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 ΛCDM cosmogony

Gerard Lemson[*†] and the Virgo Consortium

31 July 2006

ABSTRACT The Millennium Run is the largest simulation of the formation of structure within the ΛCDM cosmogony so far carried out. It uses 10^{10} particles to follow the dark matter distribution in a cubic region $500h^{-1}$Mpc on a side, and has a spatial resolution of 5 h^{-1}kpc. Application of simplified modelling techniques to the stored output of this calculation allows the formation and evolution of the $\sim 10^7$ galaxies more luminous than the Small Magellanic Cloud to be simulated for a variety of assumptions about the detailed physics involved. As part of the activities of the German Astrophysical Virtual Observatory we have used a relational database to store the detailed assembly histories both of all the haloes and subhaloes resolved by the simulation, and of all the galaxies that form within these structures for two independent models of the galaxy formation physics. We have created web applications that allow users to query these databases remotely using the standard Structured Query Language (SQL). This allows easy access to all properties of the galaxies and halos, as well as to the spatial and temporal relations between them and their environment. Information is output in table format compatible with standard Virtual Observatory tools and protocols. With this announcement we are making these structures fully accessible to all users. Interested scientists can learn SQL, gain familiarity with the database design and test queries on a small, openly accessible version of the Millennium Run (with volume 1/512 that of the full simulation). They can then request accounts to run similar queries on the databases for the full simulations.

[*]German Astrophysical Virtual Observatory, Max Planck Institutes for Astrophysics and for extraterrestrial Physics
[†]*email:* gerard.lemson@mpe.mpg.de

1 The Millennium Run

The last few years have seen the establishment of a standard model for the origin and growth of structure in the Universe, the so-called ΛCDM cosmogony. In this model small density fluctuations are generated during an early period of cosmic inflation and first become directly observable on the last scattering surface of the Cosmic Microwave Background when the Universe was about 400,000 years old. Since this time the fluctuations have grown steadily through the gravitational effects of a dominant Dark Matter component composed of some weakly interacting particle yet to be detected directly on Earth. As fluctuations become nonlinear, larger and larger objects collapse, giving rise to the galaxies and galaxy clusters we see today. This process has recently been modified as Dark Energy has come to dominate the cosmic energy density, accelerating the cosmic expansion and reducing the rate of structure growth. A major effort is currently underway, testing this paradigm and measuring its parameters. A parallel effort explores galaxy and cluster formation in this model in order to understand the physical processes which shaped observed systems.

The Millennium Run, completed in summer 2004 at the Max Planck Society's supercomputer centre in Garching, is part of the programme of the Virgo Consortium[1] and is intended as a tool to facilitate this second effort. It uses 10^{10} particles of mass $8.6 \times 10^8 h^{-1} M_\odot$ to follow the evolution of the dark matter distribution within a cubic region of side $500 h^{-1}$ Mpc from $z = 127$ until $z = 0$. The cosmological parameters assumed are $\Omega_m = \Omega_{dm} + \Omega_b = 0.25$, $\Omega_b = 0.045$, $\Omega_\Lambda = 0.75$, $h = 0.73$, $\sigma_8 = 0.9$ and $n = 1$ with standard definitions for all quantities. The initial density fluctuations correctly account for the oscillatory features introduced by the baryons, but the simulation follows the dark matter only, supplementing the mass of the simulation particles to account approximately for the neglected baryons.

The simulation was carried out using a modified version of the publicly available code GADGET-2 (Springel 2005). The positions and velocities of all simulation particles were stored at 63 times spaced approximately logarithmically from $z = 20$ to the present day. For each of these dumps the algorithm SUBFIND (Springel et al. 2001) was used to identify all self-bound halos containing at least 20 particles and all self-bound subhalos within these halos down to the same mass limit. Merger trees were then built linking each halo and its substructures at the final time to the objects at earlier times from which they formed. These trees are the input to the final stage of post-processing. This simulates the formation of the galaxies in all or a part of the volume by following simplified treatments of the baryonic physics within each tree, starting at early times and integrating down to $z = 0$. More detailed descriptions of the simulation itself and of this post-processing can be found in Springel et al. (2005).

Several different galaxy formation models have already been implemented on this structure by the Garching and Durham groups. The model used in Springel et al. (2005) to present some initial clustering and evolution results is essentially identical to that described and explored in considerably more detail by Croton et al. (2006c). The model used by De Lucia et al. (2006) to study elliptical galaxy evolution is similar in most aspects, but differs in its treatment of feedback from star formation. In

---

[1]http://www.virgo.dur.ac.uk

their study of brightest cluster galaxies De Lucia & Blaizot (2006) use a model with a different assumed IMF for star formation, an improved scheme for tracking halo central galaxies, but the same feedback scheme as Croton et al. (2006c). The model independently developed in Durham and presented in Bower et al. (2006) differs from these Garching models in many ways. The scheme for building merger trees from the halo/subhalo data is different in detail, as are many of the modelling assumptions made to deal with the baryonic physics, most notably, perhaps, those associated with the growth of and the feedback from supermassive black holes in galaxy nuclei. In this public release we are initially making available the galaxy populations produced by the models of the De Lucia & Blaizot (2006) and Bower et al. (2006) papers.

The data on the halo/subhalo and galaxy populations which have been produced by this effort can be used to address a very wide range of questions about galaxy and structure evolution in the now standard model. In the 13 months since *Nature* published the first Millennium Run paper (Springel et al. 2005) a further 24 papers have appeared on the preprint server using data derived from the simulation. Some of these are concerned with issues of dark matter structure (Gao et al. 2005; Harker et al. 2006; Gao & White 2006). Others build and test galaxy formation models, exploring the requirements for reproducing various aspects of the observed properties of galaxies and AGN (Croton et al. 2006c,b; Bower et al. 2006; De Lucia et al. 2006; Croton 2006; Wang et al. 2006; De Lucia & Blaizot 2006). Yet others concentrate on aspects of large-scale structure and galaxy clustering (Croton et al. 2006a; Noh & Lee 2006; Lee & Park 2006) and on cluster structure and gravitational lensing (Hayashi & White 2006; Moeller et al. 2006; Weinmann et al. 2006; Natarajan et al. 2006). A number of authors have used galaxy catalogues from the Millennium Run as a point of comparison in primarily observational papers (Kauffmann et al. 2006; Patiri et al. 2006; Einasto et al. 2006; Rudnick et al. 2006; Bernardi et al. 2006; Conroy et al. 2006; Li et al. 2006). Finally, the data have been used to illustrate the current state-of-the-art in a review of the field of large-scale structure (Springel et al. 2006). The goal of this release is to facilitate further such use of Millennium Run data products by making them conveniently and publicly available over the Web.

2 The Databases

The stored raw data from the Millennium Run, the positions and velocities of all 10^{10} particles in the initial conditions and at each of the 63 later output times, have a total volume of almost 20TB. This is so large that general public access and/or manipulation over the internet is not currently a viable possibility. As a result, projects which require access to the full particle data (e.g. ray-tracing projects for gravitational lensing applications) are only practicable in collaboration with Virgo scientists in Garching or Durham, where copies of the full data are stored. The Virgo Consortium welcomes suggestions for such joint projects and will try to accommodate them as far as they overlap with the interests of scientists at one of the Virgo institutions and do not conflict with existing projects.

Many projects, however, including the great majority of those listed at the end of §1, can be carried out using products from our Millennium Run post-processing

pipeline. Only $O(130\text{GB})$ are needed to store the information provided by our halo-subhalo analysis including the tree structure which describes the assembly history of all objects. A database with the corresponding galaxy information from one of our galaxy formation simulations is roughly twice as large because of the larger number of objects and the larger number of attributes ascribed to each. The variety of these attributes and the complexity of the relations between them motivate the use of relational databases, whose query engines allow complex questions to be phrased in the standard Structured Query Language (SQL) and executed in optimal fashion. These tools have become standard in the Virtual Observatory community as a means for promoting efficient and user-friendly data-mining within large observational databases. A major task for the German Astrophysical Virtual Observatory (GAVO) has been to adapt these tools for large theoretical (simulation) databases, where issues of format and quality control are less difficult than for observational archives, but where relationships can be considerably more complex, primarily because of the addition of the time dimension.

The relational database structure and the online query interface which we have set up for Millennium Run data products are modelled closely on the relevant parts of the very successful SkyServer system set up for the Sloan Digital Sky Survey.[2] Within such a database information is stored in the form of tables where rows correspond to individual objects and columns to attributes of those objects (e.g. position, velocity, mass, angular momentum, size, flattening, type , luminosity, colour, indices specifying relations to other objects...). The web interface allows users to formulate their scientific questions as SQL queries operating on these tables in a relatively simple way and to submit them remotely over the internet for execution on the database server (located at present in Garching, with a mirror to be set up in the near future in Durham). The subsequent search of the database is optimised as far as possible for "typical" queries. Results are returned to the user over the internet as tables in one of a number of formats which can then be fed into standard VO or other graphics packages or can be further manipulated by users with their own software.

The entry point for our public release of Millennium products is
http://www.mpa-garching.mpg.de/Millennium
This top page gives a brief introduction to the Millennium Run as well as links to images and movies, to papers which have used the data, and to the pages on the GAVO site which describe in detail the database structure, the SQL and the procedures for accessing and downloading data. Data on the so-called "milli-Millennium" run (hereafter milli-M) can be accessed from these pages immediately. This is a simulation which is identical to the main run in all aspects except that it is carried out in a cubic region of side $62.5h^{-1}\text{Mpc}$. It thus has 1/512 of the physical volume and its databases are 512 times smaller than those of the Millennium Run itself. Any query can be executed on the milli-M databases directly from this page, provided it executes on the host computer in less than 30 seconds. The first 10,000 lines of any output are returned to the user. These restrictions are intended to avoid inadvertently tying up the host or the internet link while developing SQL expertise.

Once users can execute their queries efficiently on the milli-M databases, they

---

can apply for password-protected accounts as specified on the web-page in order to carry out the corresponding queries on the main database. This two-stage system is intended to allow monitoring of usage patterns and to prevent accidental abuse by inexperienced users.

At present databases are accessible for the (sub)halo data and for galaxy data from the models of De Lucia & Blaizot (2006) and Bower et al. (2006). identical models and attribute lists are used for the milli-M and full Millennium versions of these databases. Further galaxy models and further data products may be added as they are generated. Examples of scientific queries that can be formulated simply and executed efficiently with the present SQL engine include:

- Find all halos (or galaxies) in a given part of the simulation at a given time and in a given mass range

- Find all companion halos (or galaxies) in some given range of separations from this previous set of objects

- Find the number of galaxies at a given time in each of a series of narrow luminosity bins (e.g. the luminosity function)

- Find the number of galaxies in high mass halos at a given time in such luminosity bins (e.g. the cluster luminosity function)

- Find all resolved progenitor halos at redshift 3 of high-mass $z = 0$ halos

- Find all galaxies at redshift 3 which are progenitors of the central galaxies of high-mass $z = 0$ halos

- Find the halo masses of all $10^{11} M_\odot$ galaxies at $z = 3$ and determine the fraction which are central galaxies of these halos

- Find the $z = 0$ descendents of all redshift 3 galaxies with stellar mass above $10^{11} M_\odot$ or with star formation rate above $10 M_\odot/\text{yr}$

- Find all halos (or galaxies) which have undergone a major merger since the previous stored output time

Clearly this capability allows a very broad range of scientific issues to be addressed in a straightforward way. Some of these are currently being studied by scientists associated with the Virgo Consortium, but we hope that this release will encourage others to use the exceptional statistics provided by the Millennium Run to explore how galactic and dark matter structures evolve in the current ΛCDM paradigm. In particular, closer comparison with a wide range of observational data should indicate how our present simple models for the formation of galaxies and AGN need to be modified to correspond better with reality.

Acknowledgements

References

Bernardi, M., Hyde, J. B., Sheth, R. K., Miller, C. J., Nichol, R. C. (2006), *The luminosities, sizes and velocity dispersions of Brightest Cluster Galaxies: Implications for formation history*, ArXiv Astrophysics e-prints

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., Lacey, C. G. (2006), *Breaking the hierarchy of galaxy formation*, Mon. Not. R. Astron. Soc., 659–+

Conroy, C., Prada, F., Newman, J. A., Croton, D., Coil, A. L., Conselice, C. J., Cooper, M. C., Davis, M., Faber, S. M., Gerke, B. F., Guhathakurta, P., Klypin, A., Koo, D. C., Yan, R. (2006), *Evolution in the Halo Masses of Isolated Galaxies between redshifts 1 and 0: From DEEP2 to SDSS*, ArXiv Astrophysics e-prints

Croton, D. J. (2006), *Evolution in the black hole mass-bulge mass relation: a theoretical perspective*, Mon. Not. R. Astron. Soc., 369, 1808

Croton, D. J., Gao, L., White, S. D. M. (2006a), *Halo assembly bias and its effects on galaxy clustering*, ArXiv Astrophysics e-prints

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. (2006b), *Erratum: The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies*, Mon. Not. R. Astron. Soc., 367, 864

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. (2006c), *The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies*, Mon. Not. R. Astron. Soc., 365, 11

De Lucia, G., Blaizot, J. (2006), *The hierarchical formation of the brightest cluster galaxies*, ArXiv Astrophysics e-prints

De Lucia, G., Springel, V., White, S. D. M., Croton, D., Kauffmann, G. (2006), *The formation history of elliptical galaxies*, Mon. Not. R. Astron. Soc., 366, 499

Einasto, J., Einasto, M., Saar, E., Tago, E., Liivamagi, L. J., Joeveer, M., Suhhonenko, I., Hutsi, G., Jaaniste, J., Heinamaki, P., Muller, V., Knebe, A., Tucker, D. (2006), *Superclusters of galaxies from the 2dF redshift survey. II. Comparison with simulations*, ArXiv Astrophysics e-prints

Gao, L., Springel, V., White, S. D. M. (2005), *The age dependence of halo clustering*, Mon. Not. R. Astron. Soc., 363, L66

Gao, l., White, S. D. M. (2006), *Asymmetries in the inner regions of LCDM haloes*, ArXiv Astrophysics e-prints

Harker, G., Cole, S., Helly, J., Frenk, C., Jenkins, A. (2006), *A marked correlation function analysis of halo formation times in the Millennium Simulation*, Mon. Not. R. Astron. Soc., 367, 1039

Hayashi, E., White, S. D. M. (2006), *How rare is the bullet cluster?*, Mon. Not. R. Astron. Soc., 370, L38

Kauffmann, G., Heckman, T. M., De Lucia, G., Brinchmann, J., Charlot, S., Tremonti, C., White, S. D. M., Brinkmann, J. (2006), *Gas infall and stochastic star formation in galaxies in the local universe*, Mon. Not. R. Astron. Soc., 367, 1394

Lee, J., Park, D. (2006), *Rotation of Cosmic Voids and Void-Spin Statistics*, ArXiv Astrophysics e-prints

Li, C., Kauffmann, G., Wang, L., White, S. D. M., Heckman, T. M., Jing, Y. P. (2006), *The clustering of narrow-line AGN in the local Universe*, ArXiv Astrophysics e-prints

Moeller, O., Kitzbichler, M., Natarajan, P. (2006), *Strong lensing statistics in large, low redshift surveys: bias in the lens galaxy population*, ArXiv Astrophysics e-prints

Natarajan, P., De Lucia, G., Springel, V. (2006), *Substructure in lensing clusters and simulations*, ArXiv Astrophysics e-prints

Noh, Y., Lee, J. (2006), *The Anisotropy in the Galaxy Velocity Field Originated from the Gravitational Pancaking Effect*, ArXiv Astrophysics e-prints

Patiri, S. G., Prada, F., Holtzman, J., Klypin, A., Betancort-Rijo, J. (2006), *The Properties of Galaxies in Voids*, ArXiv Astrophysics e-prints

Rudnick, G., Labbe, I., Foerster Schreiber, N. M., Wuyts, S., Franx, M., Finlator, K., Kriek, M., Moorwood, A., Rix, H.-W., Roettgering, H., Trujillo, I., van der Wel, A., van der Werf, P., van Dokkum, P. G. (2006), *Measuring the Average Evolution of Luminous Galaxies at z below 3: The Rest-frame Optical Luminosity Density, Spectral Energy Distribution, and Stellar Mass Density*, ArXiv Astrophysics e-prints

Springel, V. (2005), *The cosmological simulation code GADGET-2*, Mon. Not. R. Astron. Soc., 364, 1105

Springel, V., Frenk, C. S., White, S. D. M. (2006), *The large-scale structure of the Universe*, Nature, 440, 1137

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. (2005), *Simulations of the formation, evolution and clustering of galaxies and quasars*, Nature, 435, 629

Springel, V., White, S. D. M., Tormen, G., Kauffmann, G. (2001), *Populating a cluster of galaxies - I. Results at z=0*, Mon. Not. R. Astron. Soc., 328, 726

Wang, L., Li, C., Kauffmann, G., De Lucia, G. (2006), *Modelling galaxy clustering in a high resolution simulation of structure formation*, ArXiv Astrophysics e-prints

Weinmann, S. M., van den Bosch, F. C., Yang, X., Mo, H. J., Croton, D. J., Moore, B. (2006), *Properties of Galaxy Groups in the SDSS: II.- AGN Feedback and Star Formation Truncation*, ArXiv Astrophysics e-prints