parameters given in Table 1. Its numerical parameters are summarized in Table 2.

Table 2: Numerical parameters of the cosmological simulations.

Parameter	MDR1	Bolshoi	Millennium-I	Millennium-II	units
Box size	1000	250	500	100	h^{-1} Mpc
Number of particles	2048^3	2048^3	2160^3	2160^3	
Mass resolution	8.721	0.135	0.86	0.0069	10^9 h^{-1} $M_\odot$
Force resolution	7.0	1.0	5.0	1.0	h^{-1} kpc
Initial redshift	65	80	127	127	

3.1. Halo catalogues

One of the main products derived from cosmological simulations are halo catalogues, which are then used for further analysis. They contain dark matter halos (or clusters of dark matter particles) and their intrinsic properties, like position, velocity, mass and radius. Several different techniques for finding and defining such halos were developed. Two of those halo finders were applied to the data in the *MultiDark Database* and are briefly described in the appendixes A and B. The *MultiDark Database* provides results from the BDM (Bound Density Maximum) halo finder (Appendix A) which uses a spherical 3D overdensity algorithm to identify halos and subhalos. Additionally, results from the FOF halo finder (Appendix B) are provided in the database as well. The FOF halo finder uses the relative linking length - given in terms of the mean inter-particle distance - to uniquely define clusters of particles. Built on the FOF-catalogues the *MultiDark Database* further contains merger trees (see Appendix C), for tracing the history of halos. BDM-based merger trees will be provided in a later data release.

In the following sections detailed descriptions of the various halo catalogues contained in the *MultiDark Database* are given.

3.1.1. BDM catalogues in the database

Two different BDM catalogues were produced for different definitions of the halo radius:

- **BDMV**: the virial mass $M_{\rm vir}$ is defined by the solution of the top-hat model of the growth of fluctuations in an expanding Universe with a cosmological constant. We define the virial radius $R_{\rm vir}$ of halos as the radius within which the mean density is the virial overdensity $\Delta_{\rm vir}(z)$ times the mean universal matter density $\rho_{\rm m} = \Omega_{\rm m}\rho_{\rm crit}$ at that redshift. Thus, the virial mass is given by

$$M_{\rm vir} \equiv \frac{4\pi}{3}\Delta_{\rm vir}\rho_{\rm m}R_{\rm vir}^3 \,. \tag{1}$$

  For our set of cosmological parameters, at $z = 0$ the virial radius $R_{\rm vir}$ is defined as the radius of a sphere with an overdensity of 360 of the average matter density. The overdensity limit changes with redshift and asymptotically goes to 178 for high z. The overdensity $\Delta_{\rm vir}(z)$ is given by an approximation provided by (Bryan and Norman, 1998).

- **BDMW**: the halo radius is defined by the overdensity limit $\Delta_{200} = 200\,\rho_{\rm crit}$. For our set of cosmological parameters this corresponds to $740\,\rho_{\rm m}$ at $z = 0$. It approaches asymptotically the overdensity of $200\,\rho_{\rm m}$ at high redshifts. Since this density is always larger than the virial one, the halos of the BDMW catalogue are always smaller than the corresponding halos in the BDMV catalogue.

Since both halo definitions are commonly used in the literature, the BDM catalogues for both values are given.

The BDM halo catalogues provide a lot of information: each halo and subhalo is characterized with 23 parameters. The list of these parameters is given on the website of the *MultiDark Database* and is also described in Appendix A. In addition to coordinates, peculiar velocities and other halo properties, the database provides two masses: the mass of all particles inside the virial radius and the mass of gravitationally bound particles. For distinct halos the difference between the two masses is typically at most 1-2 percent. The difference is much larger for subhalos. Note that most of the parameters of both subhalos and distinct halos are defined by gravitationally bound particles.

3.1.2. FOF catalogues in the database

The nature of the FOF algorithm implies that FOF groups cannot intersect with each other (Figure B.9), which means that for a given linking length any particle is uniquely assigned to just one FOF group (such a FOF group could be the particle itself). With this property it is possible to create database tables to establish a link between FOF groups and their particles (see Section 4.5). Furthermore, substructures always lie completely within their host structure, since they are defined by smaller linking lengths. Due to the unique mapping of a particle to a FOF group one can determine unique progenitor-descendant-relations of FOF groups as the basis for the construction of the merger tree.

For the resulting FOF groups no post-processing has been applied so far. In particular, no binding/unbinding procedure was applied, i.e. a FOF group consists not necessarily of bound particles. However, for a given FOF group all particle positions and velocities can be extracted from the database for any post-processing.

Table 3: Linking lengths for the FOF catalogues provided in the database. The corresponding database table names are given in the first column, a more detailed description of these tables is provided in Table D.4. The linking lengths are also stored directly in tables `linkLength` and `linkLengthScl`. The last column contains a characteristic overdensity for the given linking length.

Database table	linking length (in units of interparticle separation)	level/sclevel in database table	overdensity
FOF	0.17	0	570.
FOF1	0.085	1	3100.
FOF2	0.0425	2	19000.
FOF3	0.02125	3	1.2×10^5
FOF4	0.010625	4	9.8×10^5
FOFc	0.20	-	390.
FOFScl	0.35	0	94.
	0.32	1	120.
	0.29	2	160.
	0.26	3	210.
	0.23	4	280.
	0.20	5	390.
	0.17	6	570.

4. Data in the MultiDark Database

The following subsections describe the available data of the MDR1 and Bolshoi simulation, how they are organized in the tables of the database, and some access examples. The names of the corresponding tables are given in brackets at each section title. A more complete overview of all tables and their relations can be found in Appendix E, Figure E.10.

4.1. Halo catalogues - (BDMV, BDMW, FOF{1, 2, 3, 4}, and FOFc)

Each BDM and FOF halo catalogue has a corresponding table in the database. The BDM halo catalogues are `BDMV` and `BDMW`. For the `FOF` tables, the numbers denote the different linking lengths used in the halo finding procedure. Furthermore, `FOFc` denotes the FOF halo catalogue with the commonly used linking length of 0.2. A complete list of the linking lengths for the various FOF halo catalogues is given in Table 3. For a given linking length the overdensity of the FOF objects depends on the concentration of the objects and therefore on mass and redshift (More et al., 2011). Moreover, it shows a large scatter. In order to give an idea of the expected overdensity for the given linking length, the last column in the table contains a characteristic overdensity corresponding to the linking length given in the second column. The halo catalogue tables contain individual records for each dark matter halo or FOF group, with all calculated properties given in the corresponding columns.

Additionally, spatial grid indexes (1024 cells per dimension) are provided, together with the computed Peano-Hilbert key for each grid cell using the Spatial3D library (Lemson et al., 2011), which enables a fast retrieval of halos or FOF groups from a given region in space.

Another feature of the database is the quick retrieval of snapshot number and mass columns, or sorting of halos/FOF groups by their mass. This is important for e.g. calculating the mass function of halos and its evolution in time, and can be done easily using database queries.

4.2. Halo profiles - (BDMVprof and BDMWprof)

For BDM halos, access to their inner structure is possible with the BDM profiles stored in tables `BDMVprof` and `BDMWprof` (`V` and `W` are defined as in Section 3.1.1). These profiles consist of logarithmically spaced shells as a function of the virial radius $R_{\rm vir}$ of the halo. They are available up to a radius of $2R_{\rm vir}$ and cover halos with more than 100 particles. Each radial bin corresponds to a row in the table and includes physical properties like e.g. local density and circular velocity. Each property is given once for all the particles enclosed by the shell, and once for the bound particles only. The halo profiles tables allow the user to study density profiles, rotation curves and other typical properties of the BDM halos.

The profile records are linked to the corresponding BDM halo from the `BDM`-table by the halo's unique identifier *bdmId*. To obtain the profile of a specific halo, one first retrieves the halo's unique identifier *bdmId* and then searches for all the entries in the profile table with the corresponding *bdmId*.

4.3. Merger trees - (FOFMtree)

Within the currently accepted ΛCDM cosmology, dark matter halos merge from small clumps to ever larger objects. This merging history can also be traced in cosmological simulations and is stored in the form of merger trees (see illustration, Figure 1). These merger trees provide a description of the assembly history of a dark matter halo and can be used as a basis for a semi-analytic model describing the baryonic properties of the galaxies evolving in a halo.

Merger trees are built for a subset of halos (or FOF groups) which exist at redshift 0 and exceed a certain mass limit. From such a root-halo (at the top in Figure 1), the branches go to each of its progenitors, reaching backwards in time, with the most massive progenitor being visited first (main branch). A typical question is to retrieve merger trees rooted in a given halo or set of halos. To answer such queries efficiently the same algorithm is employed as for the Millennium database (Lemson and Springel, 2006). Once the tree is built, its nodes are sorted according to a depth-first search. This depth-first order rank is used to construct unique identifiers and specific foreign keys, which enable quick access to the complete merger history for each halo.

The merger-tree tables contain the following merger tree identifiers for each FOF group:

- *treeRootId*: ID of the top node in the merger tree, i.e. the final descendant at redshift $z = 0$ (root halo), calculated based on the number of the halo/FOF group in the corresponding halo catalogue (start line number with 0):

$$\textit{treeRootId} = (\text{rank in file} + 1) \cdot 10^8 \tag{2}$$

- *fofTreeId*: unique identifier for each FOF group, based on *treeRootId* and rank in depth-first order:

$$\textit{fofTreeId} = \textit{treeRootId} + (\text{rank in merger tree}) \tag{3}$$

  Thus, the tree membership is encoded directly in the *fofTreeId* for each group.

- *descendantId*: identifier (*fofTreeId*) of the direct descendant of a FOF group (i.e. forward in time, into which the FOF group will grow/merge)

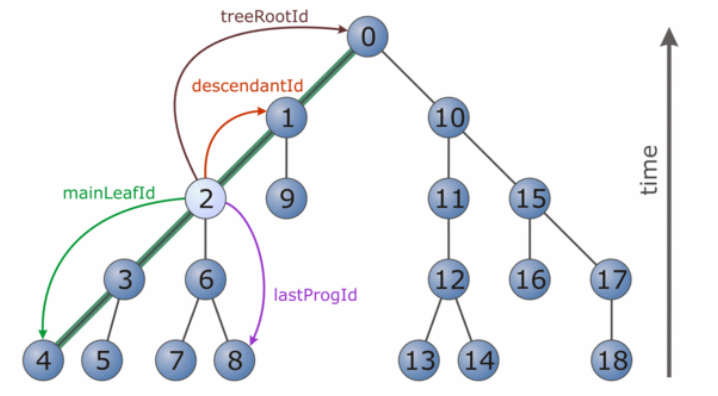

Figure 1: Merger tree: the top node (root) of the tree represents a halo or FOF group at redshift $z = 0$. From there, branches reach backwards in time to its progenitors, i.e. the timeline goes from bottom to top. The numbers at each node indicate the depth-first order, with the most massive progenitors being on the leftmost side of each sub-tree. These form the main branch (e.g. the thick green line for the tree root (0)) of the corresponding node. The identifiers (ids) drawn here for one example node (2) are stored in the database table (see text for further explanations).

- *mainLeafId*: identifier (*fofTreeId*) of the *last* FOF group of the main branch, along the most massive progenitors; enables a quick retrieval of e.g. the accretion history of a FOF group etc. by querying for all progenitors until the FOF group with the *mainLeafId* as *fofTreeId* is encountered. If the halo of the top node in Figure 1 is denoted as *topHalo*, a schematic query for the main branch of this halo would look like:

  ```
  select * from FOFMtree
      where fofTreeId between topHalo.fofTreeId and topHalo.mainLeafId
  ```

  and returns the records for FOF groups no. 0, 1, 2, 3 and 4 (for the complete query, consult the "Very useful queries" Nr. 5.2 on the *MultiDark Database* webpage).

- *lastProgId*: identifier (*fofTreeId*) of the *last* FOF group in the tree; queries for all FOF groups with *fofTreeId* between the *treeRootId* and the *lastProgId* will return the complete merger tree.

The merger trees are determined for all halos with more than 200 particles at redshift $z = 0$. They end at a certain redshift if the main progenitor of a given halo is below the detection threshold of 20 particles. Fig. 2 shows for three different mass bins of halos at $z = 0$ the fraction of halos for which the main branch of the merger tree can be followed down to redshift $z_{\rm max}$.

4.4. Substructures - (FOFScl, BDMV, and BDMW)

Small dark matter (sub)halos are embedded in larger ones, which in turn may reside in even bigger halos. Once these multi-level subhalos are found, their hierarchical structure can be represented by a substructure tree, in much the same way as the build-up process of halo formation is usually expressed with merger trees. However, this is only possible where the information of sub-substructures is available, as for FOF groups of different linking lengths. For the BDM catalogues only the *bdmId* of the host halo for each subhalo (in column *hostFlag*) is available.

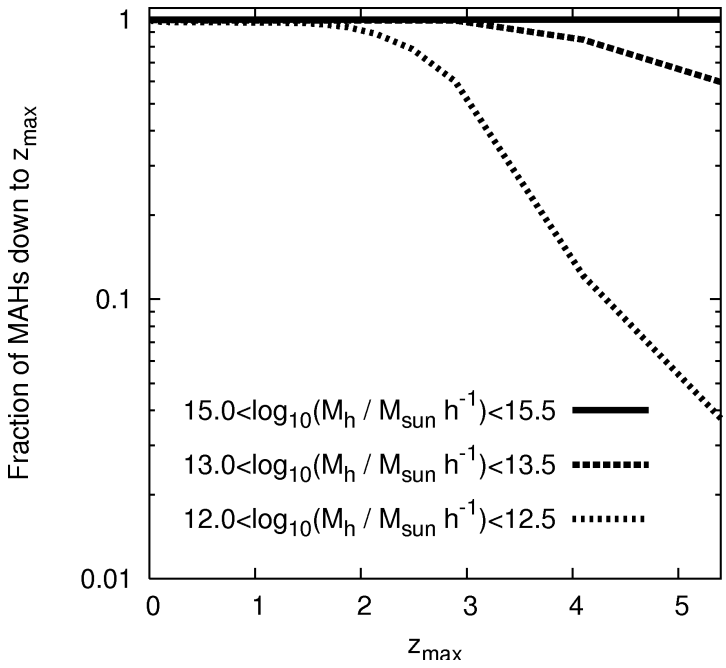

Figure 2: The fraction of FOF merger trees for the MultiDark simulation having a main branch followed down to a given redshift for three different mass bins. All trees have been constructed down to a maximum redshift of $z = 5.4$.

For the FOF halo catalogues substructure trees at redshift $z = 0$ are provided. In such a substructure tree the root node corresponds to the biggest halo (i.e. with lowest density threshold, largest linking length), followed by successively smaller halos in the next substructure level. The FOF (sub)halos of the tree are sorted in a depth-first order, with the most massive substructure as the first node of each new level, so that the main branch can be retrieved in the same way as for a merger tree (see Figure 3 and Section 4.3). The necessary tree identifiers are constructed like those for the merger trees and stored in the database tables `FOFSub` and `FOFScl`. They are only renamed to fit the substructure context (also see Figure 3):

- *fofTreeId* ↔ *fofSubId*
- *descendantId* ↔ *hostId*
- *lastProgId* ↔ *lastSubId*

4.5. Simulation particles - (`particles` and `FOFParticles{1, 2, 3, 4}`)

The *MultiDark Database* contains not only halo catalogues and many related data sets like substructures and merger trees, but as one of its main new features also the complete raw simulation data at certain redshifts. For each of these snapshots the full set of 8.6 billion particles is available along with their positions and velocities[5].

[5]The *Millennium Database* provides the particles for all 64 snapshots of the smaller (20 million particles) milli-Millennium database.

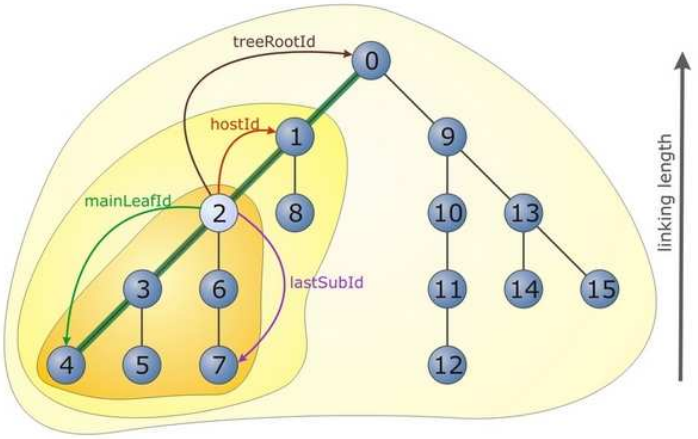

Figure 3: Substructure tree: the top node as the root of the tree represents the biggest object (with largest linking length). Each row contains FOF groups of smaller linking lengths, which are substructures of their host(s). Additionally FOF substructures are sorted by mass from left to right for each FOF group. The numbers indicate the ranking of FOF groups in a depth-first ordering. The thick green line marks the main branch of the tree root (0). The identifiers are constructed and used in the same way as for merger trees (see Section 4.3).

This particle information can be used to study the particle distribution in certain regions, e.g. in the environment of a selected dark matter halo. For accelerating such spatial queries the *Spatial3D library* is used (Lemson et al., 2011). This library is written in C# and its functions and data types are available from within T-SQL. It employs a Peano-Hilbert space-filling curve subdividing the box into a 1024^3 grid. Each particle has a column describing the grid cell it is in. The same procedure was applied to the halo catalogues. The use of the library is not completely standard and example queries are provided on the web site.

For the FOF catalogues, stored in tables `FOF` – `FOF4` and `FOFScl`, additional tables contain particles of the snapshot at redshift $z = 0$ linked to their corresponding FOF halos (tables `FOFParticles` – `FOFParticles4`). This allows users to easily extract particles for a given FOF halo and e.g. calculate additional properties, or recalculate some given quantities with alternative methods[6]. The `FOFParticles`-tables can even be used to cross-check the substructure information given in table `FOFSub`: a substructure of a FOF halo always lies completely within the host halo (since it has smaller linking length) and thus each of its particles also belongs to its host. By joining `FOFParticles`-tables for different linking lengths, one can get a list of substructures (or hosts) for a given FOF halo, independent of the `FOFSub`-table[7].

The information for the particles of a BDM halo cannot be retrieved as directly as for the FOF halos. Since no table linking BDM halos with their particles is currently provided, all the particles in the halo's region (up to its virial radius) are only accessible by using the spatial coordinates of the halo and the *Spatial3D library* to exctract a region around the halo's center.

[6]The same method of linking particles to their halos has been applied for the MMSnapshots of the Millennium database.

[7]However, such queries become often quite expensive in terms of compute time, so it is recommended to use the `FOFSub`-table for extensive substructure studies.

5. Examples of using the database

The *MultiDark Database* enables users to analyze many aspects of cosmology and galaxy evolution. It will also help to interpret large state-of-the-art observational data sets. The following list gives some possible examples of analysis using data in the database:

- properties of halos (radial profile, concentration, shapes),
- evolution of the number density of halos, essential for normalization of Press-Schechter-type models,
- evolution of the distribution and clustering of halos in real and redshift space, for comparison with large-scale galaxy/QSO surveys,
- accretion history of halos, assembly bias (variation of large-scale clustering with assembly history), and correlation with halo properties including angular momentum and shape,
- halo statistics including the mass and velocity functions, angular momentum and shapes, subhalo numbers and distribution, and correlation with environment,
- void statistics, including sizes and shapes and their evolution, and the orientation of halo spins around voids.
- quantitative descriptions of the evolving cosmic web, including applications to weak gravitational lensing,
- preparation of mock catalogs, essential for analysis of SDSS and other new survey data (SDSS-III/BOSS, DES, Planck),
- merger trees, essential for semi-analytic modeling of the evolving galaxy population, including models for the galaxy merger rate, the history of star formation and galaxy colors and morphology, the evolving AGN luminosity function, stellar and AGN feedback, recycling of gas and metals, etc.

Here we give some examples in more detail.

5.1. Example 1: Velocity function

A novel feature of the *MultiDark Database* is the access to the profiles of different physical parameters (density, velocity, etc) for each of the halos found in the BDM tables. In particular, in this example we show how to obtain the average radial velocity profile for halos of vastly different masses, from galaxy-size halos to clusters. We used many hundreds of halos for each mass range. This simple query will allow users to study the infall of material beyond the formal virial radius. For group- and cluster-sized halos there are large infall velocities, whose amplitude increases with halo mass. No infall is seen for galaxy size halos as reported by (Prada et al., 2006).

Example 1 can be written using the following SQL statement:

```
with halos as (
    select  Mvir, Rvir, bdmId
    from MDR1..BDMV
    where snapnum=85 and Mvir between 1e12 and 1.1e12
)
select  power(10.000000,(0.05*(0.5+floor(log10(p.R_Rvir)/0.05)))),
        avg(
            p.Vrad /
              ( sqrt(
                     6.67428e-8 * halos.Mvir * 1.988e33 /
                     (halos.Rvir * 3.0856e24)
                 ) / 100000
              )
        )
from halos, MDR1..BDMVprof p
where p.bdmId = halos.bdmId
group by floor(log10(p.R_Rvir)/0.05)
```

Using the data retrieved with this statement, Figure 4 was generated:

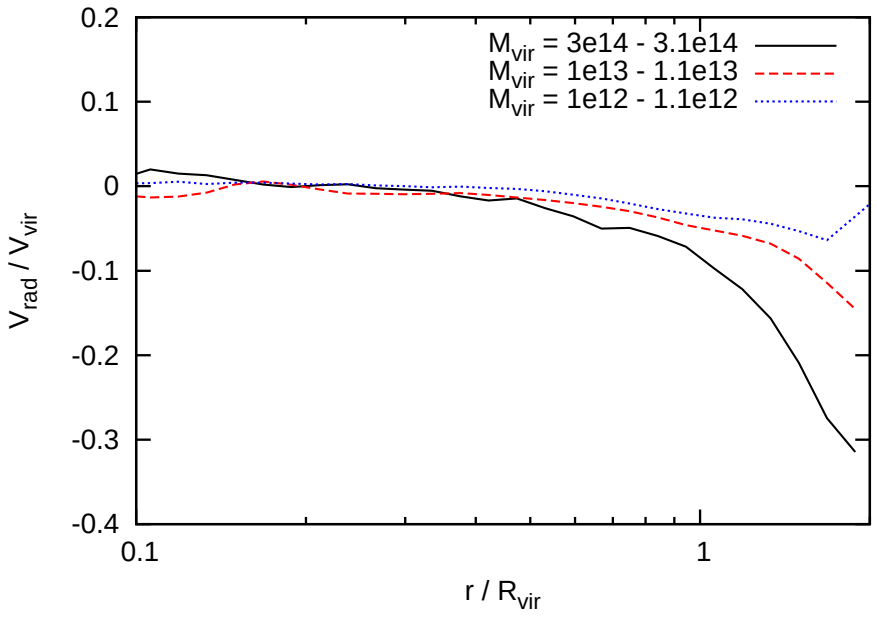

Figure 4: The plot shows the data retrieved with the statement in Example 1. We show average radial velocities for halos with different virial masses. The velocities are practically zero within 1-2 $R_{\rm vir}$ for halos with mass $10^{12}h^{-1}M_{\odot}$. The situation is different for group- and cluster-sized halos. For these massive halos significant infall velocities are found. Their amplitude increases with halo mass.

5.2. Example 2: Access to particles

Another novel feature of the *MultiDark Database* is the access to the complete particle data of snapshots. As an example we will retrieve all particles which belong to the largest FOF object (*supercluster*) found at the largest linking length, $ll = 0.35$, in the *supercluster* table of the MultiDark simulation. This object has a low overdensity of about 94 (see Table 3) and consists of 791743 particles. Its mass is $6.9 \times 10^{15} h^{-1} M_{\odot}$. These particles have been extracted from the database using the following query:

```
with mostMassiveCluster as (
    select top 1 * from
            MDR1..FOFScl
        where snapnum=85 and sclevel=0
        order by mass desc
),
fofClustParticles as (
    select fP.* from
            MDR1..FOFSclParticles_85_l0 fP,
            mostMassiveCluster mC
        where fP.fofId = mC.fofSubId and fP.snapnum=85
)
select p.* from
        fofClustParticles hP,
        MDR1..particles p
    where p.particleId = hP.particleId
```

The results of this query – positions and velocities of dark matter particles – are retrieved by the database system in less than one minute. Knowing the position and velocities of all these particles one can start individual post-processing. As an example the particles of the most massive "supercluster" in the MultiDark simulation are plotted in Figure 5 in three different projections. In this figure, the logarithm of the projected density in a grid of cell size $130 h^{-1}$ kpc was plotted. The left side of this figure shows the density distribution, whereas the right side shows the objects which the AHF halo finder (Knollmann and Knebe, 2009) has found in this particle distribution. The database allows the user to download particles of one or many objects defined at a certain mean overdensity and to run his own analysis tools, e.g. any kind of halo finder.

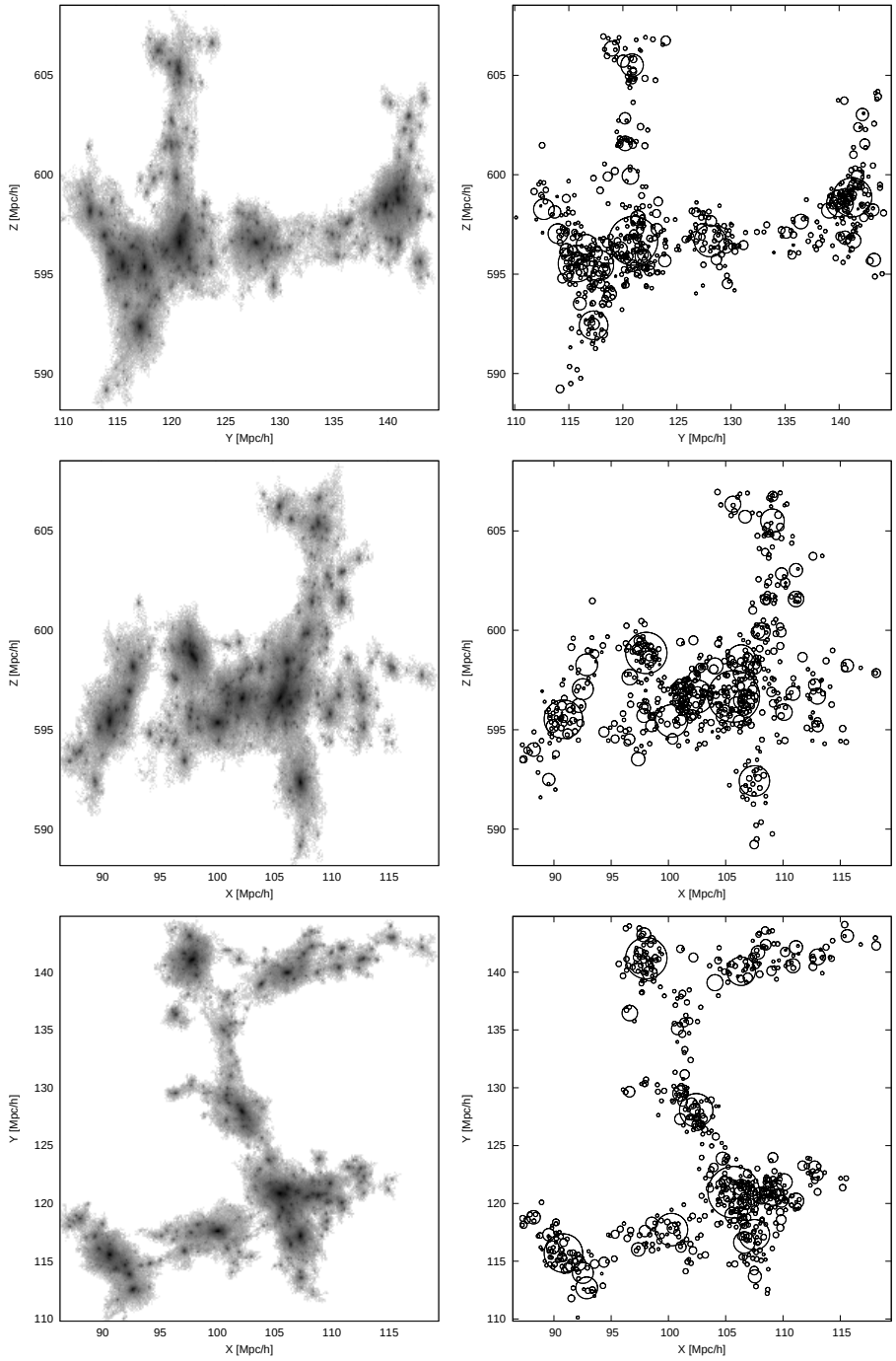

Figure 5: The left panes show density projections of the most massive supercluster in the MultiDark simulation at the highest linking length, $ll = 0.35$. The logarithm of the projected density in a grid of cell size $130h^{-1}$ kpc is plotted. The right panes show the objects which the AHF halo finder identified in the same volume. The circle's radii mark one $R_{\rm vir}$ as reported by AHF.

To analyze all objects above a certain mass threshold, all the corresponding particles from the detected "superclusters" can be downloaded. As an example of such an analysis, Figure 6 shows the cumulative fraction of particles found in FOF halos defined at overdensity 94 with masses larger than a given mass. One can also use the downloaded particles to test or to apply ones' own halo-finder. Since the mean overdensity is much lower than the virial one, these FOF halos contain (for a given mass) a complete set of spherical halos at virial overdensity. For example, to find and analyze all spherical halos with $m_{\rm vir} > 10^{15}h^{-1}M_{\odot}$ it would be sufficient to download all particles from superclusters with $m_{\rm scl} > 10^{15}h^{-1}M_{\odot}$, i.e. only about 2 % of all the particles (see Figure 6). A query which retrieves all the particles of 1000 halos with $m_{\rm scl} > 10^{15}h^{-1}M_{\odot}$ requires 6 hours and 10 minutes. Such a query needs to be split into individual queries for each halo due to the timeout limits.

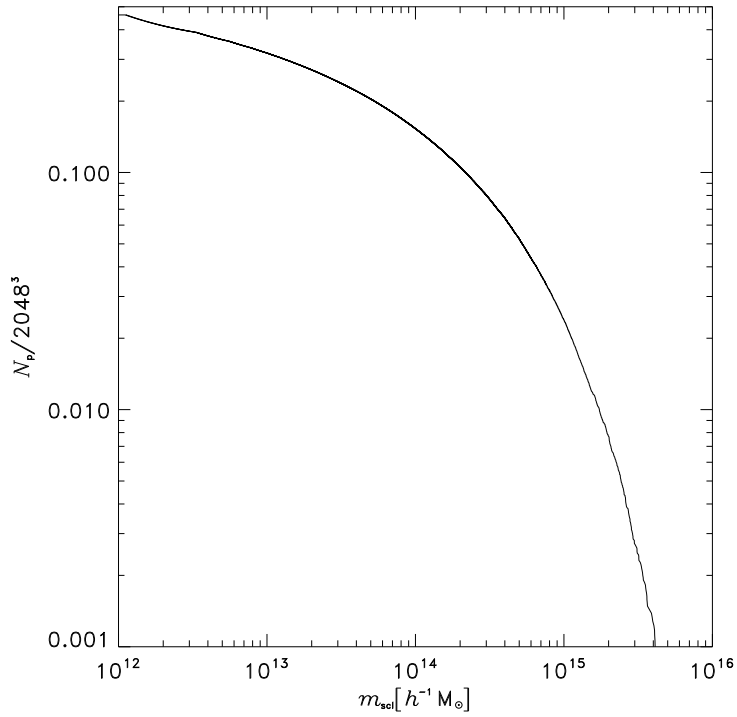

Figure 6: Cumulative fraction of particles in the MultiDark simulation at $z = 0$ in FOF halos with masses larger than $m_{\rm scl}$. A large linking length $ll = 0.171h^{-1}$ Mpc was used in this case, corresponding to a mean overdensity of 94.

5.3. Example 3: Halo Mass Function

Another very powerful feature of relational database systems is data aggregation, such as calculating the sum of a given data set, counting the number of data entries, or generating averages. For illustration and to show the strength of aggregation functions, the halo mass function is determined from the Bolshoi simulation.

Halo mass functions have been extensively studied (e.g., Jenkins et al., 2001; Tinker et al., 2008) to obtain insight into hierarchical structure formation and the build up of virialized objects. The mass function is also a key ingredient in many semi-analytical models (e.g., Somerville and Primack, 1999; Croton et al., 2006; De Lucia and Blaizot, 2007; Bower et al., 2007). In order to extract the halo mass function from the Bolshoi dataset at a given redshift and for a given halo catalogue, the following SQL statement is executed:

```
declare @boxSizeQubed as int;
set @boxSizeQubed = 250*250*250;
with redZ as (
      select snapnum
      from Bolshoi..redshifts
      where zred = 0.0
)
select  0.1 * (0.5 + floor(log10(f.mass) / 0.1)) as log_mass,
        count(*) / 0.1 / @boxSizeQubed as num from Bolshoi..FOF f,
        redZ
where f.snapnum = redZ.snapnum
group by floor(log10(f.mass) / 0.1)
order by log_mass
```

For the BDM catalogs an additional constraint selecting only distinct haloes (i.e. `f.hostFlag=-1`) needs to be added to the `where` clause of the statement.

With the results of this query, the halo mass function for the FOF and BDM halo catalogues at three different redshifts was plotted in Figure 7.

6. Summary

We present the *MultiDark Database* – a new facility to host and analyze large cosmological simulations. The first data release makes the results of two 8.6-billion particles cosmological *N*-body simulations – Bolshoi (Klypin et al., 2010) and MultiDark Run1 (Prada et al., 2011) – available for the astronomical community. Data from these simulations are organized in a relational database and are accessible through a simple web interface. SQL can be efficiently used to pose scientific questions, as shown in the examples. The same technology based on an abstraction of the data in terms of tables and relations greatly facilitates their usage and enables comparisons. It also makes the data sets fit for dissemination by standards developed in the International Virtual Observatory Alliance[8] (IVOA). In particular the *Table Access Protocol*[9] (TAP) targets the publication of, and interoperability between, datasets stored in relational databases.

For a future data release, it is planned to include raw data for more snapshots at different redshifts. We also plan to give access to galaxy mock catalogs for both simulations. These mocks are based on the halo abundance matching technique (see Trujillo-Gomez et al., 2010). Providing galaxy mock catalogs to the astronomical community is essential for analyzing large scale galaxy surveys (such as SDSS-III/BOSS, DES, Pan-Starrs), and for planning new experiments for dark energy. Finally, we plan to add at least the data of one more simulation in a larger volume than MultiDark Run1.

[8] `http://www.ivoa.net`

[9] http://www.ivoa.net/Documents/TAP/

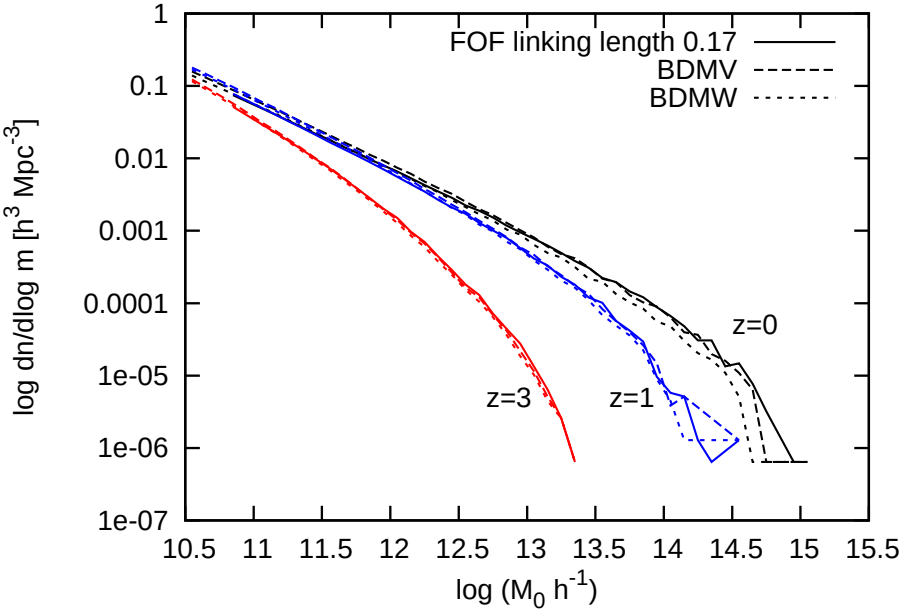

Figure 7: Halo mass function for the Bolshoi simulation derived from the FOF halo catalog with linking lengths 0.17 and the BDMW catalogs. The mass function is shown at three different redshifts $z = 0$ (black lines), $z = 1$ (blue lines), and $z = 3$ (red lines).

Acknowledgements

The Bolshoi and MultiDark (BigBolshoi) simulations were run on the NASA's Pleiades supercomputer at the NASA Ames Research Center. AK, JP, and SG are grateful to the staff of the NASA Ames Research Center for helping us with the simulations, and assisting with the analysis and visualization of the outputs. We acknowledge support of NASA and NSF grants to NMSU and UCSC for supporting this part of our research.

We thank S. Knollmann for valuable help with the AHF halo-finder and Tamás Budavari for his contribution to the Spatial 3D Library. We acknowledge the support of the Spanish MICINN Consolider-Ingenio 2010 Programme under grant MULTIDARK CSD2009-00064. We acknowledge the MoU between MultiDark and AIP for the construction of the *MultiDark Database*. The *MultiDark Database* relies on the *Millennium Simulation Database* implementation and its web application for online access, which were created by the German Astrophysical Virtual Observatory (GAVO). GAVO is funded by the German Ministry of Education and Research (BMBF).

Appendix A. Bound Density Maximum (BDM) halofinder

The basic technique of the BDM halo finder is described in Klypin and Holtzman (1997). The code was subject to major improvements since 1997. It uses a spherical 3D overdensity algorithm to identify halos and subhalos. It starts by finding the density for each individual particle. The density is defined using a top-hat filter with a given number of particles $N_{\rm filter}$, which typically is $N_{\rm filter} = 20$. The code finds all density maxima, and for each maximum it finds a sphere containing a given overdensity mass $M_\Delta = (4\pi/3)\Delta\rho_{\rm crit}R_\Delta^3$, where $\rho_{\rm crit}$ is the critical density of the Universe and Δ is the specified overdensity.

Among all overlapping spheres the code finds the one that has the deepest gravitational potential. The density maximum corresponding to this sphere is treated as the center of a distinct halo. Thus, by construction, a center of a distinct halo cannot be inside the radius of another one. However, peripheral regions can still partially overlap, if the distance between centers is less than the sum of their halo radii (see Figure A.8). Radius and mass of a distinct halo depend on whether the halo overlaps or not with other distinct halos. The code takes the largest halo and identifies all other distinct halos inside a spherical shell with distances $R = (1-2)R_{\rm center}$ from the central large halo, where $R_{\rm center}$ is the radius of the large halo. For each halo selected within this shell, the code finds two radii. The first is the distance $R_{\rm big}$ to the surface of the large halo: $R_{\rm big} = R - R_{\rm center}$. The second is the distance $R_{\rm max}$ to the nearest density maximum in the shell with the inner radius $\min(R_{\rm big}, R_\Delta)$ and the outer radius $\max(R_{\rm big}, R_\Delta)$ from the center of the selected halo. If there are no density maxima within that range, then $R_{\rm max} = R_\Delta$. The radius of the selected halo is the maximum of $R_{\rm big}$ and $R_{\rm max}$. Once all halos around the large halo are processed, the next largest halo is taken from the list of distinct halos and the procedure is applied again. This setup is designed to make a smooth transition of properties of small halos when they fall into a larger halo and become subhalos.

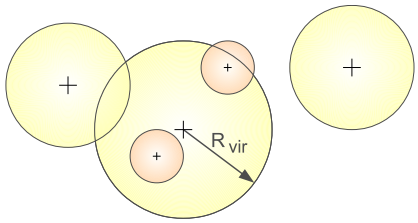

Figure A.8: Relations between BDM distinct halos and subhalos. This example has three distinct halos (yellow) and two subhalos (light red). The right-most halo is a distinct halo that does not overlap with any other distinct halo. The left-most halo is also a distinct halo, but it overlaps with the largest distinct halo at the center of the picture, and its radius is slightly reduced. Of the two subhalos, one is completely inside of its "parent", and another subhalo is partially outside of its "parent".

The bulk velocity of either a distinct halo or a subhalo is defined as the average velocity of the 100 most bound particles of that halo or by all particles, if the number of particles is less than 100. The number 100 is a compromise between the desire to use only the central (sub)halo region for the bulk velocity and the noise level.

The gravitational potential is determined by first finding the mass in spherical shells and then by integration of the mass profile. The binning is done in log radius with a very small bin size of $\Delta\log(R) = 0.01$. An algorithm based on the pair-wise summation was also tested. Just as one may expect, for relatively concentrated and not-too-aspherical halos, the difference between

spherical and direct estimates are very small for the vast majority of halos (errors are less than few percent). There are larger differences for configurations with large and dense substructure(s). In these cases the direct summation includes the potential energy of the substructure in estimating the binding energy of the whole system, which is a mistake. For this reason, a faster and more stable spherical estimator is being used.

Identification of subhalos is a more complicated procedure. Centers of subhalos can only be found among density maxima, but not all density maxima are subhalos. The code removes all "fake" subhalos: maxima of density, which do not have more than $N_{\rm filter}$ self-bound particles. These maxima are eliminated from the list of subhalo candidates. An important construct for finding subhalos are barrier points: a subhalo radius cannot be larger than the distance to the nearest barrier point times a numerical tuning factor called an overshoot factor $f_{\rm over}$, which is 1.1 for the MultiDark simulation and 1.7 for the Bolshoi simulation. The subhalo radius can be smaller than this distance. Barrier points are centers of previously identified (sub)halos. For the first subhalo, the barrier point is the center of the distinct halo. For the second subhalo, it is the first barrier point and the center of the first subhalo, and so on. The radius of a subhalo is the minimum of the distance to the nearest barrier point times $f_{\rm over}$ and the distance to its most remote bound particle.

The code starts with the density maximum and sets the first barrier point: the center of the distinct halo. Then the bulk velocity and profile of the gravitational potential of the subhalo are estimated. In the next iteration unbound particles are removed and the velocity and profiles are re-evaluated. Iterations are done until convergence is achieved or until the number of bound particles goes below $N_{\rm filter}$. Once a subhalo is found, a new barrier point is added. The procedure is repeated until all subhalo candidates have been tested.

BDM extensively uses two algorithms for rapidly finding and sorting particles. For fast search it uses two-level link-lists. The first level is a homogeneous mesh, which covers the whole volume, and its size is defined by a compromise between an optimal search radius (defined by $N_{\rm filter}$ and the overdensity limit Δ) and the available computer memory. In order to speedup the search in dense regions, a second level of the link-list is created in regions where the number of particles in the first link-list cell exceeds $15N_{\rm filter}$ particles. The cell-size of the second-level mesh is 8 times smaller than the first-level one. The code also uses a *partial* ranking algorithm for finding quantities such as the most bound particles or the particles that are closest to halo centers.

The code uses domain decomposition for MPI parallelization and OpenMP for parallelization inside each domain.

The BDM halo catalogues provide numerous parameters for each halo and subhalo: each halo is characterized with 23 parameters. In addition to coordinates and peculiar velocities, the halo finder provides two masses: the mass of all particles $M_{\rm tot}$ inside the virial radius and the mass of gravitationally bound particles $M_{\rm vir}$. Here is a list of parameters that require some explanations:

- The offset parameter $X_{\rm off}$ is defined as the distance between the center of a halo and the center of halo mass $R_{\rm cm}$. It is given in units of the halo radius: $X_{\rm off} = R_{\rm cm}/R_{\rm vir}$. This parameter is often considered as a measure of the degree of halo relaxation.

- The 3d rms velocity of particles $V_{\rm rms}$ relative to the halo center

$$V_{\rm rms}^2 = \frac{\sum_i m_i V_i^2}{\sum_i m_i}. \tag{A.1}$$

This parameter gives the kinetic energy of the halo: $E_{\rm kin} = M_{\rm bound} V_{\rm rms}^2/2$. In combination with another parameter provided by the code, the virial ratio

$$\mathrm{vir}R \equiv \frac{2E_{\rm kin}}{E_{\rm pot}} - 1, \tag{A.2}$$

one can obtain the potential energy $E_{\rm pot}$.

- The maximum circular velocity $V_{\rm circ}$ is defined using the distribution of mass $M(< R)$ inside radius R. The code bins all bound particles using very narrow spherical shells. The binning is done in constant increments of the logarithm of the radius with $\Delta \log R = 0.01$. This yields a maximum relative error in the radius of about 0.02 and even a smaller error in $V_{\rm circ}$. Then, the code searches for the maximum of circular velocity $\sqrt{GM(< R)/R}$ starting from the first bin containing at least 5 particles.

- The halo concentration C is defined by the halo mass $M_{\rm vir}$ and the maximum circular velocity $V_{\rm circ}$. The algorithm of (Prada et al., 2011) was used to find the concentration. The concentration is found by numerically solving the algebraic equation

$$\left(\frac{V_{\rm circ}}{V_{\rm vir}}\right)^2 = \frac{0.2162C}{F(C)}, \tag{A.3}$$

where $V_{\rm vir}^2 = GM_{\rm vir}/R_{\rm vir}$ and $F(C) = \ln(1 + C)/C - 1/(1 + C)$.

- The spin parameter λ is defined here as

$$\lambda \equiv \frac{JE_{\rm kin}^{1/2}}{GM_{\rm vir}^{5/2}} = \frac{jV_{\rm rms}}{\sqrt{2}GM_{\rm vir}}, \tag{A.4}$$

where J and j are respectively the total and specic angular momenta of the bound halo particles relative to the halo center. Note that the kinetic, not the total energy is used to define the spin parameter. If the total energy is wanted, it can be obtained from $V_{\rm rms}$ and $R_{\rm vir}$.

- The rms radius $R_{\rm rms}$ of bound particles:

$$R_{\rm rms}^2 = \frac{\sum_i m_i R_i^2}{\sum_i m_i}. \tag{A.5}$$

- The axis ratios and the direction of the major axis of the halo's triaxial shape. This information is obtained from diagonalization of the modified tensor of inertia $\mathcal{T}_{jk}$ for all bound particles inside the halo radius:

$$\mathcal{T}_{jk} = \sum_i \frac{x_{ij} x_{ik}}{R_i^2}, \tag{A.6}$$

where i is the particle index and j, $k = 1, 2, 3$. Here x stands for the position and r for the distance of a particles with respect to the halo's center (see also e.g. Allgood et al., 2006, equation (5)). The code does not use any corrections of the axial ratios to compensate for the fact that T_{jk} is estimated using a spherical region (Allgood et al., 2006). A correction factor for the axial ratios should be applied. However, the correction depends

on the halo concentration: it is smaller for more concentrated halos. If (c/a) and (b/a) are the small-to-large and medium-to-large axial ratios provided by the diagonalization of the modified inertia tensor, then the following corrections give the true axial ratios for halos with a flattened NFW profile:

$$\left(\frac{c}{a}\right)_{\rm cor} = \left(\frac{c}{a}\right)^s, \quad s = 1 + 2\max(q - 0.4, 0) + [5.5\max(q - 0.4, 0)]^3, \tag{A.7}$$

$$\left(\frac{b}{a}\right)_{\rm cor} = \left(\frac{b}{a}\right)^p, \quad p = 1 + 2\max(q - 0.4, 0) + [5.7\max(q - 0.4, 0)]^3 \tag{A.8}$$

$$q = \frac{R_{\rm rms}}{R_{\rm vir}}, \tag{A.9}$$

Appendix B. Friends-of-Friends halofinder (FOF)

The Friends-of-Friends (FOF) method dates back to Davis et al. (1985). This method is one of the most popular algorithms used to find objects in cosmological simulations. The great advantage of this method is its simplicity: The algorithm is based on only one free parameter - the relative linking length *ll* - which is given in terms of the mean inter-particle distance. For a given linking length the FOF algorithm uniquely defines clusters of particles that contain all particles separated by distances smaller than the linking length. In the limit of large number of particles the boundary of a cluster of particles is given by a certain isodensity surface. When the FOF algorithm was introduced into numerical cosmology, the commonly used value of the linking length was $ll = 0.2$, assuming that this value corresponds to a surface overdensity of ≈ 60, which in turn corresponds to an enclosed overdensity of 180 in an isothermal density profile, as desired for virialized halos in the standard cosmology model at that time ($\Omega_m = 1, \Lambda = 0.$). Since the modern ΛCDM model requires at $z = 0$ a higher overdensity of about 360 with respect to the mean density ($\Omega_m = 0.3$, $\Lambda = 0.7$), and since overdensities scale with (linking length)$^{-3}$, for these models a linking length of 0.17 is required at $z = 0$. However, the virial overdensity in these models changes with redshift as predicted by the spherical top-hat model (Lahav et al., 1991). It reaches again the value 180 of the Einstein-deSitter universe at high redshifts (namely when the cosmological constant is dynamically not important), thus a redshift dependent linking length would be required. Since this contradicts the idea of having only one parameter, most FOF halofinders use a fixed linking length for all redshifts. Moreover, it is well known for a long time that the overdensity of FOF objects defined with a certain linking length has a large scatter, and on average it is larger than expected. In fact, recently (More et al., 2011) have shown that for a linking length of $ll = 0.2$ of the mean interparticle distance the surface overdensity of FOF groups is equal to 81.62 times the mean density in the simulation box. Consequently, the enclosed overdensity is larger than 180. It also depends on the concentration of the objects and therefore on mass and redshift. For a linking length of $ll = 0.2$ it typically scatters between 250 and 600. For a detailed discussion of the relation between the overdensity with respect to the mean density and the linking length, see More et al. (2011). Table 3 provides a characteristic overdensity for all linking lengths available in the database.

In order to analyze the MultiDark and Bolshoi simulations with 2048^3 particles a new parallel version of the hierarchical friends-of-friends algorithm with low memory requirements was developed. The dark matter particles in the simulation are considered as an undirected graph

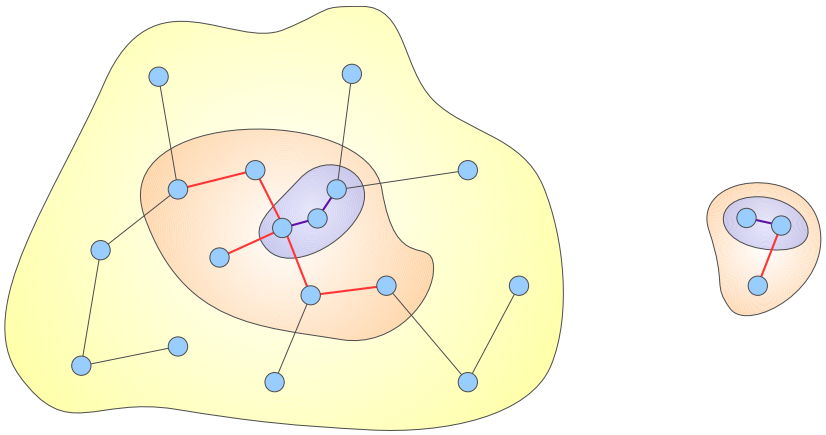

Figure B.9: FOF halos constructed with larger linking length contain halos constructed with smaller linking length.

with positive weights, namely the lengths of the segments of this graph. For simplicity, it is assumed that all weights are different. Then one can show that a unique Minimum Spanning Tree (MST) of the point distribution exists, namely the shortest graph which connects all points. If subgraphs cover the graph, then the MST of the graph belongs to the union of MSTs of the subgraphs. Thus subgraphs can be constructed in parallel. Moreover, the geometrical features of the FOF objects, namely the fact that they occupy mainly almost non-overlapping volumes, allow for the construction of fast parallel algorithms. In a second step the particles are sorted in a one-dimensional array so that each particle-cluster at any linking length is a segment of this array. This representation yields fast determination of FOF clusters at any linking length as well as the substructures of FOF clusters defined at any shorter linking length. Moreover, it is possible to determine a very fast calculation of the progenitor-descendant-relationships of FOF objects. Since at all redshifts FOF catalogues for several different linking lengths are available, one can query for the corresponding substructures. In addition, at redshift $z = 0$ large linking lengths have been used to find superclusters of FOF particles. These structure elements have been determined with a set of linking lengths representing mean overdensities down to 94 (see Table 3). For all these superclusters one has access to the raw particle data.

When the particles belonging to a given particle cluster have been determined, different properties of this FOF group can be directly calculated. The database provides the center of mass of the FOF group, its velocity, the number of particles belonging to the FOF group, the total mass, and the velocity dispersion inside the FOF halo (eq. (A.1)). Since the FOF concept is by construction aspherical, a circular velocity (as used to characterize objects found with spherical overdensity algorithms) cannot be determined here. The database provides two estimates of a "radius": (1) the coordinate dispersion (rms radius) of particles $R_{\rm rms}$ defined in the same way as in eq. (A.5), but in this case using all particles found by the FOF algorithm and (2) the radius of the sphere that has the same volume as the FOF group. This volume of the particle cluster is calculated on a grid. To save space, this volume is not included in the database. A mean overdensity of the FOF is provided, $\delta = \rho_{\rm FOF}/\rho_{\rm mean} - 1$. The database also gives the vector J of the angular momentum of each FOF group. With this vector and the total kinetic energy $E_{\rm kin}$ of the motion

of particles relative to the halo center, the spin parameter is calculated using

$$\lambda = \frac{J E_{\rm kin}^{1/2}}{G M^{5/2}} \tag{B.1}$$

Finally, the axial ratios of a FOF halo are defined as the ratios of the main axes of the tensor of inertia of FOF particles (eq. A.6). The orientation of the halo is characterized by three unit vectors pointing along these three main axes.

Appendix C. Merger trees for FOF catalogues

After finding the FOF halos in the simulation, for all the available snapshots the merging trees are determined for all halos with more than 200 particles at $z = 0$. The construction of the trees is based on the comparison of two consecutive snapshots. Starting at $z = 0$ for every FOF group in the catalog, G_0, all the FOF groups in the previous snapshots are identified that share at least 13 particle with G_0 and labeled as tentative progenitors. Then, for each tentative progenitor, all the descendants sharing at least 13 particles are determined. Only the tentative progenitors that have the group G_0 as a main descendant are labeled as confirmed progenitors at that level. This procedure is iterated for each confirmed progenitor, until the last available snapshot at high redshift is reached. By construction, each halo in the tree can have only one descendant but many progenitors.

It is also important to note, that the correspondence between FOF halos and trees is not always one to one, neither are all halos included in the trees. A possible reason is that FOF halos temporarily disappear, because of the detection threshold of 20 particles during a single snapshot. In this case, the branch is cut at the first snapshot where the halo disappears. Another possible reason is that FOF halos are sometimes linked by temporary particle bridges. In this case the algorithm detects a halo splitting when the bridge disappears, cutting the merger history of the less massive FOF clump at that time, while stitching the rest of its formation history to a tree branch of the most massive clump.

Appendix D. Database tables - Overview

Table D.4 provides an overview of tables in databases for Multidark and Bolshoi. For a complete overview of the *MultiDark Database* tables see `http://www.multidark.org/MultiDark/pages/Status.jsp`.

Table D.4: Names and description of tables in the MDR1, miniMDR1, and Bolshoi databases.

Database Table	Short description	Description
BDMV	BDM halos, $360 \cdot \rho_{back}$	Halo catalogue using the Bound Density Maximum algorithm, for all available snapshots, calculated using the standard overdensity criterion with $360 \cdot \rho_{back}$ (background density)
BDMW	BDM halos, $200 \cdot \rho_{crit}$	for all available snapshots, calculated using $200 \cdot \rho_{crit}$ (critical density) for defining the halo boundary
BDMVprof, BDMWprof	profiles for BDM halos	corresponding halo profiles for halos from tables BDMV and BDMW
FOF	FOF groups, linking length 0.17	Groups of galaxy cluster size, determined using the Friends-of-Friends analysis, for all available snapshots, level 0 (relative linking length 0.17)
FOF1 – FOF4	FOF groups, smaller linking lengths (substructures)	Friends-of-Friends catalogues for all available snapshots, same as FOF-table, but for smaller (relative) linking lengths, levels 1 (linking length 0.085) to 4 (linking length 0.010625). Thus these tables contain substructures of the FOF groups.
FOFc	FOF groups, commonly used linking length 0.2	Friends-of-Friends catalogue for all available snapshots, computed with the commonly used linking length 0.2
FOFParticles	FOF groups $\leftrightarrow$ particles	Table for connecting FOF groups from the FOF table with its particles, at the moment for redshift 0 (snapnum=85) only
FOFParticles1 – FOFParticles4	FOF groups $\leftrightarrow$ particles, for FOF1 – FOF4	Corresponding tables for connecting FOF groups from tables FOF1 – FOF4 to their particles for redshift 0
FOFMtree	merger trees for FOF	Contains identifiers to extract merger trees for galaxy clusters from FOF table
FOFSub	substructures for FOF – FOF4	Substructure tree identifiers for building a substructure tree with FOF groups from FOF – FOF4
FOFScl	FOF superclusters	Contains Friends-of-Friends catalogues for redshift 0, for 7 different (relative) linking lengths between 0.35 (sclevel 0) and 0.17 (sclevel 6), i.e. much larger objects (superclusters). It also contains identifiers for building substructure trees
particles	all particles, snapshot at $z = 0$ for both MDR1 and Bolshoi, snapshots at $z = 2.89, 1.0, 0.53$ for MDR1	All simulation particles with their positions and velocities
linkLength	linking lengths, FOF – FOF4	Overview on levels and corresponding linking lengths for FOF – FOF4 catalogues
linkLengthScl	linking lengths, FOF superclusters	Overview on sclevels for superclusters and corresponding linking lengths (for FOFScl table)
redshifts	snapshots and redshifts	Overview on available snapshots (snapnum) and corresponding redshifts

Appendix E. Diagram of the table relations in *MultiDark Database*

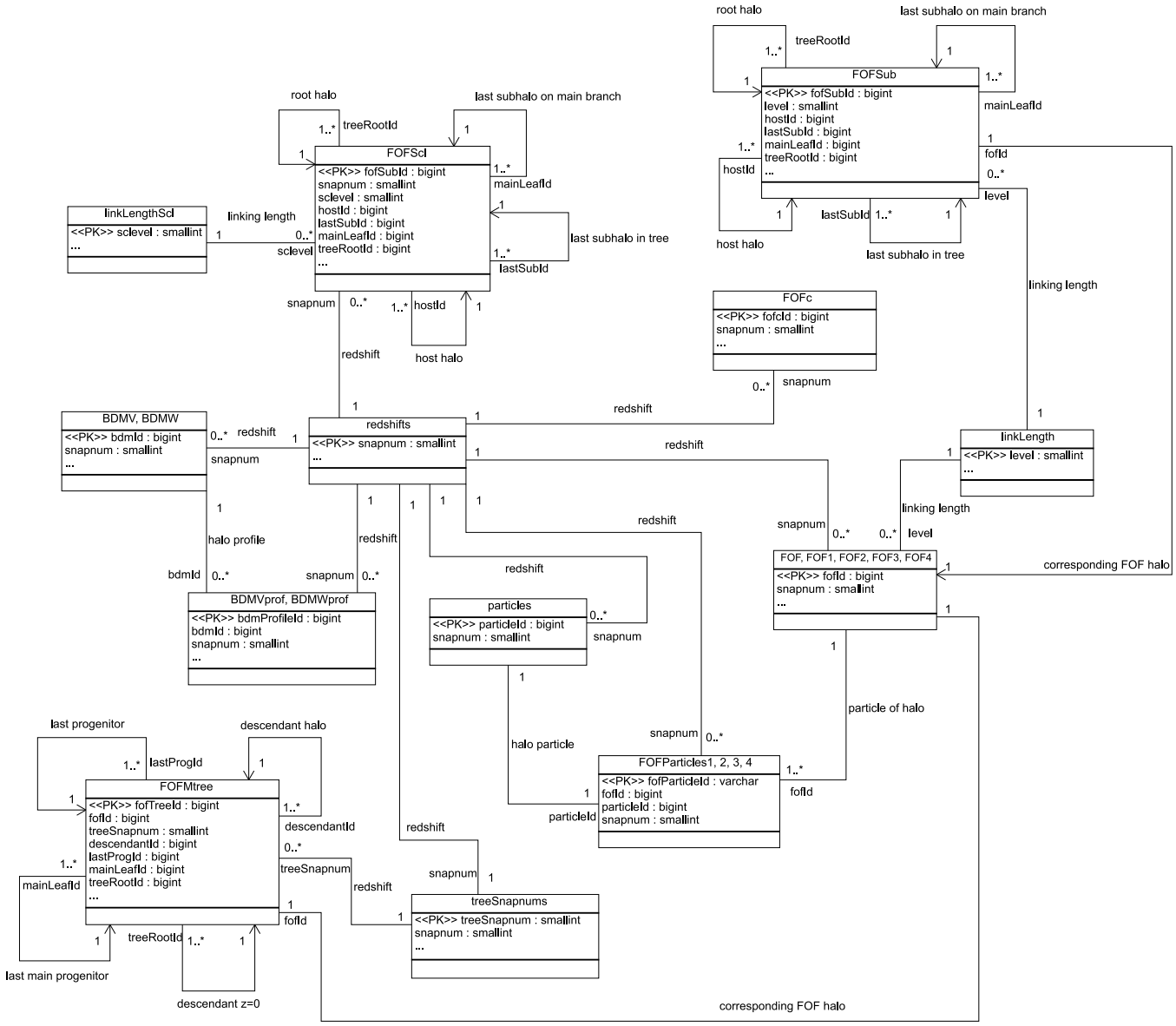

Figure E.10: Relational diagram of the available tables in the MultiDark database. The connecting lines mark relations between two tables. The connecting keys are stated along the line, together with the purpose of the relation. The numbers along the connecting lines indicate the type of the relation, for instance whether a relation is a one to one relation, one to many relation, or a many to many relation. For simplicity, only the relevant keys in the table objects are shown. For a complete UML diagram of the whole database structure, see the MultiDark webpage.

References

Aarseth, S. J., 1966. Dynamical evolution of clusters of galaxies, II. MNRAS132, 35.

Aarseth, S. J., 1969. Dynamical evolution of clusters of galaxies-III. MNRAS144, 537.

Allgood, B., Flores, R. A., Primack, J. R., Kravtsov, A. V., Wechsler, R. H., Faltenbacher, A., Bullock, J. S., Apr. 2006. The shape of dark matter haloes: dependence on mass, redshift, radius and formation. MNRAS367, 1781–1796.

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., Lacey, C. G., 2007. Breaking the hierarchy of galaxy formation. MNRAS 370, 645–655.

Boylan-Kolchin, M., Springel, V., White, S. D. M., Jenkins, A., Lemson, G., Sep. 2009. Resolving cosmic structure formation with the Millennium-II Simulation. MNRAS398, 1150–1164.

Bryan, G. L., Norman, M. L., Sep. 1998. ApJ495, 80.

Bullock, J. S., Kolatt, T. S., Sigad, Y., Somerville, R. S., Kravtsov, A. V., Klypin, A. A., Primack, J. R., Dekel, A., Mar. 2001. Profiles of dark haloes: evolution, scatter and environment. MNRAS321, 559–575.

Conroy, C., Wechsler, R. H., Kravtsov, A. V., Aug. 2006. Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time. ApJ647, 201–214.

Croton, D. J., Springel, V., White, S. D. M., De Lucia, G., Frenk, C. S., Gao, L., Jenkins, A., Kauffmann, G., Navarro, J. F., Yoshida, N., Jan. 2006. The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. MNRAS365, 11–28.

Davis, M., Efstathiou, G., Frenk, C. S., White, S. D. M., May 1985. The evolution of large-scale structure in a universe dominated by cold dark matter. ApJ292, 371–394.

De Lucia, G., Blaizot, J., Feb. 2007. The hierarchical formation of the brightest cluster galaxies. MNRAS375, 2–14.

Dubinski, J., Carlberg, R. G., Sep. 1991. The structure of cold dark matter halos. ApJ378, 496–503.

Efstathiou, G., Jones, B. J. T., Jan. 1979. The rotation of galaxies - Numerical investigations of the tidal torque theory. MNRAS186, 133.

Gao, L., Springel, V., White, S. D. M., Oct. 2005. The age dependence of halo clustering. MNRAS363, L66–L70.

GAVO, 2008. `http://www.g-vo.org/www/`.

Gott, III, J. R., Turner, E. L., Aarseth, S. J., Nov. 1979. N-body simulations of galaxy clustering. III - The covariance function. ApJ234, 13.

Gottlöber, S., Klypin, A., 2008. The ART of Cosmological Simulations High Performance Computing in Science and Engineering, Garching/Munich 2007, Transactions of the Third Joint HLRB and KONWIHR Status and Result Workshop, Eds.: Wagner, S.; Steinmetz, M.; Bode, A.;" Brehm, M., Springer-Verlag, 29–+.

Iliev, I. T., Mellema, G., Shapiro, P. R., Pen, U.-L., Mao, Y., Koda, J., Ahn, K., Jul. 2011. Can 21-cm observations discriminate between high-mass and low-mass galaxies as reionization sources? ArXiv e-prints.

Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P., Yoshida, N., Feb. 2001. The mass function of dark matter haloes. MNRAS321, 372–384.

Jing, Y. P., Aug. 1998. Accurate Fitting Formula for the Two-Point Correlation Function of Dark Matter Halos. ApJLett503, L9+.

Kauffmann, G., Colberg, J. M., Diaferio, A., White, S. D. M., Feb. 1999. Clustering of galaxies in a hierarchical universe - I. Methods and results at z=0. MNRAS303, 188–206.

Kim, J., Park, C., Gott, III, J. R., Dubinski, J., Aug. 2009. The Horizon Run N-Body Simulation: Baryon Acoustic Oscillations and Topology of Large-scale Structure of the Universe. ApJ701, 1547–1559.

Klypin, A., Holtzman, J., Dec. 1997. Particle-Mesh code for cosmological simulations. ArXiv Astrophysics e-prints.

Klypin, A., Kravtsov, A. V., Valenzuela, O., Prada, F., Sep. 1999. Where Are the Missing Galactic Satellites? ApJ522, 82–92.

Klypin, A., Trujillo-Gomez, S., Primack, J., Feb. 2010. Halos and galaxies in the standard cosmological model: results from the Bolshoi simulation. ArXiv e-prints.

Knollmann, S. R., Knebe, A., Jun. 2009. AHF: Amiga's Halo Finder. ApJS182, 608–624.

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlöber, S., Allgood, B., Primack, J. R., Jul. 2004. The Dark Side of the Halo Occupation Distribution. ApJ609, 35–49.

Kravtsov, A. V., Klypin, A. A., Aug. 1999. The Origin and Evolution of Halo Bias in Linear and Nonlinear Regimes. ApJ520, 437–453.

Kravtsov, A. V., Klypin, A. A., Khokhlov, A. M., Jul. 1997. Adaptive Refinement Tree: A New High-Resolution N-Body Code for Cosmological Simulations. ApJS111, 73–+.

Kuhlen, M., Diemand, J., Madau, P., Oct. 2008. The Dark Matter Annihilation Signal from Galactic Substructure: Predictions for GLAST. ApJ686, 262–278.

Lahav, O., Lilje, P. B., Primack, J. R., Rees, M. J., Jul. 1991. Dynamical effects of the cosmological constant. MNRAS251, 128–136.

Lemson, G., Budavari, T., Szalay, A., 2011. Implementing a General Spatial Indexing Library for Relational Databases of Large Numerical Simulations. Accepted for publication in SSDBM 2011.

Lemson, G., Springel, V., 2006. Cosmological Simulations in a Relational Database: Modelling and Storing Merger Trees. ADASS.

Lemson, G., Virgo Consortium, Aug. 2006. Halo and Galaxy Formation Histories from the Millennium Simulation: Public release of a VO-oriented and SQL-queryable database for studying the evolution of galaxies in the LambdaCDM cosmogony. ArXiv Astrophysics e-prints.

Macciò, A. V., Dutton, A. A., van den Bosch, F. C., Dec. 2008. Concentration, spin and shape of dark matter haloes as a function of the cosmological model: WMAP1, WMAP3 and WMAP5 results. MNRAS391, 1940–1954.

Moore, B., Ghigna, S., Governato, F., Lake, G., Quinn, T., Stadel, J., Tozzi, P., Oct. 1999. Dark Matter Substructure within Galactic Halos. ApJLett524, L19–L22.

More, S., Kravtsov, A., Dalal, N., Gottlöber, S., Feb. 2011. The overdensity and masses of the friends-of-friends halos and universality of the halo mass function. ArXiv e-prints.

MSDN, 2008. Transact-SQL Reference.
URL `http://msdn.microsoft.com/en-us/library/bb510741.aspx`

Navarro, J. F., Frenk, C. S., White, S. D. M., Dec. 1997. A Universal Density Profile from Hierarchical Clustering. ApJ490, 493–+.

Neto, A. F., Gao, L., Bett, P., Cole, S., Navarro, J. F., Frenk, C. S., White, S. D. M., Springel, V., Jenkins, A., Nov. 2007. The statistics of Λ CDM halo concentrations. MNRAS381, 1450–1462.

Peebles, P. J. E., Feb. 1970. Structure of the Coma Cluster of Galaxies. AJ75, 13.

Prada, F., Klypin, A., Cuesta, A., Betancort-Rijo, J., Primack, J., Apr. 2011. Halo concentrations in the standard LCDM Cosmology.

Prada, F., Klypin, A. A., Simonneau, E., Betancort-Rijo, J., Patiri, S., Gottlöber, S., Sanchez-Conde, M. A., Jul. 2006. How Far Do They Go? The Outer Structure of Galactic Dark Matter Halos. ApJ645, 1001.

Sheth, R. K., Tormen, G., Jan. 2002. An excursion set model of hierarchical clustering: ellipsoidal collapse and the moving barrier. MNRAS329, 61–75.

skyserver.sdss.org, 2008. `http://skyserver.sdss.org`.

Somerville, R. S., Gilmore, R. C., Primack, J. R., Dominguez, A., Apr. 2011. Galaxy Properties from the Ultra-violet to the Far-Infrared: Lambda-CDM models confront observations. ArXiv e-prints.

Somerville, R. S., Primack, J. R., Dec. 1999. Semi-analytic modelling of galaxy formation: the local Universe. MNRAS310, 1087–1110.

Springel, V., Dec. 2005. The cosmological simulation code GADGET-2. MNRAS364, 1105–1134.

Springel, V., Wang, J., Vogelsberger, M., Ludlow, A., Jenkins, A., Helmi, A., Navarro, J. F., Frenk, C. S., White, S. D. M., Dec. 2008. The Aquarius Project: the subhaloes of galactic haloes. MNRAS391, 1685–1711.

Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J., Pearce, F., Jun. 2005. Simulating the joint evolution of quasars, galaxies and their large-scale distribution. Nature 435, 629–636.

Stadel, J., Potter, D., Moore, B., Diemand, J., Madau, P., Zemp, M., Kuhlen, M., Quilis, V., Sep. 2009. Quantifying the heart of darkness with GHALO - a multibillion particle simulation of a galactic halo. MNRAS398, L21–L25.

Teyssier, R., Pires, S., Prunet, S., Aubert, D., Pichon, C., Amara, A., Benabed, K., Colombi, S., Refregier, A., Starck, J.-L., Apr. 2009. Full-sky weak-lensing simulation with 70 billion particles. A&A497, 335–341.

Tinker, J., Kravtsov, A. V., Klypin, A., Abazajian, K., Warren, M., Yepes, G., Gottlöber, S., Holz, D. E., Dec. 2008. Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality. ApJ688, 709–728.

Tinker, J. L., Robertson, B. E., Kravtsov, A. V., Klypin, A., Warren, M. S., Yepes, G., Gottlöber, S., Dec. 2010. The Large-scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests. ApJ724, 878–886.

Trujillo-Gomez, S., Klypin, A., Primack, J., Romanowsky, A., Mar. 2010. LCDM Correctly Predicts Basic Statistics of Galaxies: Luminosity-Velocity Relation, Baryonic Mass-Velocity Relation, and Velocity Function. ArXiv e-prints.

Vale, A., Ostriker, J. P., Sep. 2004. Linking halo mass to galaxy luminosity. MNRAS353, 189–200.

van den Bosch, F. C., Yang, X., Mo, H. J., Weinmann, S. M., Macciò, A. V., More, S., Cacciato, M., Skibba, R., Kang, X., Apr. 2007. Towards a concordant model of halo occupation statistics. MNRAS376, 841–860.

Warren, M. S., Abazajian, K., Holz, D. E., Teodoro, L., Aug. 2006. Precision Determination of the Mass Function of Dark Matter Halos. ApJ646, 881–885.

Wechsler, R. H., Zentner, A. R., Bullock, J. S., Kravtsov, A. V., Allgood, B., Nov. 2006. The Dependence of Halo Clustering on Halo Formation History, Concentration, and Occupation. ApJ652, 71–84.

Wetzel, A. R., White, M., 2010. What determines satellite galaxy disruption? MNRAS 403, 1072–1088.

White, S. D. M., Dec. 1976. The dynamics of rich clusters of galaxies. MNRAS177, 717.

Zentner, A. R., Berlind, A. A., Bullock, J. S., Kravtsov, A. V., Wechsler, R. H., May 2005. The Physics of Galaxy Clustering. I. A Model for Subhalo Populations. ApJ624, 505–525.

Zhao, D. H., Mo, H. J., Jing, Y. P., Börner, G., Feb. 2003. The growth and structure of dark matter haloes. MNRAS339, 12–24.