Cosmology II: Inhomogeneities in the Universe

7

7.1 Introduction

In Chap. 4, we discussed homogeneous world models and introduced the standard model of cosmology. It is based on the cosmological principle, the assumption of a (spatially) homogeneous and isotropic universe. Of course, the assumption of homogeneity is justified only on large scales because observations show us that our Universe is inhomogeneous on small scales—otherwise no galaxies or stars would exist.

The distribution of galaxies on the sky is not uniform or random (see Fig. 6.1), rather they form clusters and groups of galaxies. Also clusters of galaxies are not distributed uniformly, but their positions are correlated, grouped together in superclusters. The three-dimensional distribution of galaxies, obtained from redshift surveys, shows an interesting large-scale structure, as can be seen in Fig. 7.1 which shows the spatial distribution of galaxies in the two-degree-Field Galaxy Redshift Survey (2dFGRS).

Even larger structures have been discovered. The Great Wall is a galaxy structure with an extent of $\sim 100h^{-1}$ Mpc, which was found in a redshift survey of galaxies (Fig. 7.2). The largest structure discovered up to now is the Sloan Great Wall, also shown in Fig. 7.2. Such surveys also led to the discovery of the so-called *voids*, nearly spherical regions which contain virtually no (bright) galaxies, and which have a diameter of typically $30h^{-1}$ Mpc. The discovery of these large-scale inhomogeneities raises the question of whether

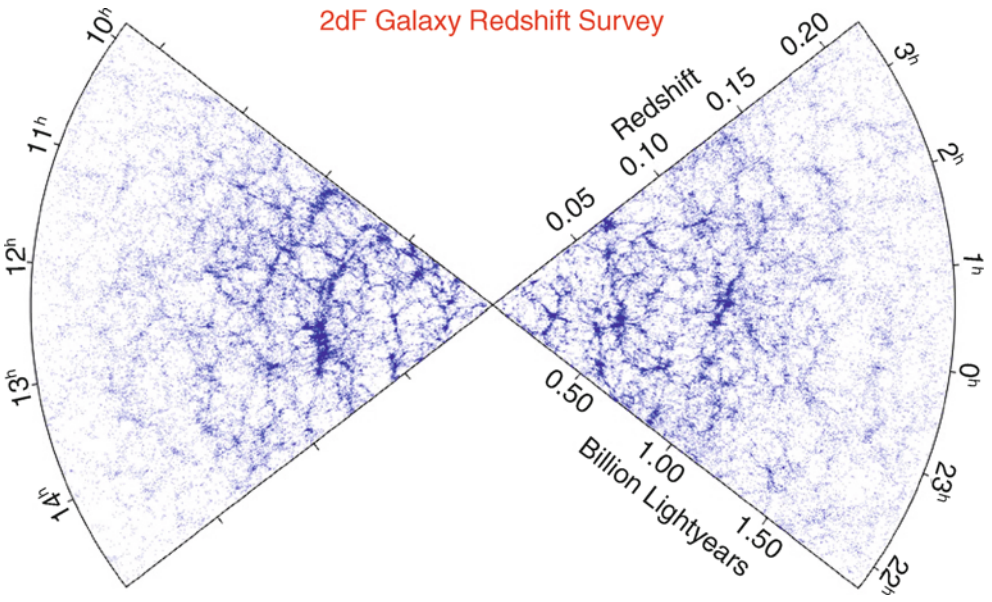

Fig. 7.1 The distribution of galaxies in the complete 2dF Galaxy Redshift Survey. In the radial direction, the escape velocity, or redshift, is plotted, and the polar angle is the right ascension. According the Hubble law, the redshift is directly related to the distance of an object, so that redshift surveys map the three-dimensional distribution of galaxies, with our Galaxy at the center of the figure. In the 2dFGRS, more than 350 000 spectra were taken between 1997 and 2002; plotted here is the distribution of more than 200 000 galaxies with reliable redshift measurements. The data from the complete survey are publicly available. Credit: M. Colless and the 2dF Galaxy Redshift Survey team

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_7,

even larger structures might exist in the Universe, or more precisely: does a scale exist, averaged over which the Universe appears homogeneous? The existence of such a scale is a requirement for the homogeneous world models to provide a realistic description of the mean cosmic behavior.

To date, no evidence of structures with linear dimension well above $100h^{-1}$ Mpc have been found, as can also be seen from Fig. 7.1. Hence, the Universe seems to be basically homogeneous if averaged over scales of $R \sim 200h^{-1}$ Mpc. This 'homogeneity scale' needs to be compared to the Hubble radius $R_{\rm H} \equiv c/H_0 \approx 3000h^{-1}$ Mpc. This implies $R \ll R_{\rm H}$, so that after averaging, $(R_{\rm H}/R)^3 \sim (15)^3 \sim 3000$ independent volume elements exist per Hubble volume. This justifies the approximation of a homogeneous world model when considering the mean cosmic history.

On small scales, the Universe is inhomogeneous. Evidence for this is the galaxy distribution projected on the sky, the three-dimensional galaxy distribution determined by redshift surveys, and the existence of clusters of galaxies, superclusters, 'Great Walls', and voids. In addition, the anisotropy of the cosmic microwave background (CMB), with relative fluctuations of $\Delta T/T \sim 10^{-5}$, indicates that the Universe already contained small inhomogeneities at redshift $z \sim 1000$, which we will discuss more thoroughly in Sect. 8.6. In this chapter, we will examine the evolution of such density inhomogeneities and their description.

7.2 Gravitational instability

7.2.1 Overview

The smallness of the CMB anisotropy suggests that the density inhomogeneities at redshift $z \sim 1000$—this is the epoch where most of the CMB photons interacted with matter for the last time—must have had very small amplitudes. Today, the amplitudes of the density inhomogeneities are considerably larger; for example, a massive cluster of galaxies contains within a radius of $\sim 1.5h^{-1}$ Mpc more than 200 times more mass than an average sphere of this radius in the Universe. Thus, these are no longer small density fluctuations.

Obviously, the Universe became more inhomogeneous in the course of its evolution; as we will see, density perturbations grow over time. One defines the *relative density contrast*

$$\boxed{\delta(\boldsymbol{r},t) := \frac{\rho(\boldsymbol{r},t)-\bar{\rho}(t)}{\bar{\rho}(t)}}\,, \tag{7.1}$$

where $\bar{\rho}(t)$ denotes the mean cosmic matter density at time t. From the definition of δ, one can immediately see that $\delta \geq -1$, because $\rho \geq 0$. The smallness of the CMB anisotropy suggests that at $z \sim 1000$, $|\delta| \ll 1$. The dynamics of the cosmic Hubble expansion is controlled by the gravitational field of the average matter density $\bar{\rho}(t)$, whereas the density fluctuations $\Delta\rho(\boldsymbol{r},t) = \rho(\boldsymbol{r},t) - \bar{\rho}(t)$ generate an additional gravitational field.

We shall here be interested only in very weak gravitational fields, for which the Newtonian description of gravity can be applied. Since the Poisson equation, which specifies the relation between matter density and the gravitational potential, is linear, the effects of the homogeneous matter distribution and of density fluctuations can be considered separately. The gravitational field of the total matter distribution is then the sum of the average matter distribution and that of the density fluctuations.

We consider a region in which $\Delta\rho > 0$, hence $\delta > 0$, so that the gravitational field in this region is stronger than the cosmic average. An overdense region produces a stronger gravitational field than that corresponding to the mean Hubble expansion. By this additional self-gravity, the overdense region will expand more slowly than the average Hubble expansion. Because of the delayed expansion, the density in this region will also decrease more slowly than in the cosmic mean, $\bar{\rho}(t) = (1+z)^3\rho_0 = a^{-3}(t)\rho_0$, and hence the density contrast in this region will increase. As a consequence, the relative density will increase, which again produces an even stronger gravitational field.... It is obvious that this situation is unstable. Of course, the argument also works the other way round: in an underdense region with $\delta < 0$, the gravitational field generated is weaker than in the cosmic mean, therefore the self-gravity is weaker than that which corresponds to the Hubble expansion. This implies that the expansion is decelerated less than in the cosmic mean, the underdense region expands faster than the Hubble expansion, and thus the local density will decrease more quickly than the mean density of the Universe. In this way, the absolute value of the density contrast increases, i.e., δ becomes more negative over the course of time.

> Density fluctuations grow over time due to their self-gravity; overdense regions increase their density contrast over the course of time, while underdense regions decrease their density contrast. In both cases, $|\delta|$ increases. Hence, this effect of *gravitational instability* leads to an increase of density fluctuations with increasing time. The evolution of structure in the Universe is described by this effect of *gravitational instability*.

Structure growth in the Universe can be understood in the framework of this model. In this chapter we will describe structure formation quantitatively. This includes the analysis of the time evolution of density perturbations, as well as a statistical description of such density fluctuations. We will

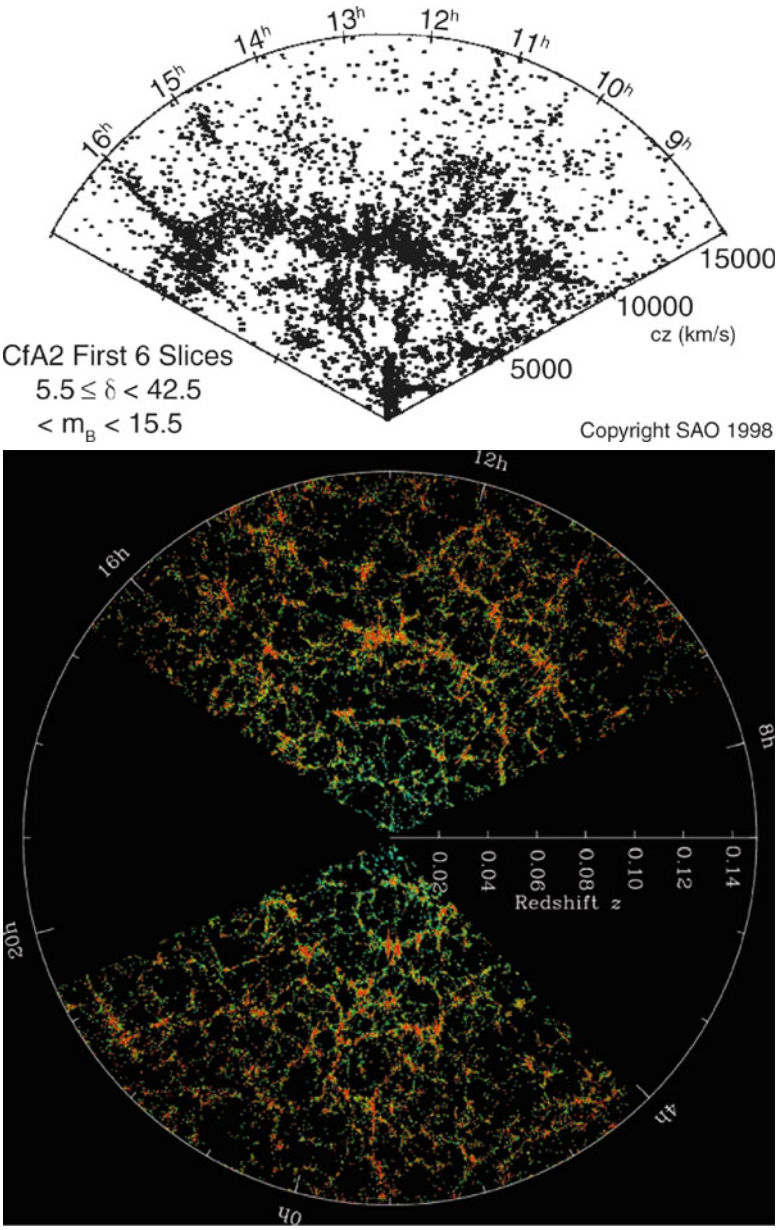

Fig. 7.2 *Top:* In the CfA galaxy redshift survey, carried out in the 1980s, a large coherent structure of galaxies was found, called the Great Wall. Shown are galaxies with radial velocities of $cz \leq 15\,000\,\mathrm{km/s}$, with declination $8.5^\circ \leq \delta \leq 42^\circ$. The Great Wall is located at a redshift of $cz \sim 6000\,\mathrm{km/s}$, extending in right ascension between $9^\mathrm{h} \leq \alpha \leq 16^\mathrm{h}$. *Bottom:* The distribution of galaxies as measured from the Sloan Digital Sky Survey. Plotted are galaxies in the narrow declination range $-1.25^\circ \leq \delta \leq 1.25^\circ$. Note that this distribution extends to considerably larger distances than the one in the CfA survey. The *color of the points* indicates the color of the galaxies. The most remarkable feature seen is the long filament of galaxies near the center of the upper part of the figure, called the Sloan Great Wall. Credit: *Top*: J. Huchra, M. Geller, Harvard-Smithsonian Center for Astrophysics. *Bottom*: M. Blanton and the Sloan Digital Sky Survey

then see that the evolution of inhomogeneities is directly observable, and that the Universe was less inhomogeneous at high redshift than it is today. Since the history of perturbations depends on the cosmological model, we need to examine whether this evolution can be used to obtain an estimate of cosmological parameters. In Chap. 8, we will give an affirmative answer to this question. Finally, we will briefly discuss the origin of density fluctuations.

7.2.2 Linear perturbation theory

We first will examine the growth of density perturbations. For this discussion, we will concentrate on length-scales that are substantially smaller than the Hubble radius. On these scales, structure growth can be described in the framework of the Newtonian theory of gravity. The effects of space-time curvature and thus of General Relativity need to be accounted for only for density perturbations on length-scales comparable to, or larger than the Hubble radius. In addition, we assume for simplicity that the matter in the Universe consists only of dust (i.e., pressure-free matter), with density $\rho(\boldsymbol{r}, t)$. The matter distribution will be described in the *fluid approximation*, where the velocity field of this fluid shall be denoted by $\boldsymbol{v}(\boldsymbol{r}, t)$.[1]

[1] Strictly speaking, the cosmic dust cannot be described as a fluid because the matter is assumed to be collisionless. This means that no interactions occur between the particles, except for gravitation. Two

Equations of motion. The behavior of this fluid is described by the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0 \,, \tag{7.2}$$

which expresses the fact that matter is conserved: the density decreases if the fluid has a diverging velocity field (thus, if particles are moving away from each other). In contrast, a converging velocity field will lead to an increase in density. Furthermore, the Euler equation applies,

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v} \cdot \nabla)\, \boldsymbol{v} = -\frac{\nabla P}{\rho} - \nabla \Phi \,, \tag{7.3}$$

which describes the conservation of momentum and the behavior of the fluid under the influence of forces. The left-hand side of (7.3) is the time derivative of the velocity as would be measured by an observer moving with the flow, because $\partial \boldsymbol{v}/\partial t$ is the derivative at a fixed point in space, whereas the total left-hand side of (7.3) is the time derivative of the velocity measured along the flow lines. The latter is affected by the gravitational field Φ and the pressure gradient. However, since we are only considering pressureless matter, the pressure vanishes, $P \equiv 0$. The gravitational potential satisfies the Poisson equation

$$\nabla^2 \Phi = 4\pi G \rho - \Lambda \,. \tag{7.4}$$

which has been modified compared to the usual Poisson equation to account for the presence of a cosmological constant Λ. We will see in a short while why this form of the Λ-term is chosen.

These three equations for the description of a self-gravitating fluid can in general not be solved analytically. However, we will show that a special, cosmologically relevant exact solution can be found, and that by linearization of the system of equations around this exact solution approximate solutions can be constructed for small relative density contrasts, $|\delta| \ll 1$.

Exact solution: The Hubble expansion. The special exact solution is the flow that we have already encountered in Chap. 4: the homogeneous expanding cosmos. By substituting into the above equations it is immediately shown that

$$\boldsymbol{v}(\boldsymbol{r}, t) = H(t)\boldsymbol{r}$$

is a solution of the equations if ρ is homogeneous and satisfies (4.11), and if the Friedmann equation (4.19) for the scale factor applies (see problem 7.1). Note that the particular form of the Λ-term in the Poisson equation was chosen because with this modification of the 'standard' Poisson equation we can obtain the second Friedmann equation from a Newtonian treatment.

As long as the density contrast $|\delta| \ll 1$, the deviations of the velocity field from the Hubble expansion will be small. We expect that in this case, physically relevant solutions of the above equations are those which deviate only slightly from the homogeneous case.

It is convenient to consider the problem in comoving coordinates; hence we define, as in (4.4),

$$\boldsymbol{r} = a(t)\, \boldsymbol{x} \,.$$

In a homogeneous cosmos, $\boldsymbol{x}$ is a constant for every matter particle, and its spatial position $\boldsymbol{r}$ changes only due to the Hubble expansion. Likewise, the velocity field is written in the form

$$\boldsymbol{v}(\boldsymbol{r}, t) = \frac{\dot{a}}{a}\boldsymbol{r} + \boldsymbol{u}\left(\frac{\boldsymbol{r}}{a}, t\right) \,, \tag{7.5}$$

where $\boldsymbol{u}(\boldsymbol{x}, t)$ is a function of the comoving coordinate $\boldsymbol{x}$. In (7.5), the first term represents the homogeneous Hubble expansion, whereas the second term describes the deviations from this homogeneous expansion. For this reason, $\boldsymbol{u}$ is called the *peculiar velocity*. Next, we will show how the above equations read in comoving coordinates.

Transforming the fluid equations to comoving coordinates. We first note that the partial derivative $\partial/\partial t$ in (7.2) means a time derivative at fixed $\boldsymbol{r}$. If the equations are to be written in comoving coordinates, this partial time derivative needs to be transformed into one where $\boldsymbol{x}$ is kept fixed. For example,

$$\begin{aligned}\left(\frac{\partial}{\partial t}\right)_r \rho(\boldsymbol{r}, t) &= \left(\frac{\partial}{\partial t}\right)_r \rho_x\left(\frac{\boldsymbol{r}}{a}, t\right) \\ &= \left(\frac{\partial}{\partial t}\right)_x \rho_x(\boldsymbol{x}, t) - \frac{\dot{a}}{a}\, \boldsymbol{x} \cdot \nabla_x \rho_x(\boldsymbol{x}, t) \,,\end{aligned} \tag{7.6}$$

where ∇_x is the gradient with respect to comoving coordinates, and where we define the function $\rho_x(\boldsymbol{x}, t) \equiv \rho(a\boldsymbol{x}, t)$. Note that $\rho_x(\boldsymbol{x}, t)$ and $\rho(\boldsymbol{x}, t)$ both describe the same *physical* density field, but that ρ and ρ_x are different *mathematical* functions of their arguments. After these transformations, (7.2) becomes

$$\frac{\partial \rho}{\partial t} + \frac{3\dot{a}}{a}\rho + \frac{1}{a}\nabla \cdot (\rho \boldsymbol{u}) = 0 \,, \tag{7.7}$$

where from now on all spatial derivatives are to be considered with respect to $\boldsymbol{x}$. For notational simplicity, from now on we also set $\rho \equiv \rho_x$ and $\delta \equiv \delta(\boldsymbol{x}, t)$, and note that the partial time derivative is to be understood to be at fixed $\boldsymbol{x}$. Writing $\rho = \bar{\rho}(1 + \delta)$ and using $\bar{\rho} \propto a^{-3}$, (7.7) reads in comoving coordinates

flows of such dust can thus penetrate each other. This situation can be compared to that of a fluid whose molecules are interacting by collisions. Through these collisions, the velocity distribution of the molecules will, at each position, assume an approximate Maxwell distribution, with a well-defined average velocity that corresponds to the flow velocity at this point. Such an unambiguous velocity does not exist for dust in general. However, at early times, when deviations from the Hubble flow are still very small, no multiple flows are expected, so that in this case, the velocity field is well defined.

$$\frac{\partial\delta}{\partial t} + \frac{1}{a}\nabla\cdot[(1+\delta)\,\boldsymbol{u}] = 0\,. \tag{7.8}$$

Accordingly, the gravitational potential Φ is written as

$$\Phi(\boldsymbol{r},t) = \left(\frac{2\pi}{3}G\bar{\rho}(t) - \frac{\Lambda}{6}\right)|\boldsymbol{r}|^2 + \phi(\boldsymbol{x},t) = \frac{\ddot{a}a}{2}\,|\boldsymbol{x}|^2 + \phi(\boldsymbol{x},t)\,; \tag{7.9}$$

the first term is the Newtonian potential for a homogeneous density field, and ϕ satisfies the Poisson equation for the density inhomogeneities,

$$\begin{aligned}\nabla^2\phi(\boldsymbol{x},t) &= 4\pi G a^2(t)\bar{\rho}(t)\delta(\boldsymbol{x},t)\\ &= \frac{3H_0^2\Omega_{\rm m}}{2a(t)}\,\delta(\boldsymbol{x},t)\,,\end{aligned} \tag{7.10}$$

where in the last step we used $\bar{\rho}\propto a^{-3}$ and the definition of the density parameter $\Omega_{\rm m}$. Then, the Euler equation (7.3) becomes

$$\frac{\partial\boldsymbol{u}}{\partial t} + \frac{\boldsymbol{u}\cdot\nabla}{a}\boldsymbol{u} + \frac{\dot{a}}{a}\boldsymbol{u} = -\frac{1}{\bar{\rho}\,a}\nabla P - \frac{1}{a}\nabla\phi\,, \tag{7.11}$$

where (4.13) has been utilized.

Linearization. In the homogeneous case, $\delta\equiv 0$, $\boldsymbol{u}\equiv 0$, $\phi\equiv 0$, $\rho=\bar{\rho}$, and (7.7) then implies $\dot{\bar{\rho}}+3H\bar{\rho}=0$, which also follows immediately from (4.17) in the case of $P=0$. Now we will look for approximate solutions of the above set of equations which describe only small deviations from this homogeneous solution. For this reason, in these equations we only consider first-order terms in the small parameters δ and $\boldsymbol{u}$, i.e., we disregard terms that contain $\boldsymbol{u}\,\delta$ or are quadratic in the velocity $\boldsymbol{u}$, i.e., the second term on the l.h.s. of (7.11). The linearized continuity and Euler equations then read

$$\frac{\partial\delta}{\partial t} + \frac{1}{a}\nabla\cdot\boldsymbol{u} = 0\,, \tag{7.12}$$

$$\frac{\partial\boldsymbol{u}}{\partial t} + \frac{\dot{a}}{a}\boldsymbol{u} = -\frac{1}{a}\nabla\phi\,, \tag{7.13}$$

where we set $P=0$, and the Poisson equation (7.10) is linear. Combining this set of equations, we can eliminate the peculiar velocity $\boldsymbol{u}$ and the gravitational potential ϕ from the equations[2] and then obtain a second-order differential equation for the density contrast δ,

$$\frac{\partial^2\delta}{\partial t^2} + \frac{2\dot{a}}{a}\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta\,. \tag{7.14}$$

It is remarkable that neither does this equation contain derivatives with respect to spatial coordinates, nor do the coefficients in the equation depend on $\boldsymbol{x}$. Therefore, (7.14) has solutions of the form

$$\delta(\boldsymbol{x},t) = D(t)\,\tilde{\delta}(\boldsymbol{x})\,,$$

i.e., the spatial and temporal dependencies factorize in these solutions. Here, $\tilde{\delta}(\boldsymbol{x})$ is an arbitrary function of the spatial coordinate, and $D(t)$ satisfies the equation

$$\ddot{D} + \frac{2\dot{a}}{a}\dot{D} - 4\pi G\bar{\rho}(t)\,D = 0\,. \tag{7.15}$$

The growth factor. The differential equation (7.15) has two linearly independent solutions. One can show that one of them increases with time, whereas the other decreases (we will see a special example of this below). If, at some early time, both functional dependencies were present, the increasing solution will dominate at later times, whereas the solution decreasing with t will become irrelevant. Therefore, we will consider only the increasing solution, which is denoted by $D_+(t)$, and normalize it such that $D_+(t_0)=1$. Then, the density contrast becomes

$$\boxed{\delta(\boldsymbol{x},t) = D_+(t)\,\delta_0(\boldsymbol{x})}\,. \tag{7.16}$$

This mathematical consideration allows us to draw immediately a number of conclusions. First, the solution (7.16) implies that in linear perturbation theory *the spatial shape of the density fluctuations is frozen in comoving coordinates*, only their amplitude increases. The *growth factor* $D_+(t)$ of the amplitude follows a simple differential equation that is readily solvable for any cosmological model. In fact, one can show (see problem 7.2) that for arbitrary values of the density parameters in matter and vacuum energy, the growth factor has the form

$$D_+(a) \propto \frac{H(a)}{H_0}\int_0^a \frac{{\rm d}a'}{\left[\Omega_{\rm m}/a' + \Omega_\Lambda a'^2 - (\Omega_{\rm m}+\Omega_\Lambda-1)\right]^{3/2}}\,, \tag{7.17}$$

where the factor of proportionality is determined from the condition $D_+(t_0)=1$.

In accordance with $D_+(t_0)=1$, $\delta_0(\boldsymbol{x})$ would be the distribution of density fluctuations today if the evolution was indeed linear until the present epoch. Therefore, $\delta_0(\boldsymbol{x})$ is denoted as the *linearly extrapolated density fluctuation field*. However, the linear approximation breaks down if $|\delta|$ is no longer $\ll 1$. In this case, the terms that have been neglected in the above derivations are no longer small and have to be included. The problem then becomes *considerably* more difficult and defies analytical treatment. Instead one needs, in general, to rely on numerical procedures for analyzing the growth of density perturbations. Furthermore, it shall be

[2] This is done by first taking the divergence of (7.13), $\nabla\cdot\dot{\boldsymbol{u}} = -(\dot{a}/a)\nabla\cdot\boldsymbol{u} - (1/a)\nabla^2\phi = \dot{a}\delta - (1/a)\nabla^2\phi$, where we made use of (7.12) in the second step. Taking the time derivative of (7.12) yields $\ddot{\delta} = (\dot{a}/a^2)\nabla\cdot\boldsymbol{u} - (1/a)\nabla\cdot\dot{\boldsymbol{u}} = -2(\dot{a}/a)\dot{\delta} + (1/a^2)\nabla^2\phi$. Finally, the Poisson equation is employed.

noted once again that, for large density perturbations, the fluid approximation is no longer valid, and that up to now we have assumed that the pressure can be neglected. At early times, i.e., for $z \gtrsim z_{\rm eq}$ [see (4.58)], this assumption becomes invalid, so that the above equations need to be modified for these early epochs.

Example: Einstein–de Sitter model. In the special case of a world model with $\Omega_{\rm m} = 1$, $\Omega_\Lambda = 0$, (7.15) can be solved explicitly. In this case, $a(t) = (t/t_0)^{2/3}$, so that

$$\left(\frac{\dot a}{a}\right) = \frac{2}{3t}\ , \text{ and } \bar\rho(t) = a^{-3}\rho_{\rm cr} = \frac{3H_0^2}{8\pi G}\left(\frac{t}{t_0}\right)^{-2}\ ;$$

furthermore, in this model $t_0\, H_0 = 2/3$, so that (7.15) reduces to

$$\ddot D + \frac{4}{3t}\dot D - \frac{2}{3t^2} D = 0\ . \tag{7.18}$$

This equation is easily solved by making the ansatz $D \propto t^q$; this ansatz is suggested because (7.18) is equidimensional in t, i.e., each term has the dimension $D/(\text{time})^2$. Inserting into (7.18) yields a quadratic equation for q,

$$q(q-1) + \frac{4}{3}q - \frac{2}{3} = 0\ ,$$

with solutions $q = 2/3$ and $q = -1$. The latter corresponds to fluctuations decreasing with time and will be disregarded in the following. So, for the Einstein–de Sitter model, the increasing solution

$$D_+(t) = \left(\frac{t}{t_0}\right)^{2/3} = a(t) \tag{7.19}$$

is found, i.e., in this case the growth factor equals the scale factor. For different cosmological parameters this is not the case, but the qualitative behavior is quite similar, which is demonstrated in Fig. 7.3 for three models. In particular, fluctuations were able to grow by a factor ~ 1000 from the epoch of recombination at $z \sim 1000$, from which the CMB photons originate, to the present day.

Evidence for dark matter on cosmic scales. At the present epoch, $\delta \gg 1$ certainly on scales of clusters of galaxies (~ 2 Mpc), and $\delta \sim 1$ on scales of superclusters (~ 10 Mpc). Hence, because of the law of linear structure growth (7.16) and the behavior of $D_+(t)$ shown in Fig. 7.3, we would expect $\delta \gtrsim 10^{-3}$ at $z = 1000$ for these structures to be able to grow to non-linear structures at the current epoch. For this reason, we should also expect CMB fluctuations to be of comparable magnitude, $\Delta T/T \gtrsim 10^{-3}$. The observed

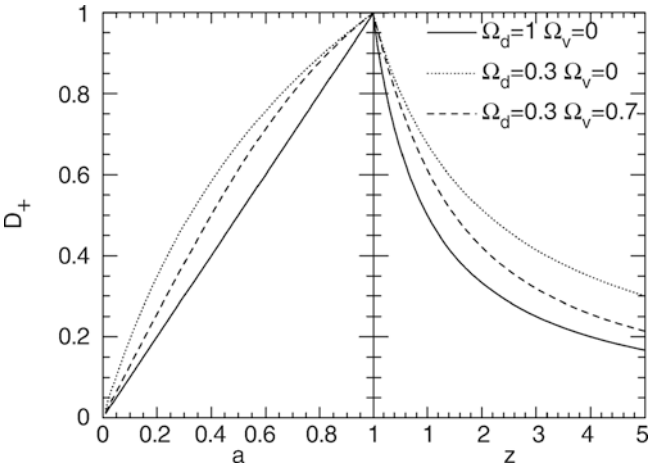

Fig. 7.3 Growth factor D_+ for three different cosmological models, as a function of the scale factor a (*left panel*) and of redshift (*right panel*). It is clearly visible how quickly D_+ decreases with increasing redshift in the EdS-model, in comparison to the models of lower density

fluctuation amplitude is much smaller, $\Delta T/T \sim 10^{-5}$, however. The corresponding density fluctuations therefore cannot have grown sufficiently strongly up to today to form non-linear structures.

This contradiction can be resolved by the dominance of dark matter. Since photons interact with baryonic matter only, the CMB anisotropies basically provide (at least on angular scales below $\sim 1°$) information on the density contrast of *baryons*. Dark matter may have had a higher density contrast at recombination, but the baryons, which are strongly coupled to the radiation field before recombination, are prevented from strong clustering due to the radiation pressure. Only after recombination, when the electrons have combined with the atomic nuclei and essentially no free electrons remain, the coupling to the radiation field ends, after which the baryons may fall into the potential wells formed by the dark matter. We will return to this issue in Sect. 7.4.3.

7.2.3 Peculiar velocities

As mentioned on several occasions before, cosmic sources do not exactly follow the Hubble expansion, but have an additional peculiar velocity. Deviations from the Hubble flow are caused by local gravitational fields, and such fields are in turn generated by local density fluctuations. These inevitably lead to an acceleration, which affects the matter and generates peculiar velocities. Indeed, we see from (7.12) that a time-dependent density contrast δ implies the presence of peculiar velocities. In numerical simulations, the peculiar velocities of individual particles are followed in the computations automatically. In this brief section, we will investigate

the large-scale peculiar velocities as they are derived from linear perturbation theory.

Since the spatial dependence of the density contrast δ is constant in time, $\delta(\boldsymbol{x},t) = \delta_0(\boldsymbol{x})\,D_+(t)$ [see (7.16)], the acceleration vector $\boldsymbol{g}$ has a constant direction in the framework of linear perturbation theory. Hence, one obtains the peculiar velocity in the form

$$\boldsymbol{u}(\boldsymbol{x}) \propto \int \mathrm{d}t\ \boldsymbol{g}(\boldsymbol{x},t)\ ,$$

i.e., parallel to $\boldsymbol{g}(\boldsymbol{x})$. On the other hand, $\boldsymbol{g}(\boldsymbol{x})$ is the gradient of the gravitational potential, $\boldsymbol{g} \propto -\nabla\phi$. This implies that $\boldsymbol{u}(\boldsymbol{x})$ is a gradient field, i.e., a scalar function $\psi(\boldsymbol{x})$ exists such that $\boldsymbol{u} = \nabla\psi$, where the gradient is taken with respect to the comoving spatial coordinate $\boldsymbol{x}$. Therefore, $\nabla\cdot\boldsymbol{g} \propto -\nabla^2\phi \propto -\delta$, where the Poisson equation (7.10) has been utilized. Thus we conclude that $\nabla\cdot\boldsymbol{u} \propto -\delta$, i.e., the divergence of the peculiar velocity field can be expressed in terms of the density contrast. In the following, we will obtain this result in a more quantitative way.

We consider the growing mode of density perturbations, for which (7.16) implies that

$$\frac{\partial\delta}{\partial t} = \frac{\dot{D}_+}{D_+}\,\delta\ .$$

Inserting this result into the linearized continuity equation (7.12) yields

$$\begin{aligned}\nabla\cdot\boldsymbol{u} &= -a\,\frac{\dot{D}_+}{D_+}\,\delta = -a\,\dot{a}\,\frac{1}{D_+}\,\frac{\mathrm{d}D_+}{\mathrm{d}a}\,\delta \\ &= -a\,H(a)\,f(a)\,\delta\ ,\end{aligned} \tag{7.20}$$

where we replaced the time derivative of D_+ by a derivative with respect to the scale factor. In the last step, we defined the function

$$f(a) := \frac{a}{D_+}\,\frac{\mathrm{d}D_+}{\mathrm{d}a} = \frac{\mathrm{d}\log D_+}{\mathrm{d}\log a}\ . \tag{7.21}$$

The function f can be calculated explicitly from (7.17). It turns out that the resulting expression can be very accurately approximated by

$$f(a) \approx \Omega_\mathrm{m}^\gamma(a)\ , \tag{7.22}$$

where $\Omega_\mathrm{m}(a)$ is the redshift-dependent matter density parameter,

$$\Omega_\mathrm{m}(a) = \frac{\rho_\mathrm{m}(a)}{\rho_\mathrm{cr}(a)} = \Omega_{\mathrm{m},0}\left[a^3\left(\frac{H}{H_0}\right)^2\right]^{-1}\ . \tag{7.23}$$

Here we used the expression for the redshift-dependent critical density given in (4.80), and we explicitly wrote $\Omega_{\mathrm{m},0}$ for the current-epoch value of the density parameter, usually simply called Ω_m. The parameter γ is in the range of 0.55–0.6 for a large variety of cosmological parameters. Therefore, one usually approximates $f(a) \approx \Omega_\mathrm{m}^{0.6}(a)$. Introducing, as before, the velocity potential ψ by $\boldsymbol{u} = \nabla\psi$, we obtain from (7.20)

$$\nabla^2\psi \approx -a\,H(a)\,\Omega_\mathrm{m}^{0.6}(a)\,\delta\ . \tag{7.24}$$

This Poisson equation for ψ can be solved, and by computing the gradient of the solution, the peculiar velocity field can be calculated,

$$\boldsymbol{u}(\boldsymbol{x},t) = \frac{\Omega_\mathrm{m}^{0.6}(a)}{4\pi}\,a\,H(a)\int \mathrm{d}^3y\ \delta(\boldsymbol{y},t)\,\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^3}\ . \tag{7.25}$$

This equation shows that the velocity field can be derived from the density field. If the density field in the Universe was observable, one would obtain a direct prediction for the corresponding velocity field from the above relations. This depends on the matter density Ω_m, so that from a comparison with the observed velocity field, one could estimate the value for Ω_m. We will come back to this in Sect. 8.1.8.

7.3 Description of density fluctuations

We will now examine the question of how to describe an inhomogeneous universe quantitatively, i.e., how to quantify the structures it contains. This task sounds easier at first sight than it is in reality. One has to realize that the aim of such a theoretical description cannot be to describe the complete function $\delta(\boldsymbol{x},t)$ for a particular universe. No cosmological model will be able to describe, for instance, the matter distribution in the vicinity of the Milky Way in detail. No model based on the laws of physics alone will be able to predict that at a distance of $\sim$ 800 kpc from the Galaxy a second massive spiral galaxy is located, because this specific feature of our local Universe depends on the specific initial conditions of the matter distribution in the early Universe. We can at best hope to predict the *statistical properties* of the mass distribution, such as, for example, the average number density of clusters of galaxies above a given mass, or the probability of a massive galaxy being found within 800 kpc of another one. Likewise, cosmological simulations (see below) cannot predict *our* Universe; instead, they are at best able to generate cosmological mass distributions that have the same statistical properties as that in our Universe.

It is quite obvious that a very large number of statistical properties exist for the density field, all of which we can

examine and which we hope can be explained quantitatively by the correct model of cosmological structure formation. To make any progress at all, the statistical properties need to be sorted or classified. How can the statistical properties of a density field best be described?

Two universes are considered equivalent if their density fields δ have the same statistical properties. One may then imagine considering a large (statistical) ensemble of universes whose density fields all have the same statistical properties, but for which the individual functions $\delta(\boldsymbol{x})$ can all be different. This statistical ensemble is called a *random field*, and any individual distribution with the respective statistical properties is called a *realization of the random field*.

An example may clarify these concepts. We consider the waves on the surface of a large lake. The statistical properties of these waves—such as how many of them there are with a certain wavelength, and how their amplitudes are distributed—depend on the shape of the lake, its depth, and the strength and direction of the wind blowing over its surface. If we assume that the wind properties are not changing with time, the statistical properties of the water surface are constant over time. Of course, this does not mean that the amplitude of the surface height as a function of position is constant. Rather, it means that two photographs of the surface that are taken at different times are statistically indistinguishable: the distribution of the wave amplitudes will be the same, and there is no way of deciding which of the snapshots was taken first. Knowing the surface topography and the wind properties sufficiently well, one is able to compute the distribution of the wave amplitudes, but there is no way to predict the amplitude of the surface of the lake as a function of position at a particular time. Each snapshot of the lake is a realization of the random field, which in turn is characterized by the statistical properties of the waves.

7.3.1 Correlation functions

Galaxies are not randomly distributed in space, but rather they gather in groups, clusters, or even larger structures. Phrased differently, this means that the probability of finding a galaxy at location $\boldsymbol{x}$ is not independent of whether there is a galaxy at a neighboring point $\boldsymbol{y}$. It is more probable to find a galaxy in the vicinity of another one than at an arbitrary location.[3] This phenomenon is described such that one considers two points $\boldsymbol{x}$ and $\boldsymbol{y}$, and two volume elements $\mathrm{d}V$ around these points. If $\bar{n}$ is the average number density of galaxies, the probability of finding a galaxy in the volume element $\mathrm{d}V$ around $\boldsymbol{x}$ is then

$$P_1 = \bar{n}\,\mathrm{d}V\ ,$$

independent of $\boldsymbol{x}$ if we assume that the Universe is statistically homogeneous. We choose $\mathrm{d}V$ such that $P_1 \ll 1$, so that the probability of finding two or more galaxies in this volume element is negligible.

The probability of finding a galaxy in the volume element $\mathrm{d}V$ at location $\boldsymbol{x}$ and at the same time finding a galaxy in the volume element $\mathrm{d}V$ at location $\boldsymbol{y}$ is then

$$P_2 = (\bar{n}\,\mathrm{d}V)^2\left[1 + \xi_{\mathrm{g}}(\boldsymbol{x},\boldsymbol{y})\right]\ . \tag{7.26}$$

If the distribution of galaxies was uncorrelated, the probability P_2 would simply be the product of the probabilities of finding a galaxy at each of the locations $\boldsymbol{x}$ and $\boldsymbol{y}$ in a volume element $\mathrm{d}V$, so $P_2 = P_1^2$. But since the distribution is correlated, the relation does not apply in this simple form; rather, it needs to be modified, as was done in (7.26). Equation (7.26) defines the *two-point correlation function* (or simply 'correlation function') of galaxies $\xi_{\mathrm{g}}(\boldsymbol{x},\boldsymbol{y})$.

By analogy to this, the correlation function for the total matter density can be defined as

$$\begin{aligned}\langle\rho(\boldsymbol{x})\,\rho(\boldsymbol{y})\rangle &= \bar{\rho}^2\,\langle[1+\delta(\boldsymbol{x})]\,[1+\delta(\boldsymbol{y})]\rangle \\ &= \bar{\rho}^2\,(1+\langle\delta(\boldsymbol{x})\,\delta(\boldsymbol{y})\rangle) \\ &=: \bar{\rho}^2\,[1+\xi(\boldsymbol{x},\boldsymbol{y})]\ ,\end{aligned} \tag{7.27}$$

because the mean (or expectation) value $\langle\delta(\boldsymbol{x})\rangle = 0$ for all locations $\boldsymbol{x}$, as can be seen from the definition (7.1).

In the above equations, angular brackets denote averaging over an ensemble of distributions that all have identical statistical properties. In our example of the lake, the correlation function of the wave amplitudes at positions $\boldsymbol{x}$ and $\boldsymbol{y}$, for instance, would be determined by taking a large number of snapshots of its surface and then averaging the product of the amplitudes at these two locations over all these realizations.

Since the Universe is considered statistically homogeneous, ξ can only depend on the difference $\boldsymbol{x}-\boldsymbol{y}$ and not on $\boldsymbol{x}$ and $\boldsymbol{y}$ individually. Furthermore, ξ can only depend on the separation $r = |\boldsymbol{x}-\boldsymbol{y}|$, and not on the direction of the separation vector $\boldsymbol{x}-\boldsymbol{y}$ because of the assumed statistical isotropy of the Universe. Therefore, $\xi = \xi(r)$ is simply a function of the separation between two points.

For a homogeneous random field, the ensemble average can be replaced by spatial averaging, i.e., the correlation function can be determined by averaging over the products of densities at pairs of points, for a large number of pairs of points with given separation r. For determining the corre-

[3] An every-day life example of clustering is the following: the population density of many European countries is of order 100 people per km^2, and so the mean separation between two people is on the order of 100 m. For those of you who live in a town or a city, the first morning view from the window shows that typically, you find many people within that distance range, an experience strengthened once you get into your car or use public transport. Obviously, people are highly clustered.

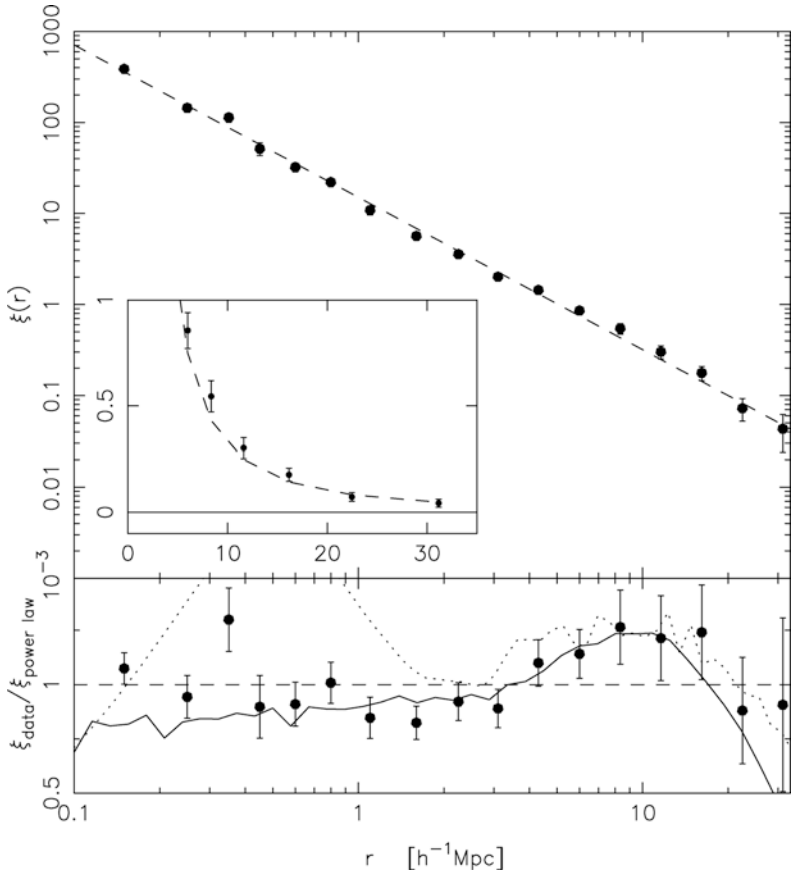

Fig. 7.4 The correlation function ξ_g of galaxies, as it was determined from the 2dF Galaxy Redshift Survey shown in Fig. 7.1. The *top panel* shows $\xi_g(r)$ as obtained from the survey (*points*), together with the best-fitting power law, $\xi_g(r) = (r/r_0)^{-\gamma}$, with correlation length $r_0 = 5.05h^{-1}$ Mpc and slope $\gamma = 1.67$. The *bottom panel* shows the ratio between the measured data points and the power-law fit (*points*), and the corresponding ratio for an earlier result obtained from a deprojection of the angular correlation function of a photometric survey (*solid curve*) and the correlation function obtained from the Hubble Volume simulation, an N-body simulation (see Fig. 7.12). Source: E. Hawkins et al. 2003, *The 2dF Galaxy Redshift Survey: correlation functions, peculiar velocities and the matter density of the Universe*, MNRAS 346, 78, p. 86, Fig. 11. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

lation function of galaxies, we note that $\xi_g(r)$ is the excess probability to find a galaxy at a separation r from another galaxy, relative to that of a random distribution. Therefore, $\xi_g(r)$ can be determined by first counting the number of galaxy pairs with separation in the interval Δr around r. Then one creates a random distribution of the same number of objects in the same volume, and again counts the pairs in the same distance interval. The ratio of these two pair counts then yields an estimate for $\xi_g(r)$.[4]

The equivalence of ensemble average and spatial average is called the ergodicity of the random field. Only by this can the correlation function (and all other statistical properties) in our Universe be measured at all, because we are able to observe only a single—namely our—realization of the hypothetical ensemble. From the measured correlations between galaxy positions, as determined from spectroscopic redshift surveys of galaxies (see Sect. 8.1.2), one finds the approximate relation

$$\xi_g(r) = \left(\frac{r}{r_0}\right)^{-\gamma} \tag{7.28}$$

for galaxies of luminosity $\sim L^*$ (see Fig. 7.4), where $r_0 \simeq 5h^{-1}$ Mpc denotes the correlation length, and where the slope is about $\gamma \simeq 1.7$. This relation is approximately valid over a range of separations $0.2h^{-1}$ Mpc $\lesssim r \lesssim 30h^{-1}$ Mpc.

Hence, the correlation function provides a means to characterize the structure of the cosmological matter distribution. Besides this two-point correlation function, correlations of higher order may also be defined, leading to general n-point correlation functions. These are more difficult to determine from observation, though. It can be shown that the statistical properties of a random field are fully specified by the set of all n-point correlations.

[4] This method is indeed used for estimating the galaxy correlation function, although with some important modifications to increase its accuracy and efficiency.

7.3.2 The power spectrum

An alternative (and equivalent) description of the statistical properties of a random field, and thus of the matter distribution in a universe, is the *power spectrum* $P(k)$. Roughly speaking, the power spectrum $P(k)$ describes the level of structure as a function of the length-scale $L \simeq 2\pi/k$; the larger $P(k)$, the larger the amplitude of the fluctuations on a length-scale $2\pi/k$. Here, k is a comoving *wave number*. Phrased differently, the density fluctuations are decomposed into a sum of plane waves of the form $\delta(\boldsymbol{x}) = \sum a_{\boldsymbol{k}} \cos(\boldsymbol{x} \cdot \boldsymbol{k})$, with a wave vector $\boldsymbol{k}$ and an amplitude $a_{\boldsymbol{k}}$. The power spectrum $P(k)$ then describes the mean of the squares, $|a_{\boldsymbol{k}}|^2$, of the amplitudes, averaged over all wave vectors with equal length $k = |\boldsymbol{k}|$. Technically speaking, this is a Fourier decomposition. Referring back to the example of waves on the surface of a lake, one finds that a characteristic wavelength L_c exists, which depends, among other factors, on the wind speed. In this case, the power spectrum will have a prominent maximum at $k = 2\pi/L_c$.

The power spectrum $P(k)$ and the correlation function are related through a Fourier transform; formally, one has[5]

$$P(k) = 2\pi \int_0^\infty \mathrm{d}x \; x^2 \, \frac{\sin(kx)}{kx} \, \xi(x) \; , \tag{7.29}$$

i.e., the integral over the correlation function with a weight factor depending on $k \sim 2\pi/L$. This relation can also be inverted, and thus $\xi(x)$ can be computed from $P(k)$.

In cosmology, one uses both concepts, correlation function and power spectrum, side-by-side. One of the two may be easier to determine from observational data, another one may be more straightforward to obtain from a model or from simulations. Also, our intuitive understanding of these two concepts may vary in different situations. Probably, for most non-cosmologists the concept of a correlation function is easier to grasp than that of a power spectrum. There are situations, however, where it is the opposite: The fluctuations of the pressure of air at a fixed point as a function of time, $\Delta P(t)$, is a function with a non-vanishing correlation $\xi(\tau) = \langle \Delta P(t) \, \Delta P(t+\tau) \rangle$, although we might have difficulties understanding the significance of this function. However, the corresponding power spectrum $P(\omega)$ is something well known, namely the frequency spectrum of sound.

Gaussian random fields. In general, knowing the power spectrum is not sufficient to unambiguously describe the statistical properties of any random field—in the same way as the correlation function $\xi(x)$ only provides an incomplete characterization. However, for certain classes of random fields, this is different. In particular, there are so-called *Gaussian random fields*, which are uniquely characterized by $P(k)$. Among the properties which characterize them, the probability distribution of the density fluctuations $\delta(\boldsymbol{x})$ at any point is a Gaussian. Such Gaussian random fields play an important role in cosmology because it is assumed that at very early epochs, the density field obeyed Gaussian statistics. This is a prediction of a large class of models of inflation which are supposed to generate the primordial density fluctuations in the Universe (see Sect. 7.9 below). Observational evidence for the Gaussian nature of the early density fluctuations comes from the observation of the anisotropy of the cosmic microwave background (see Sect. 8.6) which very strongly constrain any possible deviation from a Gaussian random field in the early Universe.

[5] This may not look like a 'standard' Fourier transform on first sight. However, the relation between $P(k)$ and $\xi(r)$ is given by a three-dimensional Fourier transform. Since the correlation function depends only on the separation $r = |\boldsymbol{r}|$, the two integrals over the angular coordinates can be performed explicitly, leading to the form of (7.29).

7.4 Evolution of density fluctuations

$P(k)$ and $\xi(x)$ both depend on cosmological time or redshift because the density field in the Universe evolves over time. Therefore, the dependence on t is explicitly written as $P(k,t)$ and $\xi(r,t)$. Note that $P(k,t)$ is linearly related to $\xi(x,t)$, according to (7.29), and ξ in turn depends quadratically on the density contrast δ. If $\boldsymbol{x}$ is the *comoving* separation vector, we then know from (7.16) the time dependence of the density fluctuations, $\delta(\boldsymbol{x},t) = D_+(t)\delta_0(\boldsymbol{x})$. Thus, within the scope of the validity of (7.16),

$$\xi(x,t) = D_+^2(t) \, \xi(x,t_0) \; , \tag{7.30}$$

and accordingly

$$P(k,t) = D_+^2(t) \, P(k,t_0) =: D_+^2(t) \, P_0(k) \; , \tag{7.31}$$

where k is a *comoving wave number*. We shall stress once again that these relations are valid only in the framework of Newtonian, linear perturbation theory in the matter dominated era of the Universe, to which we had restricted ourselves in Sect. 7.2.2. Equation (7.31) states that the knowledge of $P_0(k)$ is sufficient to obtain the power spectrum $P(k,t)$ at any time, again within the framework of linear perturbation theory.

7.4.1 The initial power spectrum

The Harrison–Zeldovich spectrum. Initially it may seem as if $P_0(k)$ is a function that can be chosen arbitrarily, but one objective of cosmology is to calculate this power spectrum and to compare it to observations. More than 30 years ago,

arguments were already developed to specify the functional form of the initial power spectrum.

At early times, the expansion of the Universe follows a power law, $a(t) \propto t^{1/2}$ in the radiation-dominated era. At that time, no natural length-scale existed in the Universe to which one might compare a wavelength. The only mathematical function that depends on a length but does not contain any characteristic scale is a power law[6]; hence for very early times one should expect

$$P(k) \propto k^{n_\mathrm{s}} \,. \tag{7.32}$$

Many years ago, Harrison, Zeldovich, Peebles and others argued that $n_\mathrm{s} = 1$, as for this slope, the amplitude of the fluctuations of the gravitational potential are constant, i.e., preferring neither small nor large scales. For this reason, the spectrum (7.32) with $n_\mathrm{s} = 1$ is called a scale-invariant spectrum, or *Harrison–Zeldovich spectrum*. With such a spectrum, we may choose a time t_i after the inflationary epoch and write

$$P(k, t_\mathrm{i}) = D_+^2(t_i)\, A\, k^{n_\mathrm{s}} \,, \tag{7.33}$$

where A is a normalization constant that cannot be determined from theory but has to be fixed by observations. As we will see in the following subsection, this is not the complete story: The result (7.33) needs to be modified to account for the different growth of the amplitude of density fluctuations in the radiation-dominated epoch of the Universe, compared to that in the later cosmic epochs from which our result (7.31) was derived.

Cold dark matter & hot dark matter. Furthermore, these modifications depend on the nature of the dark matter. One distinguishes between *cold dark matter (CDM)* and *hot dark matter (HDM)*. These two kinds of dark matter differ in the characteristic velocities of their constituents. Cold dark matter has a velocity dispersion that is negligible compared to astrophysically relevant velocities, e.g., the virial velocities of low-mass dark matter halos. Therefore, their initial velocity dispersion can well be approximated by zero, and all dark matter particles have the bulk velocity $\boldsymbol{u}$ of the cosmic 'fluid' (before the occurrence of multiple streams). In contrast, the velocity dispersion of hot dark matter is appreciable; as mentioned in Sect. 4.4.6 before, neutrinos are the best candidates for hot dark matter, in view of their known abundance, determined from the thermal history of the Universe (see Sect. 4.4), and their finite rest mass. The characteristic velocity of neutrinos is fully specified by their rest mass; despite their low temperature of $T_\nu \sim 1.9\,\mathrm{K}$ today, their thermal velocities of

$$v_\nu \sim 150\,(1+z)\left(\frac{m_\nu}{1\,\mathrm{eV}}\right)^{-1} \mathrm{km/s} \tag{7.34}$$

prevent them from forming matter concentrations at all mass scales except for the most massive ones, as their velocity is larger than the corresponding escape velocities. In other words, the finite velocity dispersion of hot dark matter is equivalent to assigning to it a pressure, which prevents them to fall into shallow gravitational potential wells. We will see below the dramatic differences between these two kinds of dark matter for the formation of structures in the Universe. In particular, this estimate shows that neutrinos cannot account for the dark matter on galaxy scales, and thus cannot explain the flat rotation curves of spiral galaxies.

If density fluctuations become too large on a certain scale, linear perturbation theory breaks down and (7.31) is no longer valid. Then the true current-day power spectrum $P(k, t_0)$ will deviate from $P_0(k)$. Nevertheless, in this case it is still useful to examine $P_0(k)$—it is then called the *linearly extrapolated power spectrum*.

7.4.2 Growth of density perturbations and the transfer function

Within the framework of linear Newtonian perturbation theory in the 'cosmic fluid', $\delta(\boldsymbol{x}, t) = D_+(t)\,\delta_0(\boldsymbol{x})$ applies. Modifications to this behavior are necessary for several reasons:

- If dark matter consists (partly) of hot dark matter, this may not be gravitationally bound to the potential well of a density concentration. In this case, the particles are able to move freely and to escape from the potential well, which in the end leads to its dissolution if these particles dominate the matter overdensity. From this argument, it follows immediately that for HDM small-scale density perturbations cannot form. For CDM this effect of *free streaming* does not occur. In our discussion of the evolution of density perturbations, we have not included the possible presence of hot dark matter, since we neglected any pressure term in the hydrodynamic equations.
- At redshifts $z \gtrsim z_\mathrm{eq}$, radiation dominates the density of the Universe. Since the expansion law $a(t)$ is then distinctly different from that in the matter-dominated phase, the growth rate for density fluctuations will also change.
- As discussed in Sect. 4.5.2, a cosmic horizon exists with comoving scale $r_\mathrm{H,com}(t)$. Physical interactions can take

[6] You can convince yourself of this by trying to find another type of function of a scale that does not involve a characteristic length; e.g., $\sin x$ does not work if x is a length, since the sine of a length is not defined; one thus needs something like $\sin(x/x_0)$, hence introducing a length-scale. The same arguments apply to other functions, such as the logarithm, the exponential etc. Also note that the sum of two power laws, e.g., $Ax^\alpha + Bx^\beta$ defines a characteristic scale, namely that value of x where the two terms become equal.

place only on scales smaller than $r_{\rm H,com}(t)$. For fluctuations of length-scales $L \sim 2\pi/k \gtrsim r_{\rm H,com}(t)$, Newtonian perturbation theory will cease to be valid, and one needs to apply linear perturbation theory in the framework of the General Relativity.

The transfer function. These effects together will lead to a modification of the shape of the power spectrum, relative to the relation (7.33); for example, the evolution of perturbations in the radiation-dominated cosmos proceeds differently from that in the matter-dominated era. The power spectrum $P(k)$ is affected by the combination of the above effects, and will be different from the primordial spectral shape, $P \propto k^{n_{\rm s}}$. The modification of the power spectrum is described in terms of the *transfer function* $T(k)$, in the form

$$P(k,t) = D_+^2(t)\, Ak^{n_{\rm s}}\, T^2(k)\ . \tag{7.35}$$

The transfer function can be computed for any cosmological model if the matter content of the universe is specified. In particular, $T(k)$ depends on the nature of dark matter.

CDM and HDM. The first of the above points immediately implies that a clear difference must exist between HDM and CDM models regarding structure formation and evolution. In HDM models, small-scale fluctuations are washed out by free-streaming of relativistic particles, i.e., the power spectrum is completely suppressed for large k, which is expressed by the transfer function $T(k)$ decreasing exponentially for large k. In the context of such a theory, very large structures will form first, and galaxies can form only later by fragmentation of large structures. However, this formation scenario is in clear contradiction with observations. For example, we observe galaxies and QSOs at $z > 6$ so that small-scale structure is already present at times when the Universe had less than 10 % of its current age. In addition, the observed correlation function of galaxies, both in the local Universe (see Fig. 7.4) and at higher redshift, is incompatible with cosmological models in which the dark matter is composed mainly of HDM.

Hot dark matter leads to structure formation that does not agree with observation. Therefore we can exclude HDM as the dominant constituent of dark matter. For this reason, it is now commonly assumed that the dark matter is 'cold'. The achievements of the CDM scenario in the comparison between model predictions and observations fully justify this assumption.

We shall elaborate on the last statement in quite some detail in Chap. 8. Anticipating these results, for most of the rest of this chapter we will neglect the possible presence of a pressure component of the dark matter, i.e., we will concentrate of cold dark matter.

Relevance of horizon size for structure growth. In linear perturbation theory, fluctuations grow at the same rate on all scales, or for all wave numbers, independent of each other. This applies not only in the Newtonian case, but also remains valid in the framework of General Relativity as long as the fluctuation amplitudes are small. Therefore, the behavior on any (comoving) length-scale can be investigated independently of the other scales. At very early times, perturbations with a comoving scale L are larger than the (comoving) horizon, and only for $z < z_{\rm enter}(L)$ does the horizon become larger than the considered scale L. Here, $z_{\rm enter}(L)$ is defined as the redshift at which the (comoving) horizon equals the (comoving) length-scale L,

$$r_{\rm H,com}(z_{\rm enter}(L)) = L\ . \tag{7.36}$$

It is common to say that at $z_{\rm enter}(L)$ the perturbation under consideration 'enters the horizon', whereas actually the process is the opposite—the horizon outgrows the perturbation. Relativistic perturbation theory shows that density fluctuations of scale L grow as long as $L > r_{\rm H,com}$, namely $\propto a^2$ if radiation dominates (thus, for $z > z_{\rm eq}$), or $\propto a$ if matter dominates (i.e., for $z < z_{\rm eq}$). Free-streaming particles or pressure gradients cannot impede the growth on scales larger than the horizon length because, according to the definition of the horizon, physical interactions—which pressure or free-streaming particles would be—cannot extend to scales larger than the horizon size.

Qualitative behavior of the transfer function. The behavior of the growth of a density perturbation on a scale L for $z < z_{\rm enter}(L)$ depends on $z_{\rm enter}$ itself. If a perturbation enters the horizon in the radiation-dominated phase, $z_{\rm eq} \lesssim z_{\rm enter}(L)$, it ceases to grow during the epoch $z_{\rm eq} \lesssim z \lesssim z_{\rm enter}(L)$. In this period, the energy density in the Universe is dominated by radiation, and the resulting expansion rate prevents an efficient perturbation growth. At later epochs, when $z \lesssim z_{\rm eq}$, the growth of density perturbations continues. If $z_{\rm enter}(L) \lesssim z_{\rm eq}$, thus if the perturbation enters the horizon during the matter-dominated epoch of the Universe, these perturbations will grow as described in Sect. 7.2.2, with $\delta \propto D_+(t)$. This implies that a length-scale L_0 is singled out, namely the one for which

$$z_{\rm eq} = z_{\rm enter}(L_0)\ , \tag{7.37}$$

so that L_0 is the comoving horizon size at matter-radiation equality. We can calculate this length-scale explicitly from (4.74),

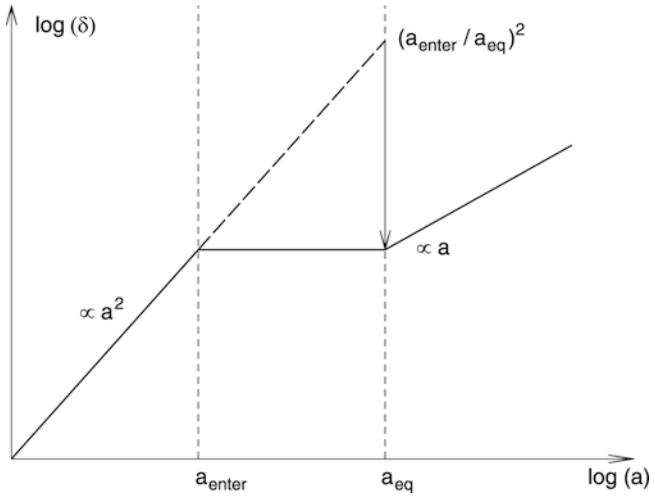

Fig. 7.5 A density perturbation that enters the horizon during the radiation-dominated epoch of the Universe ceases to grow until matter starts to dominate the energy content of the Universe. In comparison to a perturbation that enters the horizon later, during the matter-dominated epoch, the amplitude of the smaller perturbation is suppressed by a factor $(a_{\rm enter}/a_{\rm eq})^2$, which explains the qualitative behavior (7.40) of the transfer function. Adapted from: M. Bartelmann & P. Schneider 2001, *Weak Gravitational Lensing*, Phys. Rep. 340, 291

$$L_0 = r_{\rm H,com}(z_{\rm eq}) = \int_0^{a_{\rm eq}} \frac{c\,{\rm d}a}{a^2\,H(a)} = \frac{c}{H_0\sqrt{\Omega_{\rm m}}} \int_0^{a_{\rm eq}} \frac{{\rm d}a}{\sqrt{a+a_{\rm eq}}}$$
$$= \left(\sqrt{2}-1\right)\frac{2c}{H_0}\left(\frac{a_{\rm eq}}{\Omega_{\rm m}}\right)^{1/2}\,, \qquad (7.38)$$

where we made use of the Hubble function at early times where curvature and vacuum energy density are negligible, and the relation (4.30) for $a_{\rm eq}$ was used to write $H^2 = H_0^2\Omega_{\rm m}(a+a_{\rm eq})a^{-4}$. Again using (4.30), we finally obtain

$$\boxed{L_0 \approx 16(\Omega_{\rm m}h^2)^{-1}\ {\rm Mpc}}\,. \qquad (7.39)$$

Density fluctuations with $L > L_0$ enter the horizon after matter started to dominate the energy density of the Universe; hence their growth is not impeded by a phase of radiation-dominance. In contrast, density fluctuations with $L < L_0$ enter the horizon at a time when radiation was still dominating. These then cannot grow further as long as $z > z_{\rm eq}$, and only in the matter-dominated epoch will their amplitudes proceed to grow again. Their relative amplitude up to the present time has therefore grown by a smaller factor than that of fluctuations with $L > L_0$ (see Fig. 7.5): since fluctuations larger than the horizon grow $\propto a^2$ in the radiation-dominated era, a perturbation of scale $L < L_0$ has its amplitude suppressed by a factor $\left[a_{\rm enter}(L)/a_{\rm eq}\right]^2$ until matter-radiation equality, compared to a perturbation with scale $> L_0$. Since $r_{\rm H,com} \propto a$ in the radiation-dominated era [see (4.76)], these small-scale perturbations are suppressed by a factor $(L/L_0)^2$ relative to large-scale perturbations.

The quantitative consideration of these effects allows us to compute the transfer function. In general, this needs to be done numerically, but very good approximations exist. In particular, since all the physics involved at these early epochs concern small perturbations, and thus can safely be treated in a linear approximation, these numerical calculations pose no principal problems. Several codes are publicly available for calculating the transfer function for a specified set of cosmological parameters. For the limiting cases of $L \gg L_0$ and $L \ll L_0$, our previous discussion yields

$$T(k) \approx 1 \quad \text{for} \quad k \ll 1/L_0\,,$$
$$T(k) \approx (kL_0)^{-2} \quad \text{for} \quad k \gg 1/L_0\,. \qquad (7.40)$$

However, the important point is:

> In the framework of the CDM model, the transfer function can be computed, and thus, by means of (7.35), also the power spectrum of the density fluctuations as a function of length-scale and redshift. The amplitude of the power spectrum has to be obtained from observations.

The shape parameter. The transfer function depends on the combination kL_0, which is the inverse of the ratio of the length-scale under consideration ($\sim 2\pi/k$) and the horizon scale at the epoch of equality, and thus on $k(\Omega_{\rm m}h^2)^{-1}$. Since distances determined from redshift are measured in units of h^{-1} Mpc, the wave number is measured in units of $h\,{\rm Mpc}^{-1}$. Therefore, the shape of the transfer function, and thus also that of the power spectrum, depends on $\Gamma = \Omega_{\rm m}h$. Γ is called the *shape parameter* of the power spectrum. It is sometimes used as a free parameter instead of being identified with $\Omega_{\rm m}h$. A more detailed analysis shows that Γ depends also on $\Omega_{\rm b}$, but since $\Omega_{\rm b} \lesssim 0.05$ is small, according to primordial nucleosynthesis (see Sect. 4.4.5), this effect is relatively small—however, this small effect turns out to be of great importance for observational cosmology (see Sect. 7.4.3 below).

If the galaxy distribution follows the distribution of dark matter, the former can be used to determine the correlation function or the power spectrum, in particular the shape parameter Γ. Both from the distribution of galaxies projected onto the sphere (angular correlation function) and from the three-dimensional galaxy distribution (which is determined from redshift surveys), one finds that $\Gamma \sim 0.2$ (see Sect. 8.1.3). From $T(k) \approx 1$ for $kL_0 \ll 1$, and with (7.33), we find that $P(k) \propto k^{n_{\rm s}}$ for $kL_0 \ll 1$,

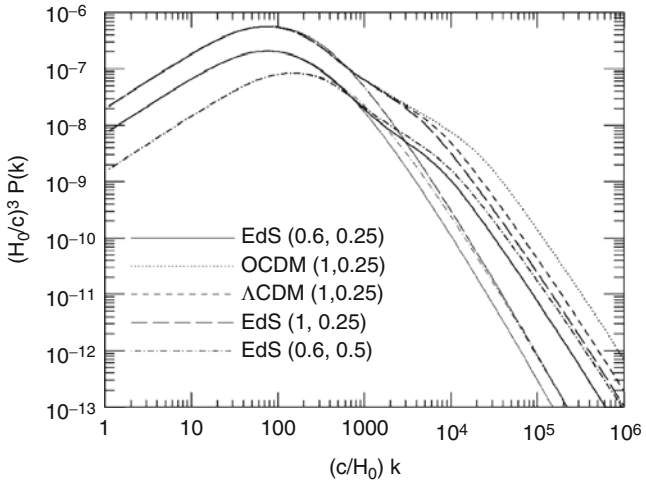

Fig. 7.6 The current power spectrum of density fluctuations for CDM models. The wave number k is given in units of H_0/c, and $(H_0/c)^3\,P(k)$ is dimensionless. The various *curves* have different cosmological parameters: EdS: $\Omega_\mathrm{m} = 1$, $\Omega_\Lambda = 0$; OCDM: $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0$; ΛCDM: $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$. The values in *parentheses* specify (σ_8, Γ), where σ_8 is the normalization of the power spectrum (which will be discussed below), and where Γ is the shape parameter. The *thin curves* correspond to the power spectrum $P_0(k)$ linearly extrapolated to the present day, and the *bold curves* take the non-linear evolution into account. For this figure, the effects of baryons on the transfer function have been neglected

with $n_\mathrm{s} \approx 1$. This behavior is compatible with the CMB anisotropy measurements on large scales, as we will discuss in detail in Sect. 8.6.

In Fig. 7.6, the power spectrum is plotted for several cosmological models that have different density parameter, shape parameter, and normalization of the power spectrum. The thin curves show $P(k)$ as derived from linear perturbation theory, and the bold curves display the power spectrum with non-linear structure evolution taken approximately into account. The power spectra displayed all have a characteristic wave number at which the slope of $P(k)$ changes, or where the peak of $P(k)$ is located. It is specified by $k \sim 2\pi/L_0 \approx (\pi\Gamma/8)h\,\mathrm{Mpc}^{-1}$, with the characteristic length L_0 being defined in (7.39). The value of the shape parameter Γ determines the location of this peak.

Transfer function for other dark matter models. As mentioned before, the evolution of density fluctuations depends on the nature of dark matter. The asymptotic relation (7.40) is valid only for cold dark matter. From our previous discussion, we infer that the free streaming of hot dark matter would erase fluctuations on nearly all scales; in particular, if the hot dark matter particles are relativistic at the epoch of matter-radiation equality, then all fluctuations on scales smaller than L_0 would be destroyed (since the length scale over which a relativistic dark matter particle can freely stream equals the horizon scale). Accordingly, the transfer function for a hot dark matter universe will have an exponential decline for $kL_0 \ll 1$.

Besides pure CDM and HDM models (the latter being excluded by observation), one can consider models which are dominated by CDM, but which have a (small) contribution by HDM; these are called mixed dark matter (MDM) models. Such a contribution has indeed now become part of the standard model, due to the detected finite rest mass of neutrinos which implies $0 < \Omega_\nu \ll 1$. With this contribution, $T(k)$ is changed in such a way that small scales (i.e., large k) are slightly damped in the power spectrum. We will see later that by observing the power spectrum we can constrain the rest mass of neutrinos very well, and cosmological observations provide, in fact, by far the most stringent mass limits for neutrinos.

More exotic dark matter models suggest the existence of particles which are intermediate between cold and hot dark matter, in that their impact on the transfer function is a damping of small-scale fluctuations. The free-streaming of the warm dark matter particles has a similar effect as that of the neutrinos, except that their mass is higher, hence the velocity is lower, and thus the cut-off scale in the power spectrum is shifted to smaller length-scales. These warm dark matter models were introduced to reduce the apparent discrepancy between the abundance of satellite galaxies and the number of dark matter subclumps predicted by cold dark matter models (see Sect. 7.8). Despite lacking a natural candidate for the constituent of warm dark matter from particle physics, its properties can be strongly constrained with recent cosmological observations; we will come back to this issue further below.

The linear theory of the evolution of density fluctuations will break down at the latest when $|\delta| \sim 1$; the above equations for the power spectrum $P(k,t)$ are therefore valid only if the respective fluctuations are small. However, accurate fitting formulae now exist for $P(k,t)$ which are also valid in the non-linear regime. For some cosmological models, the non-linear power spectrum is displayed in Fig. 7.6.

7.4.3 The baryonic density fluctuations

The evolution of density fluctuations of baryons differs from that of dark matter. The reason for this is essentially the interaction of baryons with photons: although matter dominates the Universe for $z < z_\mathrm{eq}$, the energy density of baryons remains smaller than that of the photons for a longer time, until after recombination begins, as can be seen as follows: The baryon-to-photon density ratio is

$$\frac{\rho_\mathrm{b}}{\rho_\gamma} = \frac{\Omega_\mathrm{b} a^{-3}}{\Omega_\gamma a^{-4}} = a\frac{\Omega_\mathrm{b}}{\Omega_\mathrm{m}}\frac{\Omega_\mathrm{m}}{\Omega_\mathrm{r}}\frac{\Omega_\mathrm{r}}{\Omega_\gamma} = 1.68\frac{a}{a_\mathrm{eq}}\frac{\Omega_\mathrm{b}}{\Omega_\mathrm{m}} \sim 0.28\frac{a}{a_\mathrm{eq}}\,, \tag{7.41}$$

where we used the expression (4.30) for $a_{\rm eq}$, and (4.28) for the radiation-to-photon density—the neutrinos, which contribute to the radiation density, are not coupled to the baryons. In the final step, we used the estimate $\Omega_{\rm b} \sim \Omega_{\rm m}/6$ for our Universe. Hence, if radiation-matter equality happens at $z \sim 3000$, then the photon density is larger than that of the baryons for $z \gtrsim 800$.

Since photons and baryons interact with each other by photon scattering on free electrons, which again are tightly coupled electromagnetically to protons and helium nuclei, baryons and photons are strongly coupled before recombination, and form a single fluid. Due to the presence of photons, this fluid has a strong pressure, which prevents it from falling into potential wells formed by the dark matter. Thus, the pressure prevents strong inhomogeneities of the baryon-photon fluid.

Sound waves. To discuss the evolution of baryon perturbations in a bit more detail, we consider again a perturbation of comoving scale L. As long as the perturbation is larger than the horizon size, pressure effects can not affect the behavior of the fluid, and thus baryons and photons behave in the same way as the dark matter—the amplitude of their perturbations grow. As soon as the perturbation enters the horizon, the situation changes. Although the baryons are gravitationally pulled into the density maxima of the dark matter, pressure provides a restoring force which acts against a compression of the baryon-photon fluid. As a results, this fluid will develop sound waves.

Sound horizon. The maximum distance sound waves can travel up to a given epoch is called the *sound horizon*. Loosely speaking, it is given by the product of the sound speed and the cosmic time and has a very similar meaning as the (event) horizon that we discussed before. The sound speed in this photon-dominated fluid is given by $c_{\rm s} \approx c/\sqrt{3}$, as will be shown shortly. Thus, the sound horizon is about a factor of $\sqrt{3}$ smaller than the event horizon. As soon as a perturbation enters the sound horizon, the amplitude of the baryon-photon fluctuations can not grow anymore; instead, they undergo damped oscillations.

The adiabatic sound velocity $c_{\rm s}$ of a fluid is given in general by

$$c_{\rm s}^2 = \frac{\partial P}{\partial \rho} \; .$$

The pressure of the fluid is generated by the photons, $P = c^2\rho_\gamma/3 = c^2\rho_{\rm cr}\Omega_\gamma a^{-4}/3$, and the density is the sum of that of baryons and photons, $\rho = (\Omega_{\rm b}a^{-3} + \Omega_\gamma a^{-4})\rho_{\rm cr}$. Thus, the sound velocity is

$$c_{\rm s} = \sqrt{\frac{{\rm d}P/{\rm d}a}{{\rm d}\rho/{\rm d}a}} = \frac{c}{\sqrt{3}}\sqrt{\frac{4\Omega_\gamma a^{-5}}{3\Omega_{\rm b}a^{-4} + 4\Omega_\gamma a^{-5}}} = \frac{c}{\sqrt{3(1+\mathcal{R})}}, \tag{7.42}$$

where we have defined

$$\mathcal{R} = \frac{3}{4}\frac{\rho_{\rm b}}{\rho_\gamma} = \frac{3}{4}\frac{\Omega_{\rm b}}{\Omega_\gamma}\,a \; . \tag{7.43}$$

Note that $\mathcal{R}$ is smaller than unity until recombination, and thus $c_{\rm s} \approx c/\sqrt{3}$ provides a reasonable first approximation.

At recombination, the free electrons recombined with the hydrogen and helium nuclei, after which there are essentially no more free electrons which couple to the photon field. Hence, after recombination the baryon fluid lacks the pressure support of the photons, and the sound speed drops to zero—the sound waves do no longer propagate, but get frozen in. Now the baryons are free to react to the gravitational field created by the dark matter inhomogeneities, and they can fall into their potential wells. After some time, the spatial distribution of the baryons is essentially the same as that of the dark matter.

Hence, there is a maximum wavelength of the sound waves, namely the (comoving) sound horizon at recombination, $r_{\rm s}$, which can be calculated according to (4.74),

$$r_{\rm s} = \int_0^{a_{\rm rec}} \frac{c\,{\rm d}a}{\sqrt{3(1+\mathcal{R})}a^2 H(a)} \; , \tag{7.44}$$

except that we exchanged the speed of light by the speed of sound. Using the Hubble function for the early universe, where only matter and radiation are relevant, we find

$$\begin{aligned} r_{\rm s} &= \frac{c}{H_0\sqrt{\Omega_{\rm m}}\sqrt{3(1+\mathcal{R}_{\rm eff})}} \int_0^{a_{\rm rec}} \frac{{\rm d}a'}{\sqrt{a'+a_{\rm eq}}} \\ &= \frac{2c}{H_0\sqrt{\Omega_{\rm m}}\sqrt{3(1+\mathcal{R}_{\rm eff})}} \left(\sqrt{a_{\rm rec}+a_{\rm eq}} - \sqrt{a_{\rm eq}}\right) \\ &\sim \frac{120h^{-1}{\rm Mpc}}{\sqrt{\Omega_{\rm m}}\sqrt{1+\mathcal{R}_{\rm eff}}} \; , \end{aligned} \tag{7.45}$$

where $\mathcal{R}_{\rm eff}$ is a mean of $\mathcal{R}$ over the integration range, and the final expression is a rough estimate, obtained by assuming $a_{\rm eq} \ll a_{\rm rec} \sim 10^{-3}$.

Figure 7.7 illustrates the physical significance of this length scale, showing the time evolution of an initial density peak of all four components in the Universe. The length scale $r_{\rm s}$ is the distance the baryon-photon fluid propagates outwards from the initial density peak before baryons and photons decouple, after which the density perturbation of

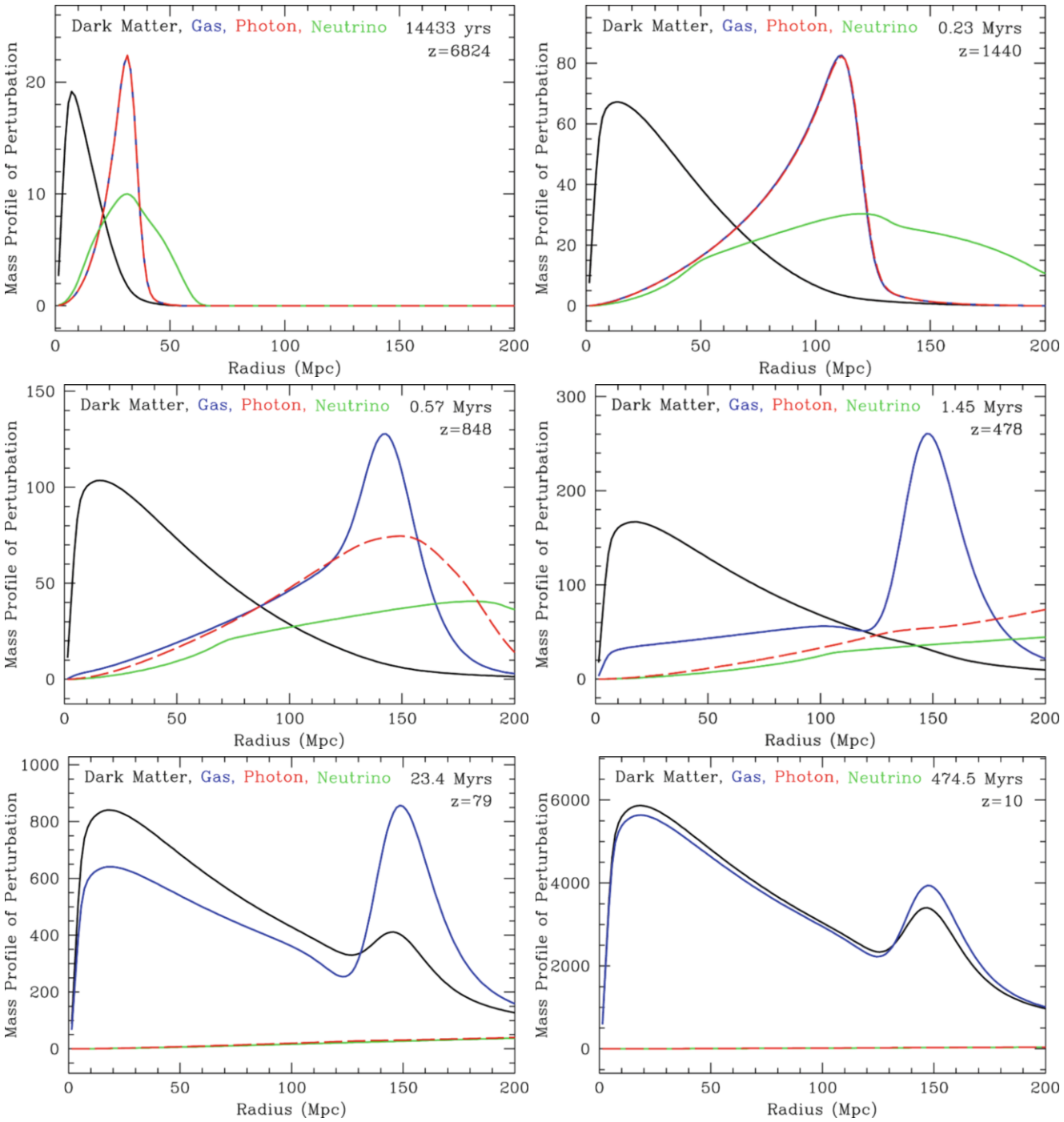

Fig. 7.7 Evolution in time of an initial density peak in all components of the cosmic matter. The x-axis shows the comoving radial coordinate, the y-axis displays the density, multiplied by (radius)^2. The different snapshots show the spatial distribution of the various species at later epochs. Neutrinos freely stream out of the perturbation at the speed of light. The photon and baryons are strongly coupled before recombination, and thus have the same spatial distribution. They move out from the initial density peak with the sound speed, $c_s \approx c/\sqrt{3}$. After recombination, the photons are no longer coupled to the baryons and freely stream away; correspondingly, the sound speed of the baryons drops to zero, and they stop propagating outwards. After that, the baryons are gravitationally attracted by the density peak of the dark matter and fall in; however, some of the matter also falls into the density peaks (in the example of this figure, it is an overdense spherical shell) created by baryons, whereas the density profile of neutrinos and photons becomes flat. At late time, the distributions of baryons and dark matter become identical (before the onset of non-linear processes such as halo formation). The central density peak, and the secondary peak have a well-defined separation, given by the distance a sound wave could travel before the baryons decoupled from the photons. Source: D.J. Eisenstein et al. 2007, *On the Robustness of the Acoustic Scale in the Low-Redshift Clustering of Matter*, ApJ 664, 660, p. 662, Fig. 1,

baryons gets frozen. Hence, this phenomenon is indeed characterized by the length-scale r_s.[7]

Baryonic acoustic oscillations. The sound waves in the baryon-photon fluid, the baryonic acoustic oscillations (BAOs), are observable today. Since at recombination, the photons interacted with matter for the last time, the cosmic microwave background radiation provides us with a picture of the density fluctuations at the epoch of recombination. Our observable cosmic microwave sky essentially is a picture of a two-dimensional cut at fixed time (the time of last scattering) through the density field of the baryons. A cut through an ensemble of sound waves shows an instantaneous picture of these waves. Hence, they are expected to be visible in the temperature distribution of the CMB. As we will see in Sect. 8.6, this is indeed the case: These BAOs imprint one of the most characteristic features on the CMB anisotropies. Since the sound waves are damped once they are inside the sound horizon, the largest amplitude waves are those whose wavelength equals the sound horizon at recombination.

BAOs in the low-redshift Universe. We have argued that the baryons, once they are no longer coupled to radiation and thus become pressureless, fall into the potential wells of the dark matter. This happens because the dark matter fluctuations can grow while the baryonic fluctuations could not due to the photon pressure, and because the mean density of dark matter is substantially larger than that of the baryons. This is almost the full story, but not entirely: Baryons make about 15 % of the total matter density, and are therefore not negligible. After recombination, the BAOs are frozen, like standing waves, and thus the total matter fluctuations are a superposition of the dark matter inhomogeneities and these standing waves. Whereas the dark matter dominates the density fluctuations, a small fraction of the matter also follows the inhomogeneities created by the standing waves. Since these waves have a characteristic length scale—the sound horizon at recombination—this characteristic length scale should be visible in the properties of the matter distribution even today. As we shall see in Sect. 8.1, the correlation function of galaxies contains a characteristic feature at the length scale r_s. Hence, relics of the sound waves in the pre-recombination era are even visible in the current Universe. The effects of the BAOs are included in the transfer function $T(k)$, which thus shows some low-amplitude oscillations, often called 'wiggles' (they not seen in Fig. 7.6, since there the transfer function of a dark matter-only models was used).

[7] One may wonder why the neutrinos in Fig. 7.7 have a broad distribution in space, and not simply a sharp shell—they all stream out with velocity c. The reason is that the initial conditions were chosen such that they correspond to a growing mode. For those, we argued that any density perturbation is associated with a velocity perturbation. As can be seen from (7.25), the relation between density and peculiar velocity is non-local, i.e., the velocity field associated with a density peak at the center is non-zero at all radii. This velocity field also causes the dark matter distribution to expand—once the neutrinos and the photon-baryon fluid have moved out, the gravitational field inside the dark matter peak is weaker than it would have been predicted from a pure dark matter distribution, yielding an expanding peculiar velocity field near the center.

7.5 Non-linear structure evolution

Linear perturbation theory has a limited range of applicability; in particular, the evolution of structures like clusters of galaxies cannot be treated within the framework of linear perturbation theory. One might imagine that one can evolve the system of (7.8), (7.10) and (7.11) to higher order in the small variables δ and $|\boldsymbol{u}|$, and thus consider non-linear perturbation theory. In fact, a quite extensive literature exists on this topic in which such calculations are performed. However, while this higher-order perturbation theory indeed allows us to follow density fluctuations to somewhat larger values of $|\delta|$, the mathematical effort required is substantial. In addition, the fluid approximation is no longer valid if gravitationally bound systems form because, as mentioned earlier, multiple steams of matter will occur in this case.

However, for some interesting limiting cases, analytical descriptions exist which are able to represent the characteristics of the non-linear evolution of the mass distribution in the Universe. We shall now investigate a special and very important case of such a non-linear model. In general, studying the non-linear structure evolution requires the use of numerical methods. Therefore, we will also discuss some aspects of such numerical simulations.

7.5.1 Model of spherical collapse

Assumptions. We consider a spherical region in an expanding universe, with its density $\rho(t)$ enhanced compared to the mean cosmic density $\bar{\rho}(t)$,

$$\rho(t) = [1 + \delta(t)]\,\bar{\rho}(t)\ , \tag{7.46}$$

where we use the density contrast δ as defined in (7.1). For reasons of simplicity we assume that the density within the sphere is homogeneous although, as we will later see, this is not really a restriction. The density perturbation is assumed to be small for small t, so that it will grow linearly at first, $\delta(t) \propto D_+(t)$, as long as $\delta \ll 1$. If we consider a time t_i which is sufficiently early such that $\delta(t_i) \ll 1$, then according to the definition of the growth factor D_+, $\delta(t_i) = \delta_0\, D_+(t_i)$, where δ_0 is the density contrast linearly extrapolated to the present day. It should be mentioned once again that $\delta_0 \neq$

$\delta(t_0)$, because the latter is in general affected by the non-linear evolution.

Let R be the initial *comoving* radius of the overdense sphere; as long as $\delta \ll 1$, the *comoving* radius will change only marginally. The mass within this sphere is

$$M = \frac{4\pi}{3} R^3 \rho_0 (1 + \delta_\mathrm{i}) \approx \frac{4\pi}{3} R^3 \rho_0 \,, \tag{7.47}$$

because the physical (or proper) radius is $R_\mathrm{phys} = aR$, and $\bar{\rho} = \rho_0/a^3$. This means that a unique relation exists between the initial *comoving* radius and the mass of this sphere, independent of the choice of t_i and δ_0, if only we choose $\delta(t_\mathrm{i}) = \delta_0 \, D_+(t_\mathrm{i}) \ll 1$.

Evolution. Due to the enhanced gravitational force, the sphere will expand slightly more slowly than the universe as a whole, which will lead to an increase in its density contrast. This then decelerates the expansion rate even further, relative to the cosmic expansion rate. Indeed, the equations of motion for the radius of the sphere are identical to the Friedmann equations for the cosmic expansion, only with the sphere having an effective Ω_m different from that of the mean universe. If the initial density is sufficiently large, the expansion of the sphere will come to a halt, i.e., its proper radius $R_\mathrm{phys}(t)$ will reach a maximum; after this, the sphere will recollapse.

If t_max is the time of maximum expansion, then the sphere will, theoretically, collapse to a single point at time $t_\mathrm{coll} = 2t_\mathrm{max}$. The relation $t_\mathrm{coll} = 2t_\mathrm{max}$ follows from the time reversal symmetry of the equation of motion: the time to the maximum expansion is equal to the time from that point back to complete collapse.[8] The question of whether the expansion of the sphere will come to a halt depends on the density contrast $\delta(t_\mathrm{i})$ or δ_0—compare the discussion of the expansion of the Universe in Sect. 4.3.1—and on the model for the cosmic background.

Special case: Einstein–de Sitter model. In the special case of $\Omega_\mathrm{m} = 1$ and $\Omega_\Lambda = 0$, this behavior can easily be quantified analytically; we thus treat this case separately. In this cosmological model, any sphere with $\delta_0 > 0$ is a "closed universe" and will therefore recollapse at some time. For the collapse to take place before t_1, $\delta(t_\mathrm{i})$ or δ_0 needs to exceed a threshold value. For instance, for a collapse at $t_\mathrm{coll} \le t_0$, a linearly extrapolated overdensity of

$$\delta_0 \ge \delta_\mathrm{c} = \frac{3}{20} (12\pi)^{2/3} \simeq 1.69 \tag{7.48}$$

is required. More generally, one finds that $\delta_0 \ge \delta_\mathrm{c} \, (1 + z)$ is needed for the collapse to occur before redshift z.

One can calculate δ_c also for other values of the density parameters. It turns out that the modifications are relatively small, and thus the value (7.48) is a useful approximation also for other cosmological models.

Violent relaxation and virial equilibrium. Of course, the sphere will not really collapse to a single point. This would only be the case if the sphere was perfectly homogeneous and if the particles in the sphere moved along perfectly radial orbits. In reality, small-scale density and gravitational fluctuations will exist within such a sphere. These then lead to deviations of the particles' tracks from perfectly radial orbits, an effect that becomes more important as the density contrast of the sphere increases. The particles will scatter on these fluctuations in the gravitational field and will virialize; this process of *violent relaxation* has already been described in Sect. 6.3.3 and occurs on short time-scales—roughly the dynamical time-scale, i.e., the time it takes the particles to fully cross the sphere. In this case, the virialization is essentially complete at t_coll. After that, the sphere will be in virial equilibrium, and its average density will be[9]

$$\langle\rho\rangle = (1 + \delta_\mathrm{vir}) \; \bar{\rho}(t_\mathrm{coll}), \quad \text{where} \quad (1 + \delta_\mathrm{vir}) \simeq 18\pi^2 \approx 178 \,, \tag{7.49}$$

where the final expression is obtained for an EdS model. With the same assumption, corresponding expressions can be derived for other models as well. However, since the numerical factor for the overdensity of a virialized halo depends on the idealized assumptions made in the spherical collapse model, its exact value is of little importance. Nevertheless, the foregoing relation forms the basis for the statement that the virialized region, e.g., of a cluster, is a sphere with an average density $\sim$ 200 times the critical density ρ_cr of the Universe at the epoch of collapse. Another conclusion from this consideration is that a massive galaxy cluster with a virial radius of $1.5\,h^{-1}$Mpc must have formed from the collapse of a region that originally had a comoving radius larger by about an order of magnitude (see problem 7.3). Such a virialized mass concentration of dark matter is called a *dark matter halo*.

[8] This occurs for the same reason that it takes a stone thrown up into the air the same time to reach its peak altitude as to fall back to the ground from there.

[9] This result is obtained from conservation of energy and from the virial theorem. The total energy E_tot of the sphere is a constant. At the time of maximum expansion, it is given solely by the gravitational binding energy of the system since then the expansion velocity, and thus the kinetic energy, vanishes. On the other hand, the virial theorem implies that in virial equilibrium $E_\mathrm{kin} = -E_\mathrm{pot}/2$, and by combining this with the conservation of energy $E_\mathrm{tot} = E_\mathrm{kin} + E_\mathrm{pot}$ one is then able to compute E_pot in equilibrium and hence the radius and density of the collapsed sphere. For an EdS model, $r_\mathrm{vir} = r_\mathrm{max}/2$.

Up to now, we have considered the collapse of a homogeneous sphere. From the above arguments one can easily convince oneself that the model is still valid if the sphere has a radial density profile with a density that decreases outwards. In this case, the initial density contrast will also decrease as a function of radius. The inner regions of such a sphere will then collapse faster than the outer ones; a halo of lower mass will form first, and only later, when the outer regions have also collapsed, will a halo with higher mass form. From this it follows that halos of low initial mass will grow in mass by further accretion of matter.

The spherical collapse model is a simple model for the non-linear evolution of a density perturbation in the Universe. Despite being simplistic, it represents the fundamental principles of gravitational collapse and yields approximate relations, e.g., for the collapse time and mean density inside the virialized region, as they are found from numerical simulations.

7.5.2 Number density of dark matter halos

Press–Schechter model. The model of spherical collapse allows us to approximately compute the number density of dark matter halos as a function of their mass and redshift; this model is called the *Press–Schechter model*.

We consider a field of density fluctuations $\delta_0(\boldsymbol{x})$, featuring fluctuations on all scales according to the power spectrum $P_0(k)$. Assume that we smooth this field with a *comoving* smoothing length R, by convolving it with a filter function of this scale. In our example of the waves on a lake, we could examine a picture of its surface taken through a pane of milk-glass, by which all the contours on small scales would be blurred. Then, let $\delta_R(\boldsymbol{x})$ be the smoothed density field, linearly extrapolated to the present day. This field does not contain any fluctuations on scales $\lesssim R$, because these have been smoothed out. Each maximum in $\delta_R(\boldsymbol{x})$ corresponds to a peak with characteristic scale $\gtrsim R$ and, according to (7.47), each of these maxima corresponds to a mass peak of mass $M \sim (4\pi R^3/3)\rho_0$. If the amplitude δ_R of the density peak is sufficiently large, a sphere of (comoving) radius R around the peak will decouple from the linear growth of density fluctuations and will begin to grow non-linearly. Its expansion will come to a halt, and then it will recollapse. This process is similar to that in the spherical collapse model and can be described approximately by this model. The density contrast required for the collapse, $\delta_R \geq \delta_{\min}$, can be computed for any cosmological model and for any redshift.

If the statistical properties of $\delta_0(\boldsymbol{x})$ are Gaussian—which is expected for a variety of reasons—the statistical properties of the fluctuation field δ_0 are completely defined by the power spectrum $P(k)$. Then the abundance of density maxima with $\delta_R \geq \delta_{\min}$ can be computed, and hence the (comoving) number density $n(M, z)$ of relaxed dark matter halos in the Universe as a function of mass M and redshift z can be determined.

The mass spectrum. The most important results of the Press–Schechter model are easily explained (see Fig. 7.8). The number density of halos of mass M depends of course on the amplitude of the density fluctuation δ_0—i.e., on the normalization of the power spectrum $P_0(k)$. Hence, the normalization of $P_0(k)$ can be determined by comparing the prediction of the Press–Schechter model with the observed number density of galaxy clusters, as we will discuss further in Sect. 8.2.1 below. The corresponding result is called the "cluster-normalized power spectrum".

Furthermore, we find that $n(M, z)$ is a decreasing function of halo mass M. This follows immediately from the previous argument, since a larger M requires a larger smoothing length $R \propto M^{1/3}$, together with the fact that the number density of mass peaks of a given amplitude $\delta_{\min}$ decreases with increasing smoothing length. For large M, $n(M, z)$ decreases exponentially because sufficiently high peaks become very rare for large smoothing lengths. Therefore, *very* few clusters with mass $\gtrsim 2 \times 10^{15} M_\odot$ exist today. At higher redshift, the cut-off in the abundance is at smaller masses, so that massive clusters are expected to be increasingly rare at higher z. From Fig. 7.8, we can see that the number density of clusters with $M \gtrsim 10^{15} M_\odot$ today is about $10^{-7}\,\mathrm{Mpc}^{-3}$, so the average separation between two such clusters is larger than 100 Mpc, which is compatible with the observation that the most nearby massive cluster (Coma) is about 90 Mpc away from us.

The density contrast $\delta_{\min}$ required for a collapse before redshift z is a function of z, as we have seen above. In particular, for the Einstein–de Sitter model we have $\delta_{\min} \simeq 1.69(1+z)$. In general, $\delta_{\min} \approx 1.69/D_+(z)$. This means that the redshift dependence of $\delta_{\min}$ depends on the cosmological model and is basically described by the growth factor $D_+(z)$. Since $D_+(z)$ is, at fixed z [we recall that, by definition, $D_+(0) = 1$], larger for smaller Ω_m (see Fig. 7.3), the ratio of the number density of halos at redshift z to the one in the current Universe, $n(M, z)/n(M, 0)$, is larger the smaller Ω_m is. For cluster masses ($M \sim 10^{15} M_\odot$), the evolution of this ratio in the Einstein–de Sitter model is dramatic, whereas it is less strong in open and in flat, Λ-dominated universes (see Fig. 7.8).

Density fluctuations on a given mass scale. We considered above the smoothed density field $\delta_R(\boldsymbol{x})$, which corresponds to a mass $M = (4\pi/3)R^3\rho_0$. Like before, we here interpret δ and δ_R as the linear density field. As was true for the original density field, the expectation value—or the spatial average—of the smoothed density field vanishes, $\langle \delta_R(\boldsymbol{x}) \rangle =$

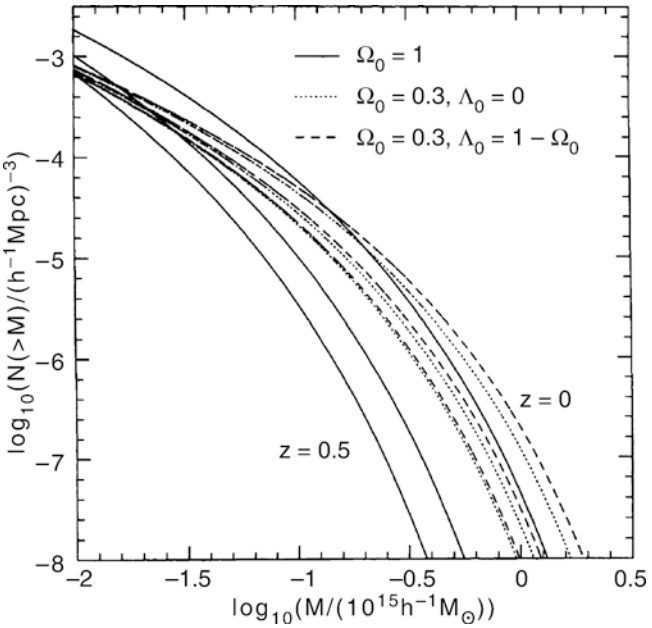

Fig. 7.8 Number density of dark matter halos with mass $> M$, computed from the Press–Schechter model. The comoving number density is shown for three different redshifts, $z = 0$ (*upper curves*), $z = 0.33$, and $z = 0.5$ (*lower curves*), for three different cosmological models: an Einstein–de Sitter model (*solid curves*), a low-density open model with $\Omega_m = 0.3$ and $\Omega_\Lambda = 0$ (*dotted curves*), and a flat universe of low density with $\Omega_m = 1 - \Omega_\Lambda = 0.3$ (*dashed curves*). The normalization of the density fluctuation field has been chosen such that the number density of halos with $M > 10^{14}h^{-1}M_\odot$ at $z = 0$ in all models agrees with the local number density of galaxy clusters. Note in particular the dramatic redshift evolution in the EdS model. Source: V.R. Eke et al. 1996, *Cluster evolution as a diagnostic for* Ω, MNRAS 282, 263, p. 269, Fig. 4. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

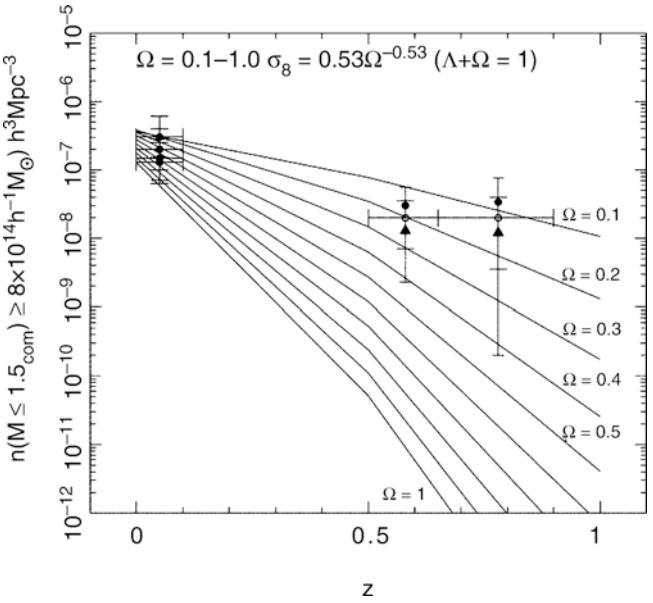

Fig. 7.9 Expected (comoving) number density of galaxy clusters with mass $> 8 \times 10^{14}h^{-1}M_\odot$ within a (comoving) radius of $R < 1.5h^{-1}$ Mpc, for flat cosmological models and different values of the density parameter Ω_m. The normalization of the power spectrum in the models has been chosen such that the current cluster number density is approximately reproduced. The *points* with error bars show results from observations of galaxy clusters at different redshift—although the error bars at high redshift are very large, a high density Universe is seen to be excluded. Source: N.A. Bahcall & X. Fan 1998, *The Most Massive Distant Clusters: Determining* Ω *and* σ_8, ApJ 504, 1, p. 2, Fig. 1. ©AAS. Reproduced with permission

0. The amplitude of fluctuations of the smoothed field can be characterized by the dispersion of the field,

$$\sigma^2(M) := \left\langle |\delta_R(\boldsymbol{x})|^2 \right\rangle , \tag{7.50}$$

which depends on the smoothing scale, and thus on the mass (these two variables are therefore interchangeable). The larger the smoothing scale, the smaller are the relative fluctuations of the resulting smoothed field. Hence, $\sigma(M)$ is a monotonically decreasing function of the mass. The larger $\sigma(M)$, the more abundant are peaks with an amplitude above some threshold at this mass scale. This is particularly true for the density threshold δ_c, required for collapse until today. Indeed, the halo abundance as predicted by Press-Schechter theory depends only on the ratio ν between the density threshold required for collapse and the dispersion of fluctuations on a given mass scale,

$$\nu := \frac{\delta_c}{D_+(z)\,\sigma(M)} , \tag{7.51}$$

where we accounted for the fact that collapse before redshift z requires a fluctuation amplitude of $\delta_c / D_+(z)$.

A first application. By comparing the number density of galaxy clusters at high redshift with the current abundance, we can thus obtain constraints on Ω_m, and in some sense also on Ω_Λ. Even a few very massive clusters at $z \gtrsim 0.5$ are sufficient to rule out the Einstein–de Sitter model by this argument. As a matter of fact, the existence of the cluster MS 1054−03 (Fig. 6.21) alone, the mass of which was determined by dynamical methods, from its X-ray emission, and by the gravitational lensing effect, is already sufficient to falsify the Einstein–de Sitter model (see Fig. 7.9).[10] However, at least one problem exists in the application of this method, namely making a sufficiently accurate mass determination for distant clusters and, in addition, determining whether they are virialized and thus are described by the Press–Schechter model. Also the completeness of the local cluster sample is a potential, though smaller problem.

[10] Until about 2000, this cluster was the highest-redshift massive cluster known.

A special case. To get a more specific impression of the Press–Schechter mass spectrum, we consider the special case where the power spectrum $P_0(k)$ is described by a power law, $P_0(k) \propto k^n$. From Fig. 7.6, we can see that this provides quite a good description over a large range of k if one concentrates on scales either clearly above or far below the maximum of P_0. The length-scale at which P_0 has its maximum is specified roughly by (7.39). As we can also see from Fig. 7.6, the non-linear evolution that the Press–Schechter model refers to is relevant only for scales considerably smaller than this maximum, rendering the power law a useful first approximation, with $n \sim -1.5$. In this case, the mass function can be written in closed form,

$$\frac{\mathrm{d}n}{\mathrm{d}M}(M,z) = \frac{\rho_{\mathrm{cr}}\Omega_{\mathrm{m}}}{\sqrt{\pi}}\,\frac{\gamma}{M^2}\left(\frac{M}{M^*(z)}\right)^{\gamma/2} \times \exp\left[-\left(\frac{M}{M^*(z)}\right)^{\gamma}\right] , \qquad (7.52)$$

where $(\mathrm{d}n/\mathrm{d}M)\mathrm{d}M$ is the comoving number density of halos with mass in the interval between M and $M + \mathrm{d}M$, $\gamma = 1 + n/3 \sim 0.5$, and $M^*(z)$ is the z-dependent mass-scale above which the mass spectrum is exponentially cut off. More specifically, $M^*(z)$ is defined as the mass where the parameter ν [see (7.51)] is unity, i.e., it is given implicitly by

$$\sigma(M^*(z)) = \delta_{\mathrm{c}}/D_+(z) . \qquad (7.53)$$

For masses considerably smaller than $M^*(z)$, the Press–Schechter mass spectrum is basically a power law in M. The characteristic mass-scale $M^*(z)$ for this particular model can be derived explicitly,

$$M^*(z) = M_0^*\,[D_+(z)]^{2/\gamma} = M_0^*\,(1+z)^{-2/\gamma} , \qquad (7.54)$$

where the final expression applies to an Einstein–de Sitter universe only. Hence, the characteristic mass-scale grows over time, and it describes the mass-scale on which the mass distribution in the universe is just becoming non-linear for a particular redshift. This mass-scale at the current epoch, M_0^*, depends on the normalization of the power spectrum; it approximately separates groups from clusters of galaxies, and explains the fact that clusters are (exponentially) less abundant than groups.

Hierarchical structure formation. Furthermore, the Press–Schechter model describes a very general property of structure formation in a CDM model, namely that low-mass structures—like galaxy-mass dark halos—form at early times, whereas large mass accumulations evolve only later. The explanation for this is found in the shape of the power spectrum $P(k)$ as described in (7.35) together with the asymptotic form (7.40) of the transfer function $T(k)$. A model like this is also called a *hierarchical structure formation* or a 'bottom-up'-scenario. In such a model, small structures that form early later merge to form large structures.

Comparison with numerical simulations. The Press–Schechter model is a very simple model, based on assumptions that are not really justified in detail. Nevertheless, its predictions are in astounding agreement with the number density of halos determined from simulations, and this model, published in 1974, has for nearly 25 years predicted the halo density with an accuracy that was difficult to achieve in numerical simulations. Only since the mid-1990s have the precision and statistics of numerical simulations of structure formation reached a level on which significant discrepancies with the Press–Schechter model became clearly noticeable. However, the analytical description was also improved; generalizing a spherical collapse, the more realistic ellipsoidal collapse has been investigated, by which the number density of halos is modified relative to the Press–Schechter model. This advanced model is found to be in better agreement with numerical results. Furthermore, simple fit functions have been found which fit the dark matter halo abundance from numerical simulations very well, as demonstrated in Fig. 7.10, so that today we have a good description of $n(M,z)$ that very accurately resembles the results from numerical simulations. These mass functions share the property that the halo abundance depends only on the 'peak height' ν, defined in (7.51).

7.5.3 Numerical simulations of structure formation

Analytical considerations—such as, for instance, linear perturbation theory or the spherical collapse model—are only capable of describing limiting cases of structure formation. In general, gravitational dynamics is too complicated to be analytically examined in detail. For this reason, experiments to simulate structure formation by means of numerical methods have been performed for some time already. The results of these simulations, when compared to observations, have contributed very substantially to establishing the standard model of cosmology, because only through them did it become possible to quantitatively distinguish the predictions of this model from those of other models. Of course, the enormous development in computer hardware rendered corresponding progress in simulations possible; in addition, the continuous improvement of numerical algorithms has allowed a steadily improved spatial resolution of the simulations.

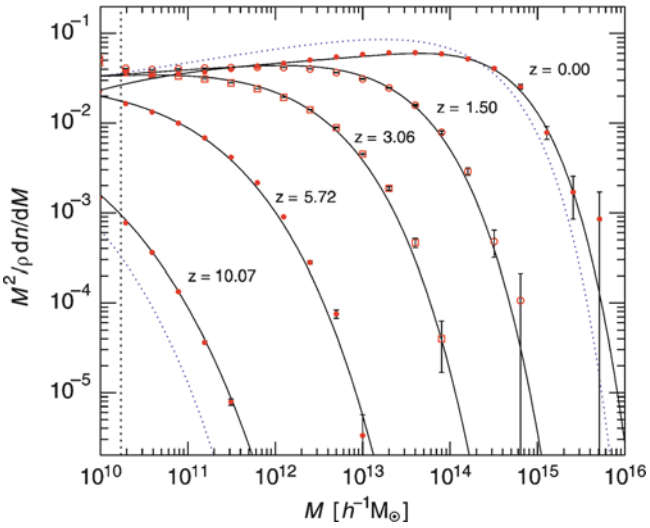

Fig. 7.10 The mass spectrum of dark matter halos is plotted for five different redshifts (data points with error bars), as measured in the Millennium Simulation (which we will discuss more extensively below—see Fig. 7.13). The *solid curves* describe an approximation for the mass spectrum, which was obtained from *different* simulations, and which obviously provides an excellent description of the simulation results. For $z = 0$ and $z = 10$, the prediction of the Press–Schechter model is indicated by the *dotted curves*, underestimating the abundance of very massive halos and overestimating the density of lower-mass halos. The *vertical dotted line* indicates the lowest halo mass which can still be resolved in these simulations. Source: V. Springel 2005, *Simulating the joint evolution of quasars, galaxies and their large-scale distribution*, Nature 435, 629, Fig. 2. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2005

Since the Universe is dominated by dark matter, for many purposes it is sufficient to compute the behavior of this matter component and thus to consider solely gravitational interactions. Only in recent years has computing power increased to a level where hydrodynamical processes can also approximately be taken into account, so that the baryonic component of the Universe can be traced as well. In addition, radiative transfer can be included in such simulations, hence the influence of radiation on the heating and cooling of the baryonic component can also be examined. We will describe such simulations in Sect. 10.6.1, when we discuss the cosmic evolution of the galaxy population.

The principle of simulations. Representative dark matter particles. We will now give a brief description of the principle of such simulations, where we confine ourselves to dark matter. Of course, no individual particles of dark matter are traced in the simulations: since it presumably consists of elementary particles, which therefore have a high number density, one would only be able to simulate an extremely small, microscopic section of the Universe. Rather, one examines the behavior of dark matter in the expanding Universe by representing its particles by bodies of mass M, and by then assuming that these "macroscopic particles" behave like the dark matter particles in a volume $V = M/\rho$. Effectively, this corresponds to the assumption that dark matter consists of particles of mass M. Since this assumption cannot be valid in detail, we will later need to modify the resulting equations of motion.

Choice of simulation volume. The next point one needs to realize right from the start is that one cannot simulate the full spatial volume of the Universe (which may be infinite) but only a representative section of it. Typically, a *comoving* cube with side length L is chosen. For this section to be representative, the linear extent L should be larger than the largest observed structures in the Universe. Otherwise, the effects of the large-scale structure would be neglected. For example, hardly any structure is found in the Universe on scales $\gtrsim 200\,h^{-1}$ Mpc, so that $L \gtrsim 200\,h^{-1}$ Mpc is a reasonable value for the comoving size of the cube. However, in a box of that size one cannot expect to get a representative result for the largest structures in the Universe (such as 'Great Walls') or for the abundance of very massive clusters, as their space density is of order $\sim 10^{-6}\left(h^{-1}\,\mathrm{Mpc}\right)^{-3}$, i.e., one expects to find at most a handful of very massive clusters in such a volume.

Since the numerical effort scales with the number of grid points at which the gravitational force is computed, $N_{\mathrm{grid}} = (L/\Delta x)^3$, and which is limited by the computer's speed and memory, the choice of L also immediately implies the length-scale of the numerical resolution. Furthermore, the total mass within the numerical volume is $\propto \Omega_{\mathrm{m}} L^3$, so that for a given maximum number of particles, the minimum mass that can be resolved in the simulation is also known.

Periodic boundary conditions. Since particles close to the boundaries of the cube also feel gravitational forces from matter outside the cube, one cannot simply assume the region outside the cube to be empty. We need to make assumptions about the matter distribution outside the numerical volume. Since one assumes that the Universe is essentially homogeneous on scales $> L$, the cube is extended periodically—for instance, a particle leaving the cube at its upper boundary will immediately re-enter the cube from the lower side, with the same velocity vector. The mass distribution (and with it also the force field) is periodic in these simulations, with a period of L. This assumption of periodicity has an effect on the results for the mass distribution on scales comparable to L; the quantitative analysis of the results from these simulations should therefore be confined to scales $\lesssim L/2$.

Computation of the force field. With the above assumptions, the equation of motion for all particles can now be set up. The force on the i-th particle is

$$\boldsymbol{F}_i = \sum_{j \neq i} \frac{M^2\,(\boldsymbol{r}_j - \boldsymbol{r}_i)}{\left|\boldsymbol{r}_j - \boldsymbol{r}_i\right|^3}\,, \tag{7.55}$$

thus the sum of forces exerted by all the other particles, where these are periodically extended. This aspect may appear at first sight more difficult than it actually is, as we will see next.

The computation of the force acting on individual particles by the summation (7.55) is not feasible in practice. For example, assume the simulation to trace 10^{10} particles, then in total 10^{20} terms need to be calculated using (7.55)—for each time step. Even on the most powerful computers this is not feasible today. To handle this problem, one evaluates the force in an approximate way. One first notes that the force experienced by the i-th particle, exerted by the j-th particle, is not very sensitive to small variations in the separation vector $\boldsymbol{r}_i - \boldsymbol{r}_j$, as long as these variations are much smaller than the separation itself. Except for the nearest particles, the force on the i-th particle can then be computed by introducing a grid into the cube and shifting the particles

in the simulation to the closest grid point.[11] With this, a discrete mass distribution on a regular grid is obtained.

The force field of this mass distribution can then be computed by means of a Fast Fourier Transform (FFT), a fast and very efficient algorithm. However, the introduction of the grid establishes a lower limit to the spatial force resolution. Because the size of the grid cells also defines the spatial resolution of the force field, it is chosen to be roughly the mean separation between two particles, so that the number of grid points is typically of the same order as the number of particles. This is called the PM (particle-mesh) method.

To achieve better spatial resolution, the interaction of closely neighboring particles can considered separately. This is done by splitting the gravitational potential $\Phi(r) = -GM/r$ of a particle into a short- and long-range part, $\Phi(r) = \Phi_\mathrm{s}(r) + \Phi_\mathrm{l}$. For example, one can choose $\Phi_\mathrm{s}(r) = \Phi(r)\, f(r/r_\mathrm{s})$, where the function f smoothly declines from $f(0) = 1$ to $f(1) = 0$, and $f(x) = 0$ for $x > 1$. Thus, the short-range gravitational potential $\Phi_\mathrm{s}(r)$ vanishes for $r > r_\mathrm{s}$. The long-range potential then is $\Phi_\mathrm{l}(r) = \Phi(r)\,[1 - f(r/r_\mathrm{s})]$, and hence vanishes at $r = 0$, whereas for $r > r_\mathrm{s}$, $\Phi_\mathrm{l} = \Phi$. The force on a particle is then given by the sum of the gradients of the short- and long-range potential. For the former, only those particles with separation $\le r_\mathrm{s}$ contribute, and this can be calculated by a sum of pairwise forces. On the other hand, the force field corresponding to Φ_l is smooth and is calculated by the grid method, as explained before. This kind of calculation of the force is called the $\mathrm{P^3M}$ (particle-particle particle-mesh) method.

Softening length. The force law (7.55) also describes strong collisions of particles, e.g., where a particle changes its velocity direction by $\sim 90°$ in a collision if it comes close enough to another particle. Of course, this effect is a consequence of replacing the dark matter constituents by macroscopic 'particles' of mass M. As we have seen in Sect. 3.2.4, the typical relaxation time-scale for a system is $\propto N/\ln N$, and since the mass in the numerical volume is defined by L, one has $N \propto 1/M$. Reducing the particles' mass and increasing N accordingly, the abundance of strong collisions would decrease, but computer power and memory is then a limiting factor. Thus to correct for the artefact of strong collisions, the force law is modified for small separations such that strong collisions no longer occur. The length-scale below which the force equation is modified ('softened') and deviates from $\propto 1/r^2$ is called *softening length*. Its choice depends on the method with which the force field is evaluated. If the force is calculated on a grid, as in the PM method, where the force resolution is limited by the size of grid cells, the softening length is typically chosen to be of similar size. On the other hand, for $\mathrm{P^3M}$-like methods, the softening length can be chosen substantially smaller, however at the expense of requiring smaller time steps. Of course, the softening length defines the smallest length scale on which the results of the simulation can be trusted: scales below or comparable to the softening length are not resolved, and the behavior on these small scales is affected by numerical artefacts.

Initial conditions and evolution. The initial conditions for the simulation are set at very high redshift. The particles are then distributed such that the power spectrum of the resulting mass distribution resembles a Gaussian random field with the theoretical (linear) power spectrum $P(k, z)$ of the cosmological model. The equations of motion for the particles with the force field described above are then integrated in time. The choice of the time step is a critical issue in this integration, as can be seen from the fact that the force on particles with relatively close neighbors will change more quickly than that on rather isolated particles. Hence, the time step is either chosen such that it is short enough for the former particles—which requires substantial computation time—or the time step is varied for different particles individually, which is clearly the more efficient strategy. For different times in the evolution, the particle positions and velocities are stored; these results are then available for subsequent analysis.

Examples of simulations. The size of the simulations, measured by the number of particles considered, has increased enormously in recent years with the corresponding increase in computing capacities and the development of efficient algorithms. In modern simulations, 1024^3 or even more particles are traced. One example of such a simulation is presented in Fig. 7.11, where the structure evolution was computed for four different cosmological models. The parameters for these simulations and the initial conditions (i.e., the initial realization of the random field) were chosen such that the resulting density distributions for the current epoch (at $z = 0$) are as similar as possible; by this, the dependence of the redshift evolution of the density field on the cosmological parameters can be recognized clearly. Comparing simulations like these with observations has contributed substantially to our realizing that the matter density in our Universe is considerably smaller than the critical density.

Massive clusters of galaxies have a very low number density, which can be seen from the fact that the massive cluster closest to us (Coma) is about 90 Mpc away. This is directly related to the exponential decrease of the abundance of dark matter halos with mass, as described by the Press–Schechter model (see Sect. 7.5.2). In simulations such as that shown in Fig. 7.11, the simulated volume is still too small to derive statistically meaningful results on such sparse mass concentrations. This difficulty has been one of the reasons for simulating considerably larger volumes. The Hubble Volume Simulations (see Fig. 7.12) use a cube with a side length of $3000h^{-1}$ Mpc, not much less than the currently visible Universe. This simulation is particularly well-suited to studying the statistical properties of very massive structures, like, e.g., the distribution of galaxy clusters. On the other hand, this large volume, together with the limited total number of particles that can be followed, means that the mass and spatial resolution of this simulation are insufficient for studying smaller-scale objects like galaxies.

The Millennium Simulation (MS) was performed in 2004, assuming a cosmological model with $\Omega_\mathrm{m} = 0.25$, $\Omega_\Lambda = 0.75$, a power spectrum normalization of $\sigma_8 = 0.9$, and a Hubble constant of $h = 0.73$. A cube of side length $500h^{-1}$ Mpc was considered, in which $(2160)^3 \approx 10^{10}$ particles with a mass of $8.6 \times 10^8 h^{-1}\, M_\odot$ each were traced. With this choice of parameters, one can spatially resolve the halos of galaxies. At the same time, the volume is large enough for the simulation to contain a large number of massive clusters whose evolutionary history can be followed.

[11] In practice, the mass of a particle is distributed to all eight neighboring grid points, with the relative proportion of the mass depending on the distance of the particle to each of these grid points.

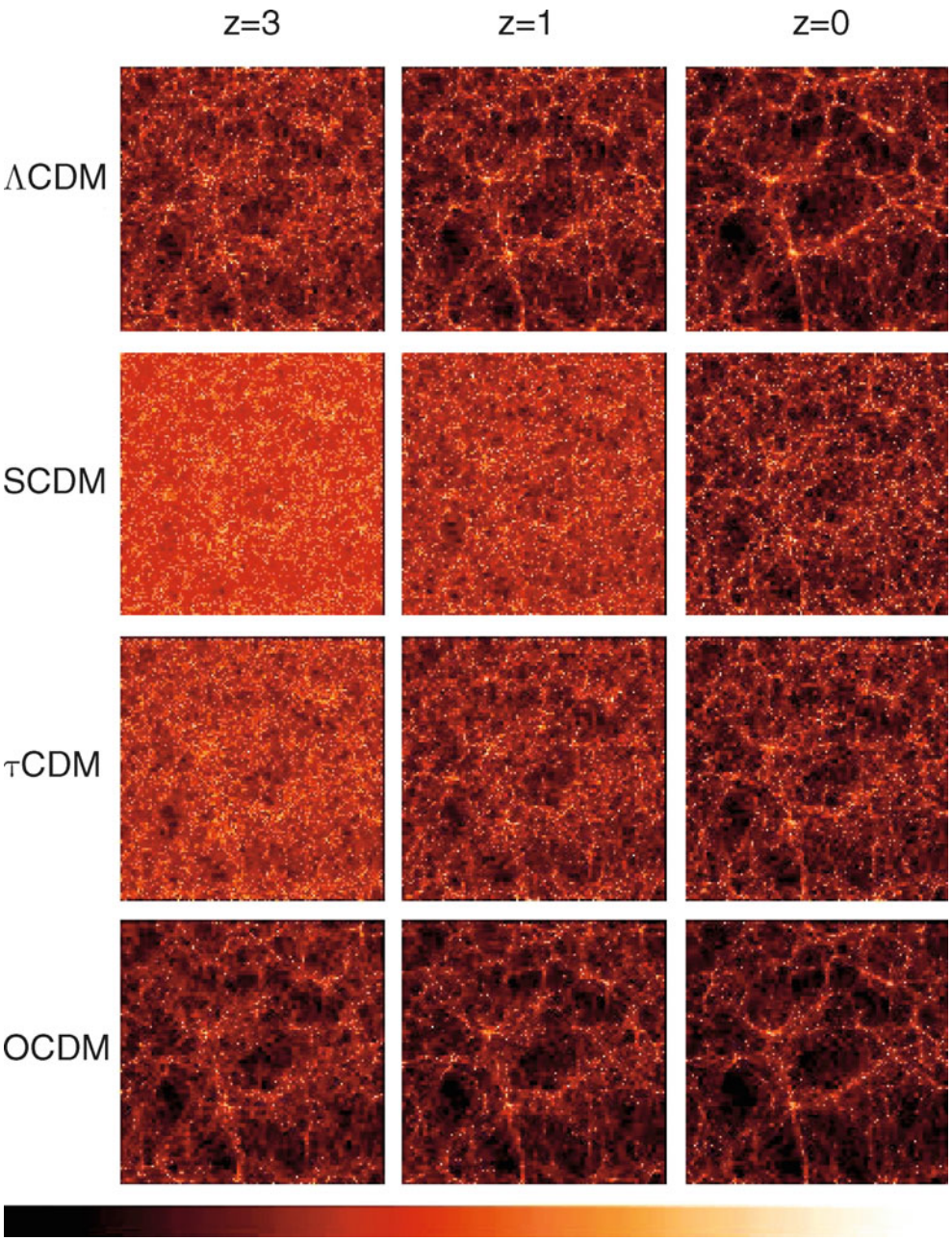

Fig. 7.11 Simulations of the dark matter distribution in the Universe for four different cosmological models: $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$ (ΛCDM), $\Omega_\mathrm{m} = 1.0$, $\Omega_\Lambda = 0.0$ (SCDM and τCDM), and $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0$ (OCDM). The two Einstein–de Sitter models differ in their shape parameter Γ which specifies the shape of the power spectrum $P(k)$ through the location of its peak. For each of the models, the mass distribution is presented for three different redshifts, $z = 3$, $z = 1$, and today, $z = 0$. Whereas the current mass distribution is quite similar in all four models (the model parameters were chosen as such), they clearly differ at high redshift. We can see, for instance, that significantly less structure has formed at high redshift in the SCDM model compared to the other models. From the analysis of the matter distribution at high redshift, one can therefore distinguish between the different models. In these simulations by the VIRGO Consortium, 256^3 particles were traced; the side length of the simulated volume is $\sim 240h^{-1}$ Mpc. Credit: VIRGO Collaboration, J. Colberg/MPA Garching. The simulations were carried out by the Virgo Supercomputing Consortium using computers based at the Computing Centre of the Max-Planck Society in Garching and at the Edinburgh parallel Computing Centre. Research article: A. Jenkins et al. 1998, *Evolution of Structure in Cold Dark Matter Universes*, ApJ 499, 20

The spatial resolution of the simulation is $\sim 5h^{-1}$ kpc, yielding a linear dynamic range of $\sim 10^5$. The resulting mass distribution at $z = 0$ is displayed in Fig. 7.13 in slices of $15h^{-1}$ Mpc thickness each, where the linear scale changes by a factor of four from one slice to the next. The images zoom in to a region around a massive cluster that becomes visible with its rich substructure in the uppermost slice, as well as filaments of the matter distribution, at the intersections of which massive halos form. The mass distribution in the Millennium Simulation is of great interest for numerous different investigations. We will discuss some of its results further in Chap. 10.

The MS was complemented by two related simulations, the Millennium-II (MS-II) and the Millennium-XXL (MXXL). Both assume the same cosmological parameters as the original Millennium Simulation, but differ in the volume considered. MS-II has the same number of particles as the original MS, but a five times smaller box size, i.e., $L = 100h^{-1}$ Mpc, yielding 125 times better mass resolution. MXXL treats a considerably larger box with $L = 3000h^{-1}$ Mpc, yielding the same comoving volume as the whole observable Universe within a redshift of $z = 0.72$, and $6720^3 \approx 3 \times 10^{11}$ particles, which makes it one of the largest N-body simulations up to now (2014). Together,

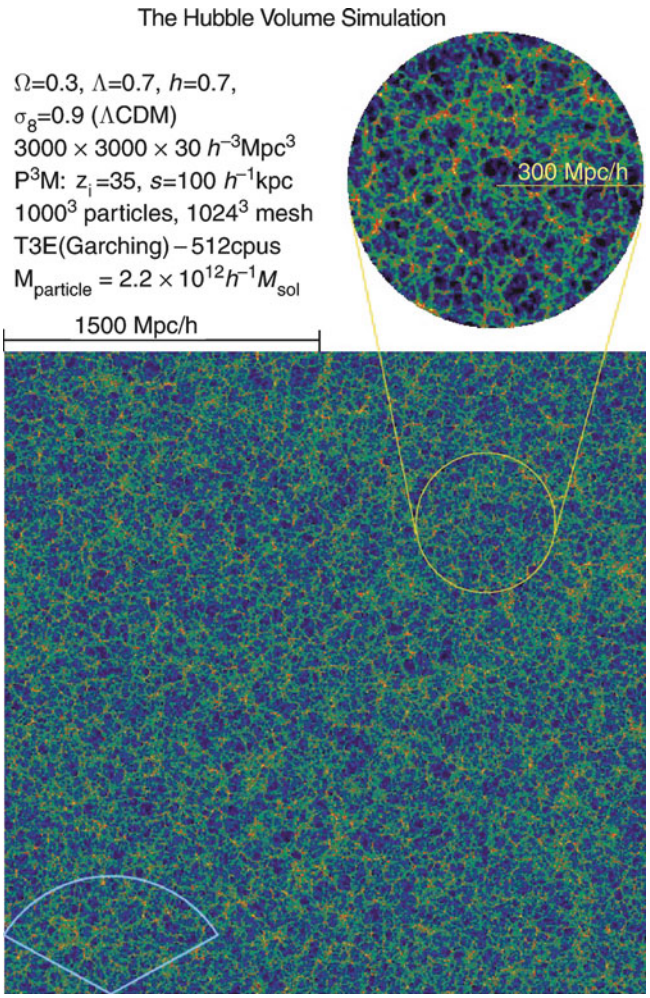

Fig. 7.12 The Hubble Volume Simulations: simulated is a box of volume $(3000h^{-1}\,\mathrm{Mpc})^3$, containing 10^9 particles, where a ΛCDM model with $\Omega_\mathrm{m} = 0.3$ and $\Omega_\Lambda = 0.7$ was chosen. Displayed is the projection of the density distribution of a $30h^{-1}$ Mpc thick slice of the cube. Simulations like this can be used to analyze the statistical properties of the mass distribution in the Universe on large scales. The sector in the lower left corner represents roughly the size of the CfA redshift survey (see Fig. 7.2). Credit: VIRGO Collaboration. The simulations were carried out by the Virgo Supercomputing Consortium using computers based at the Computing Centre of the Max-Planck Society in Garching and at the Edinburgh parallel Computing Centre. Research article: J.M. Colberg et al. 2000, *Clustering of galaxy clusters in CDM universes*, MNRAS, 319, 209

these three simulations can be used to study the effects of numerical resolution. For example, the combination of the simulations can measure the abundance of dark matter halos over nearly eight orders of magnitude in mass (see Fig. 7.14).

Analysis of numerical results. The analysis of the numerical results is nearly as intricate as the simulation itself because the positions and velocities of some 10^{10} particles alone do not provide any new insights. However, properly displaying the resulting mass distribution offers a first important insight into the large-scale structure in a CDM universe. In Fig. 7.15, we can see strong concentrations of the matter distribution, interlinked by large filaments. The overall structure resembles that of a web—often called the 'Cosmic Web'. The mass is assembled in sheets; where two sheets intersect, filaments form, and at the intersection of filaments are the more massive dark matter halos. These halos themselves are not smooth, but contain a rich substructure, as one recognizes in the smaller-scale zooms of Fig. 7.15. More insight is provided by looking at the time evolution of the density field; Fig. 7.16 shows four snapshots of the same region in the Millennium-II simulation. First, the density contrast increases with time, as expected from gravitational instability. At high redshift, only small mass concentrations have formed; more massive ones show up only at smaller redshift, in accordance to what we discussed in Sect. 7.5.2: Low-mass halos form first, massive ones only later. This is the hierarchical built-up of structure in a CDM universe. The halos grow in mass due to two different processes, by merging with other halos and by accretion of matter. The first process leads to the substructure (subhalos) of halos, which are the relics of earlier mergers. The accretion of matter is due to the gravitational pull of the mass concentrations on the surrounding matter. This accretion process happens predominantly through the filaments which are connected to the halo—it is therefore not a spherically symmetric effect.

The output of the simulation needs to be analyzed with respect to specific questions. Obviously, the (non-linear) power spectrum $P(k, z)$ of the matter distribution can be computed from the spatial distribution of particles, i.e., the density field; the corresponding results have led to the construction of the analytic fit formulae presented in Fig. 7.6. Furthermore, one can search for voids in the resulting particle distribution, which can then be compared to the observed abundance and typical size of voids.

Identification of dark matter halos. One of the main applications is the identification of collapsed mass concentrations (i.e., dark matter halos); their number density can be compared to predictions from the Press–Schechter model and to observations. For this, one needs to specify what a dark matter halo is and how this specification can be applied to the output of simulations. The spherical collapse model suggests that a halo is a spherical region inside of which the mean density is $\sim$ 200 times the critical density, but we recall that this particular value of 200 was based on a number of idealized assumptions. Furthermore, the dark matter concentrations usually deviate quite strongly from spherical symmetry.

Different methods for the identification of halos are used in simulations; for example, based on the position and velocities of particles, one can consider a halo to consist of

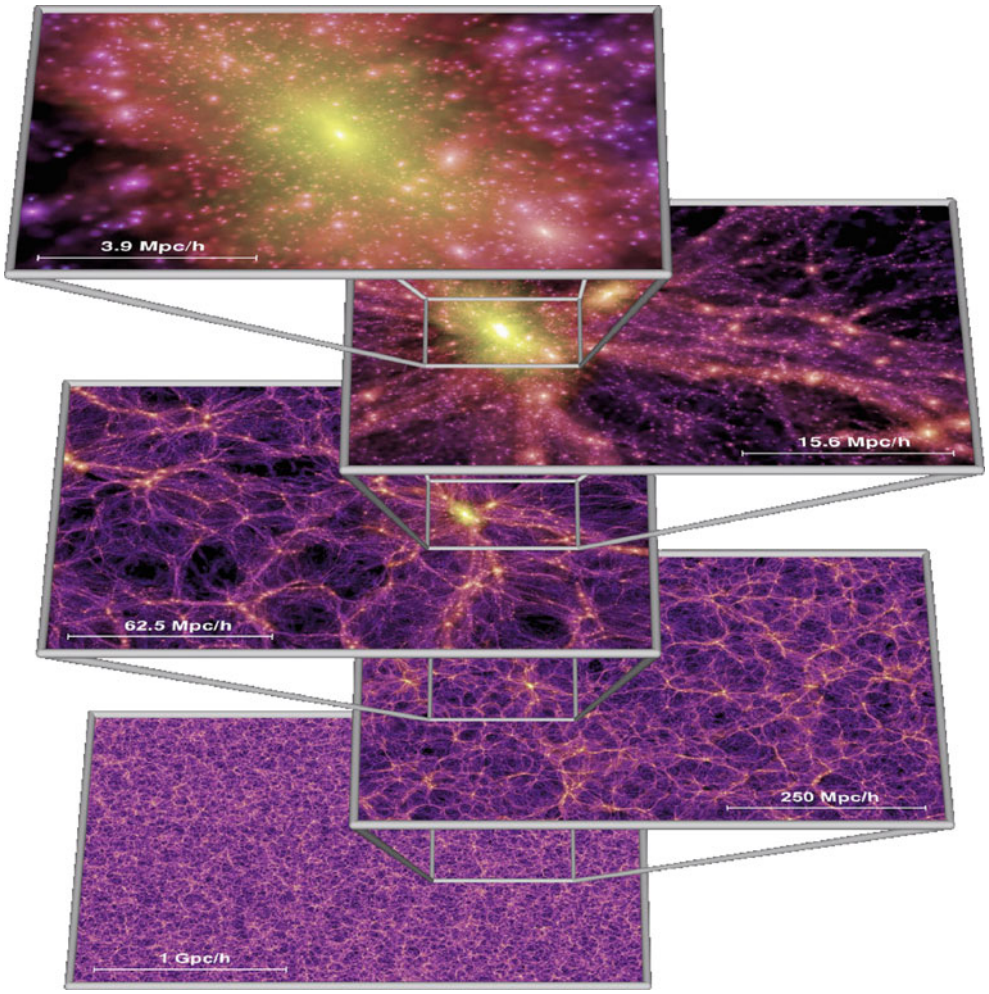

Fig. 7.13 Distribution of matter in slices of thickness $15h^{-1}$ Mpc each, computed in the Millennium Simulation. This simulation took about a month in 2004, running on 512 CPU processors. The output of the simulation, i.e., the position and velocities of all 10^{10} particles at 64 times steps, has a data volume of ~ 27 TB. The region shown in the two lower slices is larger than the simulated box which has a sidelength of $500h^{-1}$ Mpc; nevertheless, the matter distribution shows no periodicity in the figure as the slice was cut at a skewed angle to the box axes. Source: V. Springel 2005, *Simulating the joint evolution of quasars, galaxies and their large-scale distribution*, Nature 435, 629, Fig. 1. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2005

all particles which are gravitationally bound to it. Perhaps the most frequently used method consists of linking all particles whose separation is smaller than a fraction b of the mean particle separation $\sqrt{1/n}$, where $n = N/L^3$ is the number density of particles. Then those particles which are connected by a link are considered to be members of a halo; this is called the friends-of-friends algorithm. One finds from numerical experiments that by choosing $b = 0.2$, the characteristic density of these halos is about 200 times the critical density. Obviously, the way a halo is defined will affect the resulting mass spectrum, as can be seen in Fig. 7.14, where results from the 'gravitationally bound' and the friends-of-friends method are compared. The same ambiguity arises when the abundance of dark matter halos from simulations are compared to observed abundances of astronomical objects, such as galaxies or clusters.

It has been found that the Press–Schechter mass function represents the basic aspects of the mass spectrum astonishingly well, but a comparison with more recent simulations has shown significant deviations. More accurate formulae for the mass spectrum of halos have been constructed from simulations (see Figs. 7.10 and 7.14).

The direct link between the results from dark matter simulations and the observed properties of the Universe requires an understanding of the relation between dark matter and luminous matter. Dark matter halos in simulations cannot be compared to the observed galaxy distribution without further assumptions, e.g., on the mass-to-light ratio. We will return to these aspects later.

7.6 Properties of dark matter halos

The bound and virialized structures formed in the evolution of the cosmic density field, i.e., the dark matter halos, are of particular interest since they are believed to host the luminous objects in the Universe, galaxies and clusters. Therefore, we shall describe their properties in more detail here.

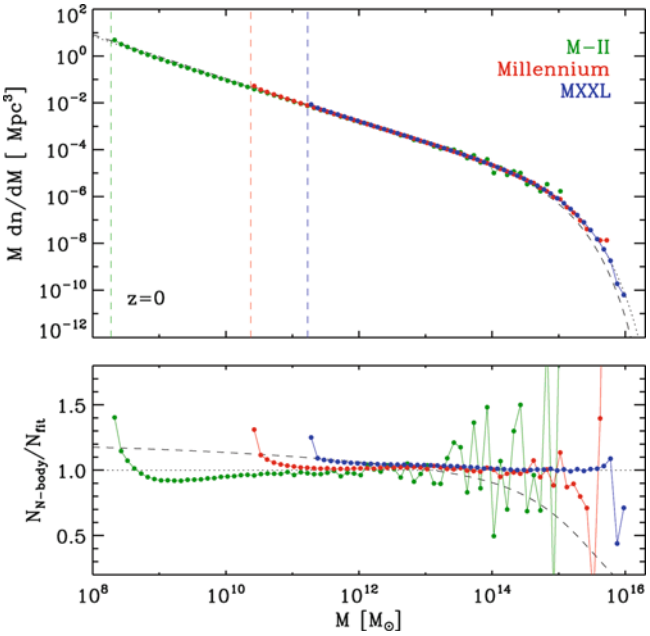

Fig. 7.14 The *upper panel* shows the comoving abundance of dark matter halos at $z = 0$, as obtained from the Millennium Simulation, the Millennium-II and the Millennium-XXL, shown in *red*, *green* and *blue*, respectively. Due to the different size and mass resolution, each of these simulations can determine the abundance best in a limited range of masses; together, they cover almost a factor 10^8 in mass. The *vertical dashed lines* indicates the minimum mass for each simulation, taken to be the mass of 20 simulation particles. The *dotted curve* shows a fit to the mass spectrum. In the *lower panel*, the ratio of the halo abundance and the fit is shown. The agreement between the results of these three simulations is clearly seen. Here, halos are identified by the friends-of-friends method. The *dashed curve* shows a fit to the halo abundance when a halo is identified by the set of all self-bound particles. Obviously, the mass spectrum depends on the details of the characterization of a halo. Source: R.E. Angulo et al. 2012, *Scaling relations for galaxy clusters in the Millennium-XXL simulation*, MNRAS 426, 2046, Fig. 2. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

7.6.1 Profile of dark matter halos

As already mentioned above, dark matter halos can be identified in mass distributions generated by numerical simulations. Besides the abundance of halos as a function of their mass and redshift, their radial mass profile can also be analyzed if individual halos are represented by a sufficient number of dark matter particles. The ability to obtain halo mass profiles depends on the mass resolution of a simulation. A surprising result has been obtained from these studies, namely that halos seem to show a universal density profile. We will briefly discuss this result in the following.

If we define a halo as described above, i.e., as a spherical region within which the average density is $\sim$ 200 times the critical density at the respective redshift, the mass M of the halo is related to its (virial) radius r_{200} by

$$M = \frac{4\pi}{3} r_{200}^3 \, 200 \, \rho_{\rm cr}(z) \; .$$

Since the critical density at redshift z is specified by $\rho_{\rm cr}(z) = 3H^2(z)/(8\pi G)$, we can write this as

$$M = \frac{100 r_{200}^3 H^2(z)}{G} \; , \tag{7.56}$$

so that at each redshift, a unique relation exists between the halo mass and its radius. We can also define the virial velocity V_{200} of a halo as the circular velocity at the virial radius,

$$V_{200}^2 = \frac{GM}{r_{200}} \; . \tag{7.57}$$

Combining (7.56) and (7.57), we can express the halo mass and virial radius as a function of the virial velocity,

$$M = \frac{V_{200}^3}{10GH(z)} \; , \quad r_{200} = \frac{V_{200}}{10H(z)} \; . \tag{7.58}$$

Since the Hubble function $H(z)$ increases with redshift, the virial radius at fixed virial velocity decreases with redshift. From (7.56) we also see that r_{200} decreases with redshift at fixed halo mass. Hence, halos at a given mass (or given virial velocity) are more compact at higher redshift than they are today, because the critical density was higher in the past.

The NFW profile. The *density profile of halos* averaged over spherical shells seems to have a universal functional form, which was first reported by Julio Navarro, Carlos Frenk & Simon White in a series of articles in the mid-1990s. This *NFW-profile* is described by

$$\boxed{\rho(r) = \frac{\rho_{\rm s}}{(r/r_{\rm s})(1 + r/r_{\rm s})^2}} \; , \tag{7.59}$$

where $r_{\rm s}$ specifies a characteristic radius, and $\rho_{\rm s} = 4\rho(r_{\rm s})$ determines the amplitude of the density profile. For $r \ll r_{\rm s}$ we find $\rho \propto r^{-1}$, whereas for $r \gg r_{\rm s}$, the profile follows $\rho \propto r^{-3}$. Therefore, $r_{\rm s}$ is the radius at which the slope of the density profile changes (see Fig. 7.17). $\rho_{\rm s}$ can be expressed in terms of $r_{\rm s}$, since, according to the definition of r_{200},

$$\begin{aligned} \bar{\rho} = 200\rho_{\rm cr}(z) &= \frac{3}{4\pi r_{200}^3} \int_0^{r_{200}} 4\pi r^2 \, {\rm d}r \, \rho(r) \\ &= 3\rho_{\rm s} \int_0^1 \frac{{\rm d}x \; x^2}{c \; x \; (1 + cx)^2} \; , \end{aligned} \tag{7.60}$$

where in the last step the integration variable was changed to $x = r/r_{200}$, and the *concentration index*

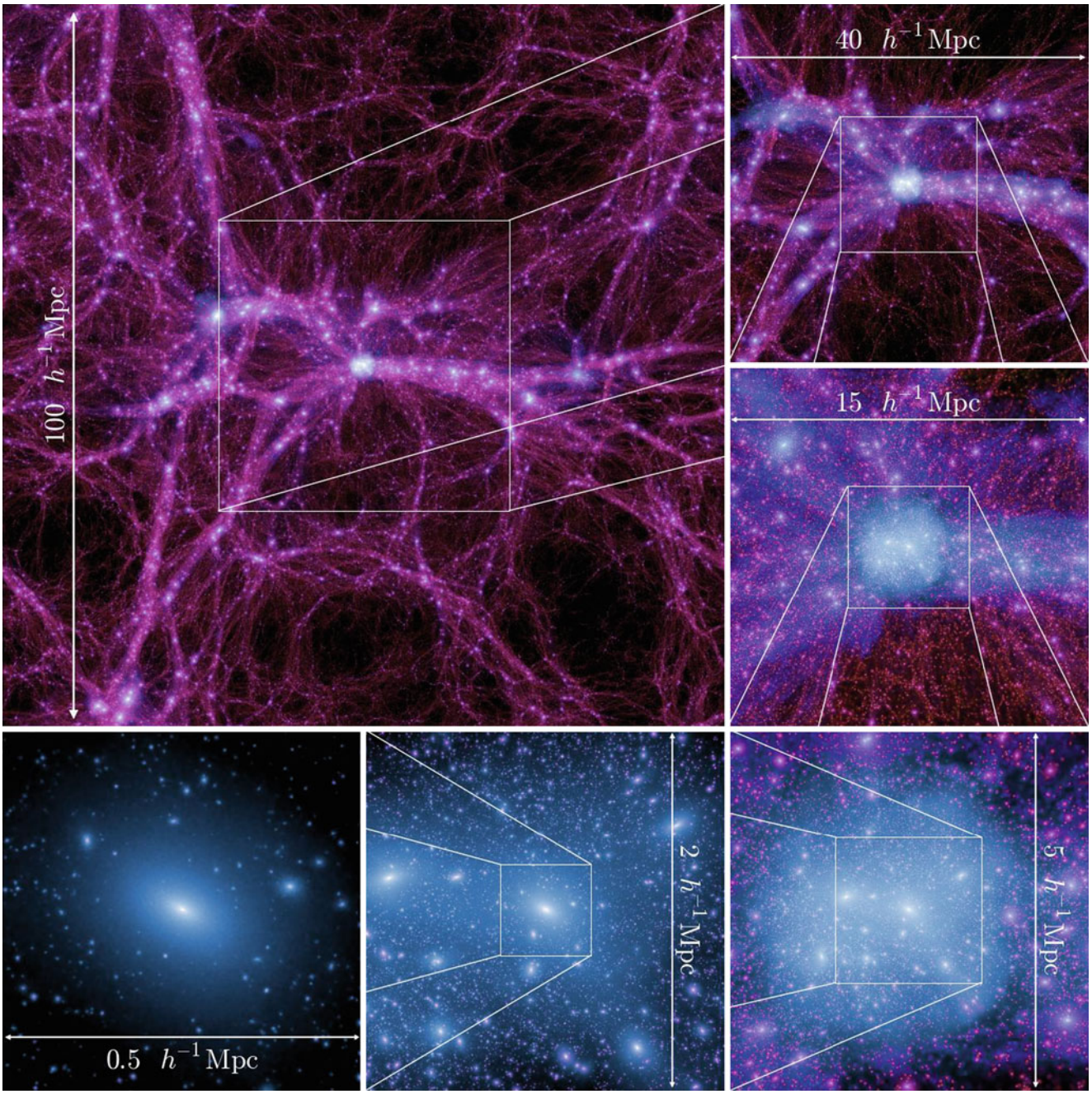

Fig. 7.15 The Millennium-II simulation. The *large upper-left panel* shows a $15h^{-1}$ Mpc slice through the full simulation, while the other *panels* display subsequent zooms of the central region, where the most massive halo in the simulation is located, with decreasing thickness of the slices. Source: M. Boylan-Kolchin et al. 2009, *Resolving cosmic structure formation with the Millennium-II Simulation*, MNRAS 398, 1150, p. 1153, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

$$\boxed{c := \frac{r_{200}}{r_s}} \tag{7.61}$$

was defined. The larger the value of c, the more strongly the mass is concentrated towards the inner regions. Equation (7.60) implies that ρ_s can be expressed in terms of $\rho_{cr}(z)$ and c, and performing the integration in (7.60) yields

$$\rho_s = \frac{200}{3}\rho_{cr}(z)\,\frac{c^3}{\ln(1+c) - c/(1+c)}\,.$$

Since M is determined by r_{200}, the NFW profile is parametrized by r_{200} (or by the mass of the halo) and by the concentration c that describes the shape of the distribution. Simulations show that the concentration index c is strongly

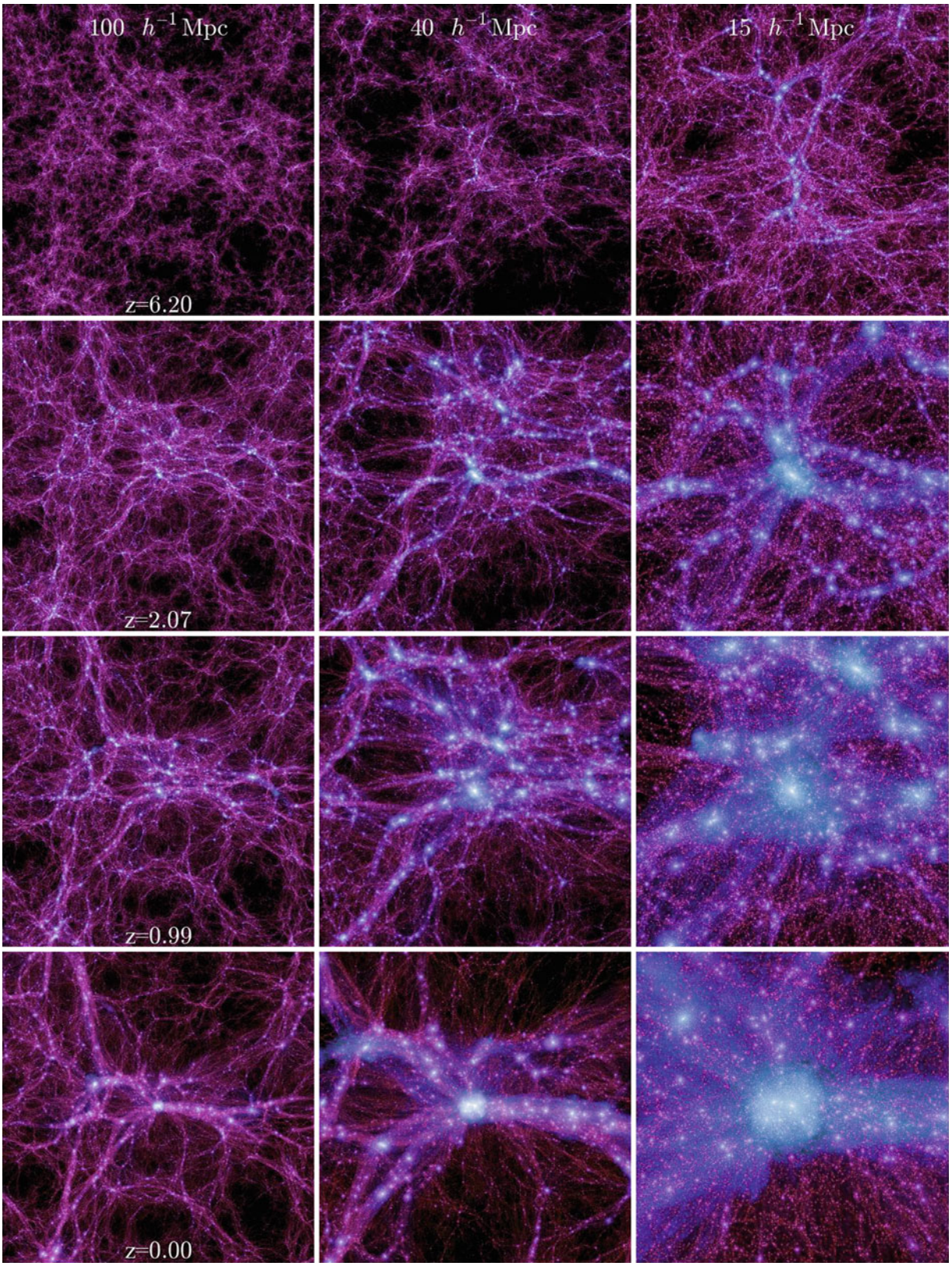

Fig. 7.16 Time evolution in the Millennium-II simulation. The most massive halo in the simulation is shown at four different redshifts, and three spatial resolutions. The thickness of the slices from left to right is 15, 10, and $6h^{-1}$ Mpc, respectively. Source: M. Boylan-Kolchin et al. 2009, *Resolving cosmic structure formation with the Millennium-II Simulation*, MNRAS 398, 1150, p. 1155, Fig. 2. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

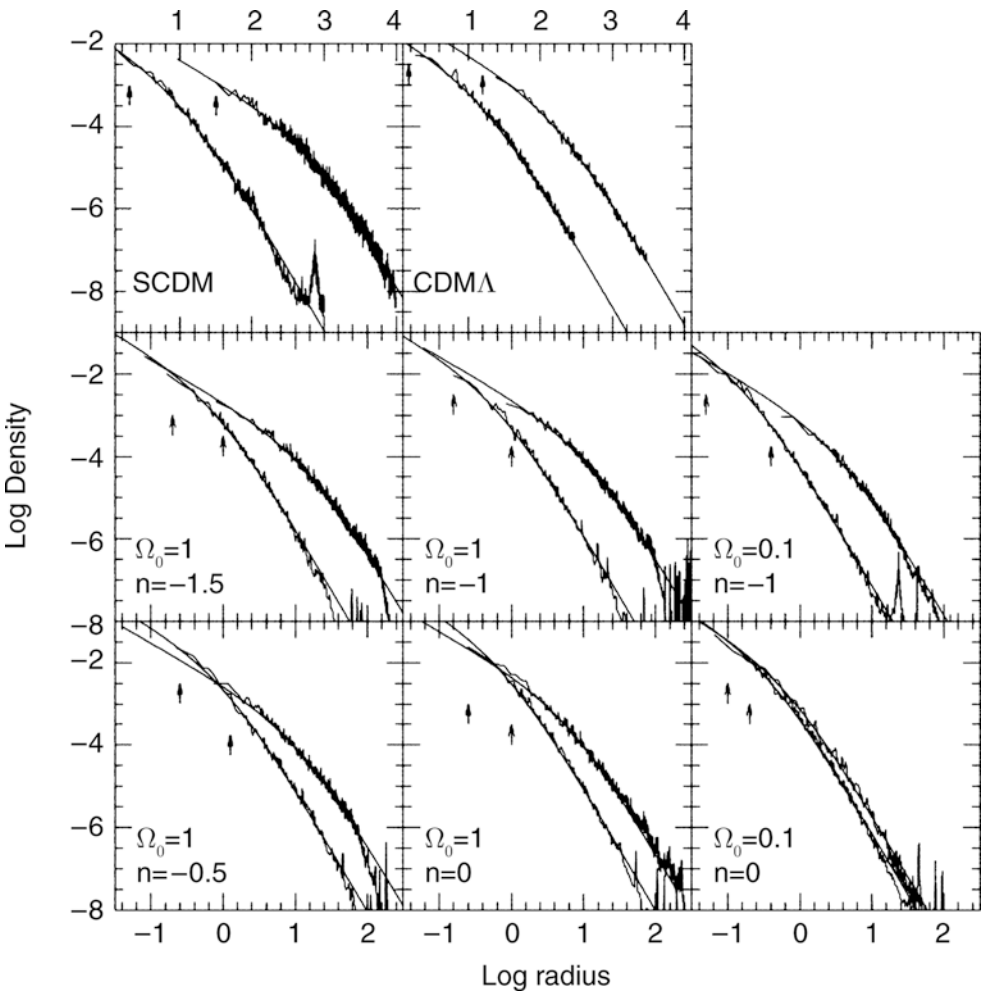

Fig. 7.17 For eight different cosmological simulations, the density profile is shown for the most massive and the least massive halo, each as a function of the radius, together with the best fitting density profile (7.59). The cosmological models represent an EdS model (here denoted by SCDM, *top left*), a ΛCDM model (*top right*), and different models with power spectra that are assumed to be power laws locally, $P(k) \propto k^n$. The *arrows* indicate the softening length in the gravitational force for the respective halos; thus, the major part of the profiles is numerically well resolved. Source: J.F. Navarro et al. 1997, *A Universal Density Profile from Hierarchical Clustering*, ApJ 490, 493, p. 496, Fig. 2. ©AAS. Reproduced with permission

correlated with the mass and the redshift of the halo; one finds approximately

$$c \approx 6.7 \left(\frac{M}{2 \times 10^{12} h^{-1} M_\odot} \right)^{-0.1} (1+z)^{-0.5} \qquad (7.62)$$

for relaxed halos. A similar result can also be obtained from analytical scaling arguments, under the assumption of the existence of a universal density profile. In Fig. 7.18, the density profile of dark matter halos is plotted as a function of the scaled radius r/r_{200}, where the similarity in the profile shapes for the different simulations becomes clearly visible, as well as the dependence of the concentration index on the halo mass. The range over which the density distribution of numerically simulated halos is described by the profile (7.59) is bounded above by the virial radius r_{200}, whereas in the central region of halos the numerical resolution of the simulations is too low to test (7.59) for very small r. The latter comment concerns the inner $\sim 1\,\%$ of the halo mass.

Generalization. No good analytical argument has yet been found for the existence of such a universal density profile, in particular not for the specific functional form of the NFW profile. As a matter of fact, other numerical simulations found slightly different density profiles, in particular towards the center. The reason for the differences between different simulations may be related to resolution issues. More recent numerical results have established a slight deviations of the mean halo profile from the NFW law, showing that a better fit is provided by the so-called Einasto profile,

$$\rho(r) = \rho_s \exp\left(\frac{-2}{\alpha} \left[\left(\frac{r}{r_s} \right)^\alpha - 1 \right] \right) , \qquad (7.63)$$

where r_s is a scale radius, ρ_s is the density at the scale radius, and α determines the overall shape; to good accuracy, $\alpha \sim 0.17$, though its value depends somewhat on the halo mass. The slope of an Einasto profile is a power law in radius, since

$$\frac{\mathrm{d} \ln \rho}{\mathrm{d} \ln r} = -2 \left(\frac{r}{r_s} \right)^\alpha . \qquad (7.64)$$

Thus, at the scale radius, the slope is -2, and it gradually decreases towards the center. This profile is not truly cuspy, since the slope approaches zero as $r \to 0$, however, due to the smallness of α, it does so very slowly.

Comparison with observations. The comparison of these theoretical profiles with an observed density distribution is

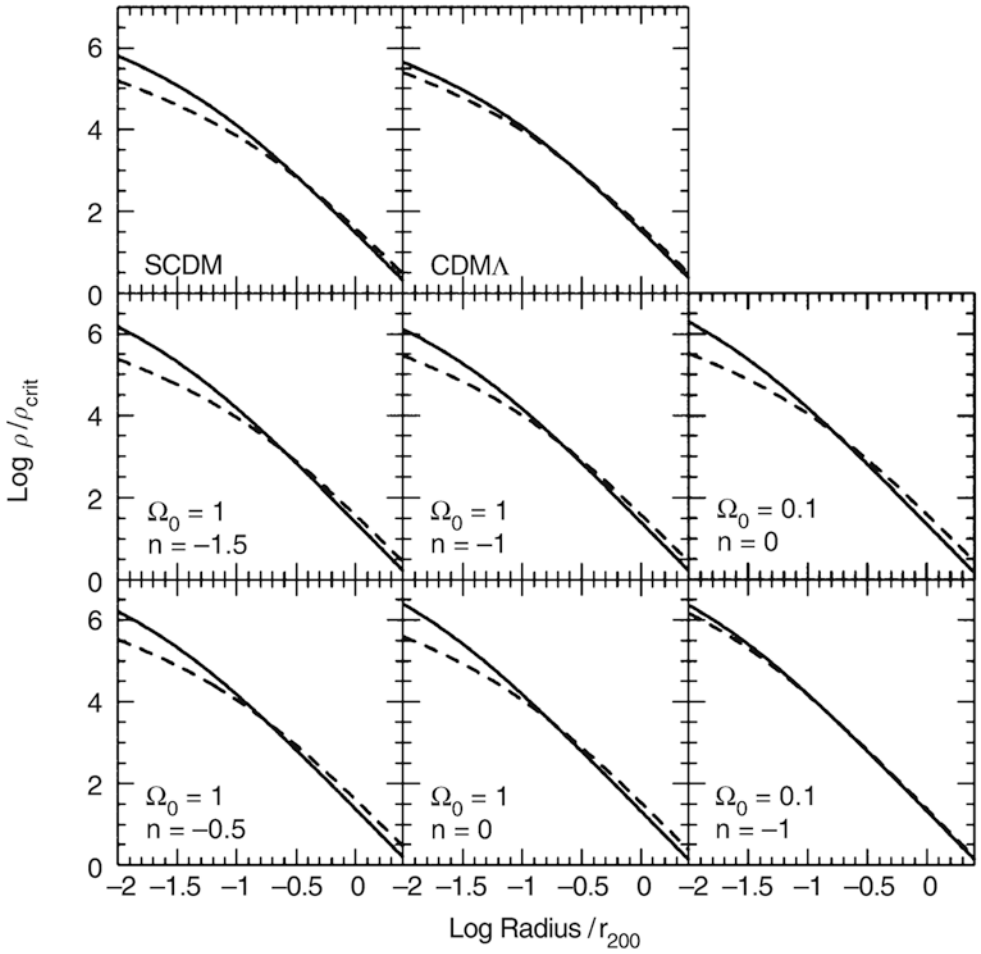

Fig. 7.18 The density profiles from Fig. 7.17, but now the density is scaled by the critical density, and the radius scaled by r_{200}. *Solid* (*dashed*) *curves* correspond to halos of low (high) mass—thus, halos of low mass are relatively denser close to the center, and they have a higher concentration index c. Source: J.F. Navarro et al. 1997, *A Universal Density Profile from Hierarchical Clustering*, ApJ 490, 493, p. 497, Fig. 3. ©AAS. Reproduced with permission

by no means simple because the density profile of dark matter is of course not directly observable. For instance, in normal spiral galaxies, $\rho(r)$ is dominated by baryonic matter at small radii. In the Milky Way, roughly half of the matter within R_0 consists of stars and gas, so that only little information is provided on $\rho_{\rm DM}$ in the central region. It is often assumed that galaxies with very low surface brightness (LSBs) are dominated by dark matter well into the center. The rotation curves of LSB galaxies are apparently *not* in agreement with the expectations from the NFW model shown in Fig. 7.19; in particular, they provide no evidence of a cusp in the central density distribution ($\rho \to \infty$ for $r \to 0$).

Part of this discrepancy may perhaps be explained by the finite angular resolution of the 21 cm line measurements of the rotation curves; however, the discrepancy remains if higher-resolution rotation curves are measured using optical long-slit and integral-field spectroscopy. As an additional point, the kinematics of these galaxies may be more complicated, and in some cases their dynamical center is difficult to determine. The orbits of stars and gas in these galaxies may show a more complex behavior than expected from a smooth density profile. The mass distribution in (the inner parts of) a dark matter halo is neither smooth nor axially symmetric, and stars and gas do not move on circular orbits in a thin plane of symmetry. Instead, simulations show that the pressure support of the gas, together with non-circular motions and projection effects systematically underestimate the rotational velocity in the center of dark matter halos, thereby creating the impression of a constant density core. Nevertheless, the observed rotation curves of LSB galaxies may prove to be a major problem for the CDM model—hence, this potential discrepancy must be resolved.

An additional complication is the fact that not only is baryonic matter present in the inner regions of galaxies (and clusters), thus contributing to the density, but also these baryons have modified the density profile of dark matter halos in the course of cosmic evolution. Baryons are dissipative, they can cool, form a disk, and accrete inwards. Also the opposite can happen: the explosion of supernovae can push some of the gas to larger radii, or even drive it out of the halo, in particular for low-mass ones. The changes in the resulting density distribution of baryons by dissipative processes cause a change of the gravitational potential over time, to which dark matter also reacts. The dark matter profile in real galaxies is thus modified compared to pure dark matter simulations.

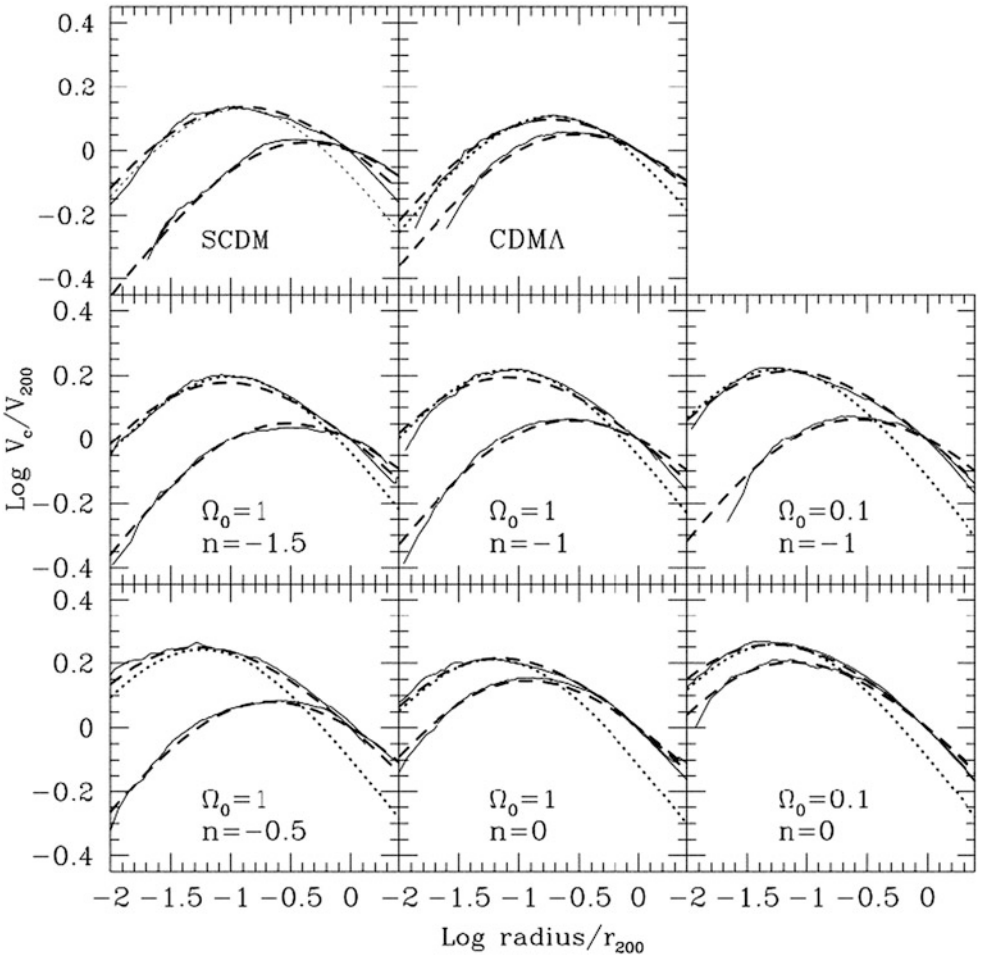

Fig. 7.19 The rotation curves in the NFW density profiles from Fig. 7.17, in units of the rotational velocity at r_{200}. All curves initially increase, reach a maximum, and then decrease again; over a fairly wide range in radius, the rotation curves are approximately flat. The *solid curves* are taken directly from the simulation, while *dashed curves* indicate the rotation curves expected from the NFW profile. The *dotted curve* in each panel presents a fit to the low-mass halo data with the so-called Hernquist profile, a mass distribution frequently used in modeling—it fits the rotation curve very well in the inner part of the halo, but fails beyond $\sim 0.1 r_{200}$. In these scaled units, halos of low mass have a relatively higher maximum rotational velocity. Source: J.F. Navarro et al. 1997, *A Universal Density Profile from Hierarchical Clustering*, ApJ 490, 493, p. 498, Fig. 4. ©AAS. Reproduced with permission

For galaxy clusters, the situation is more favorable, since the mass fraction of stars in them is smaller than in galaxies, and the stars are less concentrated towards the center. Indeed, it has been found that the X-ray data of many clusters are compatible with an NFW profile. Analyses based on the gravitational lensing effect also show that an NFW mass profile provides a very good description for the strong and weak lensing data; we will elaborate on this in Sect. 7.7 below. Additionally, Fig. 7.20 shows that the radial profile of the galaxy density in clusters on average follows an NFW profile, where the mean concentration index is $c \approx 3$, i.e., smaller than expected for the *mass* profile of clusters. One interpretation of this result is that the galaxy distribution in clusters is less strongly concentrated than the density of dark matter.

7.6.2 The shape and spin of halos

Halo shapes. Whereas the spherical collapse model made the simplifying assumption that the overdense regions are spherical, there is no reason for halos to have that symmetry. In fact, if one considers the peaks of a random density field, these maxima in general have different curvature along different directions. Correspondingly, the resulting halos are expected to deviate from sphericity.

Approximating the surfaces of constant density in a halo by an ellipsoid of semi-axes $a_1 \leq a_2 \leq a_3$, the shape of a halo is characterized by the ratios $s = a_1/a_3$ and $q = a_2/a_3$. If $s = q < 1$, the shape is similar to that of a cigar, and the ellipsoid is called 'prolate'. A halo with $s < 1$ and $q = 1$ has the shape of a hamburger, termed 'oblate'. In general, all three axes are different; such objects are called triaxial.

Numerical simulations show that the shape of dark matter halos depends strongly on their formation time and their merger history. If two halos of comparable mass collide and merge, the shape of the resulting halo will be strongly prolate. On the other hand, halos that form early and experience no such strong mergers tend to be more spherical.

Angular momentum of halos. During their evolution, dark matter halos can obtain a finite angular momentum, which can be measured from the output of simulations. The origin of this angular momentum can be traced back to tidal torques.

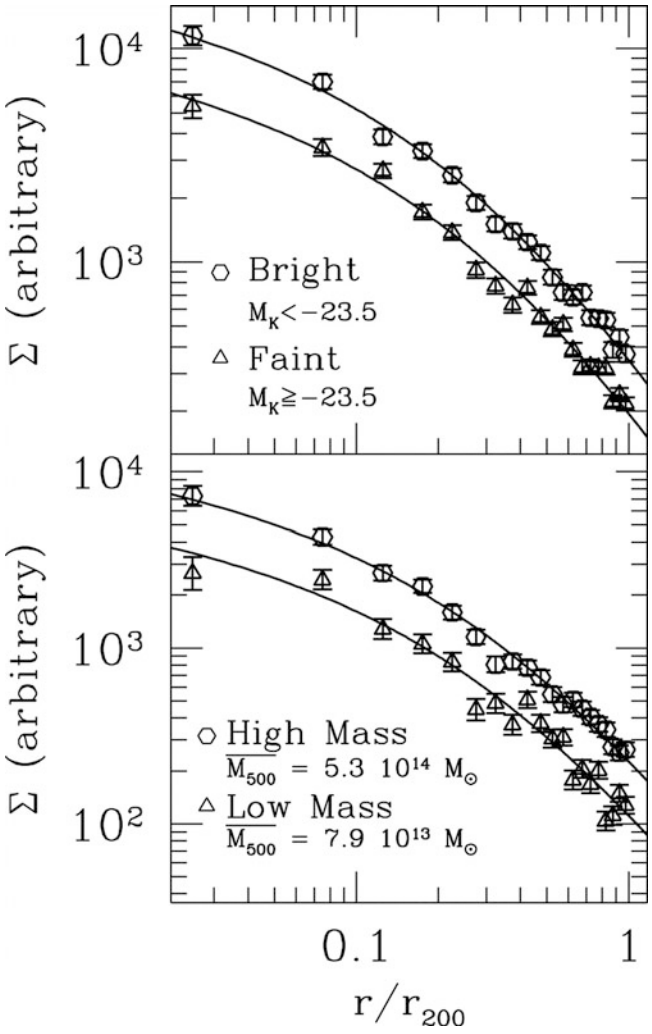

Fig. 7.20 The galaxy distribution averaged over 93 nearby clusters of galaxies, as a function of the projected distance to the cluster center. Galaxies were selected in the NIR, and cluster masses, and thus r_{200}, were determined from X-ray data. Plotted is the projected number density of cluster galaxies, averaged over the various clusters, versus the scaled radius r/r_{200}. In the *top panel* the galaxy sample is split into luminous and less luminous galaxies, while in the *bottom panel* the cluster sample is split according to the cluster mass. The *solid curves* show a fit of the projected NFW profile, which turns out to be an excellent description in all cases. The concentration index is, with $c \approx 3$, roughly the same in all cases, and somewhat smaller than expected for the mass profile of clusters. Source: Y.-T. Lin et al. 2004, *K-Band Properties of Galaxy Clusters and Groups: Luminosity Function, Radial Distribution, and Halo Occupation Number*, ApJ 610, 745, p. 756, Fig. 8. ©AAS. Reproduced with permission

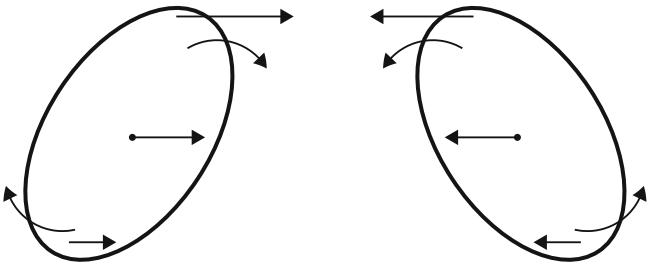

Fig. 7.21 Sketch of two non-spherical density concentrations and their mutual gravitational attraction. Points located closer to the neighboring overdensity feel a stronger force than the center of mass, leading to the tidal torque and a rotation of the body. This is how mass concentrations attain their angular momentum

As we have seen, halos are not spherical in general. If an ellipsoidal body is located in a tidal gravitational field, a torque acts on it, trying to align the body with the direction of the tidal field (see Fig. 7.21). This causes the body to rotate, yielding an angular momentum.

Usually, the angular momentum of a halo is quantified by the so-called spin parameter λ. To motivate its definition, consider a rigidly rotating sphere of radius R, angular velocity ω, and mass M. The sphere has a gravitational binding energy of

$$|E| \sim \frac{G\,M^2}{R}$$

and an angular momentum of

$$J \sim M\,R^2\,\omega\,,$$

with the constants of proportionality dependent on the density distribution inside the sphere. In order for the sphere to be rotationally supported, the gravitational acceleration on its surface should be balanced by the centrifugal acceleration, so that

$$\omega^2\,R \sim \frac{G\,M}{R^2}\,, \quad \text{or}$$

$$J \sim M\,R^2\,\omega \sim M\,R^2\sqrt{\frac{G\,M}{R^3}}$$

$$\sim M^{5/2}\,G\,|E|^{-1/2}\,,$$

where we have expressed R in terms of $|E|$. Hence, one defines the dimensionless *spin parameter*

$$\boxed{\lambda := \frac{J\,|E|^{1/2}}{G\,M^{5/2}}}\,. \tag{7.65}$$

For plausible density profiles, one finds that $\lambda \sim 0.4$ corresponds to rotational support.

The spin parameter of halos measured in simulations is typically an order of magnitude smaller than required for rotational support. This shows that the deviation from sphericity is *not* due to their rotation, but by the distribution of orbits of the dark matter particles. More quantitatively, one finds that the spin parameter of halos has a probability distribution of the form

$$p(\lambda)\,\mathrm{d}\lambda = \frac{1}{\sigma_\lambda\sqrt{2\pi}}\exp\left(-\frac{\ln^2(\lambda/\bar{\lambda})}{2\sigma_\lambda^2}\right)\frac{\mathrm{d}\lambda}{\lambda}\,,$$

with $\bar{\lambda} \sim 0.04$ and $\sigma_\lambda \sim 0.5$. Furthermore, there is the tendency that halos in denser environments have larger than average values of λ, as is expected from the above argument: on average, the tidal gravitational field is stronger in overdense regions.

In the early stages of halo formation, the baryons have about the same spatial distribution as that of the dark matter. Therefore, the gas in dark matter halos will attain a similar specific angular momentum as the halo itself. This angular momentum will turn out to be a key element for the formation of galaxies, as will be discussed in Chap. 10.

7.6.3 The bias of dark matter halos

Since dark matter halos host the observable tracers of the Universe, i.e., galaxies and groups and clusters of galaxies, it is interesting to study the properties of their spatial distribution and to relate this to the observable spatial distribution of galaxies and clusters. We shall consider this latter aspect in Sect. 8.1, but describe here the clustering properties of halos.

Halo number density contrast. We start by considering the density field in the universe, smoothed over spheres of radius R. As before, we call the smoothed density field $\delta_R(\boldsymbol{x})$. If R is large, any such sphere will contain many halos. Furthermore, we consider dark matter halos with mass between M and $M + \mathrm{d}M$. Their mean number density is $\bar{n} = \frac{\mathrm{d}n}{\mathrm{d}M}\mathrm{d}M$, as given by the halo abundance discussed before. Let the number of halos within a sphere of radius R centered on $\boldsymbol{x}$ be $N(\boldsymbol{x};M)\mathrm{d}M$; then we can define the local density of halos as $n(\boldsymbol{x};M)\mathrm{d}M = N(\boldsymbol{x};M)\mathrm{d}M/V$, where $V = (4\pi/3)R^3$ is the volume of the sphere. In analogy to the definition of the density contrast of matter, we can consider the relative density contrast of halos,

$$\delta_\mathrm{h}(\boldsymbol{x};M) = \frac{n(\boldsymbol{x};M) - \bar{n}(M)}{\bar{n}(M)}\,. \qquad (7.66)$$

We can now ask whether the two density fields, $\delta_R(\boldsymbol{x})$ and $\delta_\mathrm{h}(\boldsymbol{x};M)$ are similar. For example, suppose that $\delta_R(\boldsymbol{x}) = 1$ at one point, i.e., the matter density there is twice the cosmic mean. Does one expect to find also twice as many halos in the sphere surrounding $\boldsymbol{x}$ as one finds on average in a sphere of this radius?

Halo biasing. In general, one expects the number density of halos to be large in those regions of space where the matter density is also high. However, we have no good reason to assume that the number density of halos follows exactly the matter distribution. In fact, we can argue that in general, these two distributions should be different: Consider the density fluctuations of the matter to be divided into large- and small-scale fluctuations (see Fig. 7.22). The spherical collapse model predicts that a halo forms when the linear density contrast exceeds $\delta_\mathrm{c} = 1.69$. If we now use the decomposition $\delta(\boldsymbol{x}) = \delta_\mathrm{s}(\boldsymbol{x}) + \delta_\mathrm{l}(\boldsymbol{x})$, where the 's' and 'l' stand for small- and large-scale fluctuations, then in regions of positive δ_l, a halo can collapse even if δ_s is less than the threshold value δ_c, whereas in underdense regions with $\delta_\mathrm{l} < 0$, the small-scale fluctuations must reach a higher value than δ_c for the collapse to occur, namely $\delta_\mathrm{s} > \delta_c - \delta_\mathrm{l}$. In other words, in regions of overdense large-scale fluctuations, the smaller-scale fluctuations get a head-start for their gravitational collapse. As a matter of fact, this so-called peak-background split of the density fluctuation field can be used to analytically predict the relation between δ_h and δ.

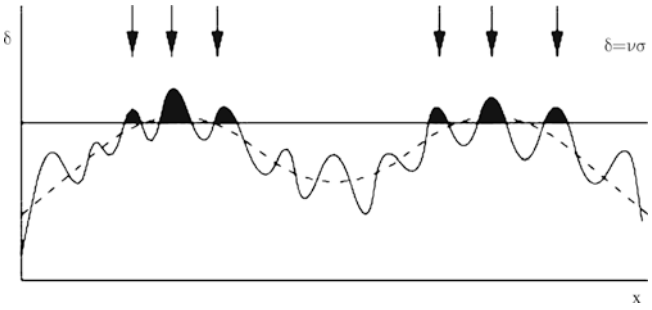

Fig. 7.22 The sketch represents a particular model of biasing. Let the one-dimensional density profile of matter be specified by the *solid curve*, which results from a superposition of a large-scale (represented by the *dashed curve*) and a small-scale fluctuation. Assuming that halos can form only at locations where the density field exceeds a certain threshold—plotted as a *straight line*—the halos in this density profile will be localized at the positions indicated by the *arrows*. Obviously, the locations of the halos are highly correlated; they only form near the peaks of the large-scale fluctuation. In this picture, the correlation of halos on small scales is much stronger than the correlation of the underlying density field. Source: J.A. Peacock 2003, *Large-scale surveys and cosmic structure*, astro-ph/0309240, Fig. 8

The fact that in general $\delta_\mathrm{h}(\boldsymbol{x};M) \neq \delta(\boldsymbol{x})$ is called halo biasing. The most simple way this biasing could be modeled is by assuming that these two density fluctuation fields are simply proportional to each other, i.e.,

$$\delta_\mathrm{h}(\boldsymbol{x};M) = b_\mathrm{h}(M)\,\delta(\boldsymbol{x})\,, \qquad (7.67)$$

where $b_\mathrm{h}(M)$ is called the *halo bias factor*. The ansatz (7.67) is called linear deterministic biasing. Whereas it cannot be valid in detail, this ansatz provides a useful description for large scales, i.e., when R is chosen sufficiently large. In particular, averaged over scales larger than $\sim 10h^{-1}\,\mathrm{Mpc}$, (7.67) is rather well satisfied, whereas on smaller scales, it ceases to be valid.

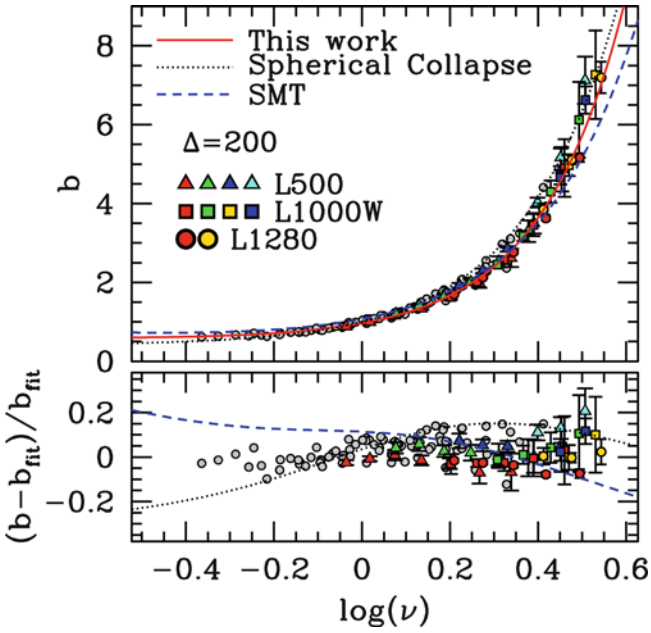

Fig. 7.23 *Top panel*: The large-scale bias of dark matter halos, as determined from a set of three different simulations (indicated by the three different types of symbols), as a function of the 'peak height' ν [see (7.51)]. For each type of symbol, different *colors* indicate different redshifts at which the bias was determined. The three *curves* show halo bias as obtained in the framework of Press–Schechter theory (*black dotted*) and the models of spheroidal collapse (*blue dashed*), as well as a fitting function to the data (*red solid*). The *bottom panel* shows the relative deviation of the measured halo bias from the fit, indicating its accuracy. Source: J.L. Tinker et al. 2010, *The Large-scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests*, ApJ 724, 878, p. 880, Fig. 1. ©AAS. Reproduced with permission

Dependence on M and z. As mentioned before, using the peak-background split, the bias factor can be estimated analytically, and can also be determined from N-body simulations; these two estimates mutually agree rather well (see Fig. 7.23). One finds that $b_{\rm h}(M)$ is a monotonically increasing function of halo mass, with $b_{\rm h} < 1$ for $M \lesssim M^*(z)$ [or $\nu \lesssim 1$, where the 'peak height' ν is given in (7.51)] with $M^*(z)$ given by (7.53), and $b_{\rm h}(M) > 1$ for $M \gtrsim M^*(z)$ ($\nu \gtrsim 1$). In fact, as was the case for the halo abundance, the halo bias essentially depends solely on the value of ν. Towards low masses, the halo bias approaches a constant value ~ 0.7, and it rather steeply increases beyond M^*, reaching values of a few on the largest mass scales. Furthermore, at fixed mass, the halo bias increases with redshift, since according to (7.51), ν increases with redshift for fixed M.

These qualitative properties can be easily understood in terms of the peak-background split picture: the little head-start for the collapse of a peak in a region of large-scale overdensity is the more important, the rarer peaks of amplitude $> \delta_{\rm c}$ are. Hence, at rather high masses, this little additional push may actually be necessary for such halos to form at all. Conversely, for smaller masses, the abundance of peaks is high, and their number density with amplitude $\delta_{\rm c} + \epsilon$ is rather the same as that with $\delta_{\rm c} - \epsilon$. Hence, for them the impact of the large-scale fluctuation is negligible. As a general rule, the rarer density peaks are, the larger is the bias.

As an immediate consequence, (7.67) implies that the correlation function of halos is different from the correlation function of the matter distribution. Since the correlation function is quadratic in the density field, we readily find

$$\xi_{\rm h}(\boldsymbol{y}; M) := \langle \delta_{\rm h}(\boldsymbol{x})\, \delta_{\rm h}(\boldsymbol{x}+\boldsymbol{y}) \rangle = b_{\rm h}^2(M)\, \xi(\boldsymbol{y})\,. \quad (7.68)$$

Thus, massive halos with $b_{\rm h} > 1$ are more strongly clustered than the underlying matter distribution, whereas low-mass halos are clustered less. One therefore expects that galaxies are less clustered than galaxy clusters; we will see later that this is indeed the case. A similar expression applies of course also to the power spectrum. Indeed, the halo bias in the simulations (Fig. 7.23) was measured from the ratio of the power spectra of halos and the overall matter distribution, restricted to the largest scales.

7.7 Weak gravitational lensing studies of dark matter halos

Whereas rotation curves probe the inner part of galaxy halos—typically out to a radius not larger than 1/10 of the virial radius—obtaining information for the mass profile at larger radii is difficult, due to the lack of luminous tracers. Weak gravitational lensing (see Sect. 6.6.2) offers the possibility to study the mass profile out to very large radii. In fact, it is by far the most powerful method for probing the outer parts of (galaxy-mass) halos. Combined with strong lensing at smaller radii, the halo mass profile of massive clusters can be studied over a broad range in radii (Sect. 7.7.1).

Apart from very massive clusters, the weak lensing signal of individual objects is not strong enough to be studied individually. The ellipticity dispersion of the faint galaxies which act as background sources for the lensing effect provides a noise component which is too large compared to the lensing signal. However, one can combine the weak lensing signal of a large number of lensing galaxies, and thus study their mean mass profiles. Provided the area of the imaging survey is large enough, the lens galaxies can be grouped into samples of similar properties (like redshift, color, luminosity, etc.), and thus the average mass profiles of galaxies can be investigated in dependence on these parameters.

In this section, we describe the basic method of this *galaxy-galaxy lensing (GGL)* technique and present some of the recent results (Sect. 7.7.2). We then discuss a method

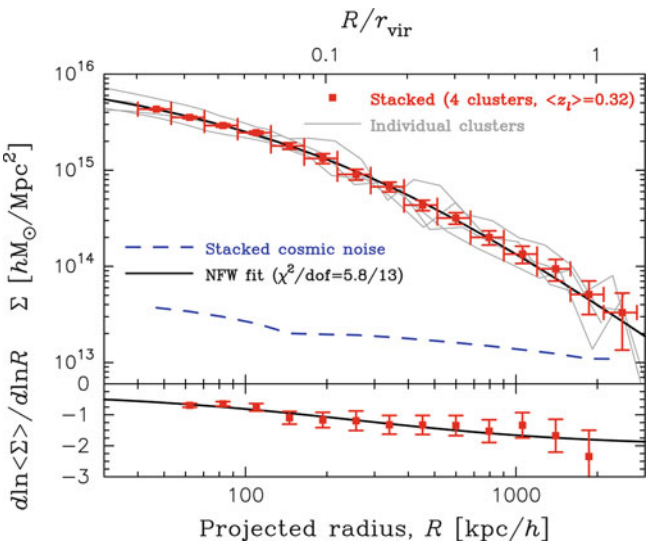

Fig. 7.24 The average surface mass density profile (*red points* with error bars) obtained from a strong and weak lensing analysis of four massive clusters of galaxies for which very high quality data are available. The four individual mass profiles are shown as *thin grey curves*. The *thick solid curve* is the best fitting line-of-sight projected NFW-profile to the mean mass profile; it provides an exceptionally good fit. The *blue dashed curve* shows the noise of the mean mass profile that is expected from the lensing effects of the large-scale structure between us and the clusters, and between the clusters and the source population. The *bottom panel* shows the slope of the surface mass density as a function of radius. There is a continuous steepening towards larger radii, again compatible with the NFW-profile, shown as solid curve. Source: K. Umetsu et al. 2011, *A Precise Cluster Mass Profile Averaged from the Highest-quality Lensing Data*, ApJ 738, 41, p. 6, Fig. 1. ©AAS. Reproduced with permission

to interpret them, introducing the so-called *halo model* in Sect. 7.7.3. Finally, we generalize this technique to the statistical study of the mass distribution of galaxy groups in Sect. 7.7.4.

7.7.1 Massive clusters

Very massive clusters show a sufficiently strong lensing signal for their mass distribution to be studied individually. In order to get information from a wide range in radii, the best results are obtained from combining strong lensing (multiple images and arcs) in the inner region with weak lensing for large radii.

Figure 7.24 shows results for four strong lensing clusters, supplemented by weak lensing information for large R. It is seen that the resulting mass profile is very well fit with an NFW-profile, a result also obtained by other studies. Therefore, it appears that lensing studies of clusters support the prediction of the CDM model for the existence of a universal mass profile.

Whereas the lensing data are well fit by the functional form of the NFW-profile, the resulting concentration parameters are found to be larger than the CDM prediction (7.62), even if the predicted spread of the c-M-relation is taken into account. It thus appears that strong lensing clusters are 'over-concentrated'. However, there are severe selection effects at work. First, the more concentrated a mass distribution is, the more likely it is that it produces giant arcs and multiple images (because the area over which the surface mass density exceeds the critical surface mass density for lensing is then larger). Thus, a strong lensing selection favors halos which have a higher-than-average concentration. Second, there are projection effects. The c-M-relation is obtained by considering the spherically-averaged mass distribution of halos. Since halos are triaxial in general, the concentration fitted to the projected mass profile will depend on the projection direction. Geometrically it is obvious that the largest concentration is obtained if the projection occurs along the direction of the largest axis, which then also maximizes the projected mass density, and hence the strong lensing strength. Thus, again, this leads to a selection effect for strong lensing clusters which biases the concentration to high values. Finally, the strong lensing probability increases substantially when a cluster is in the process of a merger; the resulting asymmetry of the mass distribution renders the occurrence of spectacular strong lensing features particularly likely. Indeed, we have seen in Sect. 6.6.1 that many of the famous giant arc clusters show a bimodal distribution. These clusters may therefore be extreme outliers in the c-M-relation. For these reasons, the 'over-concentration' is not regarded to be a serious issue for CDM models.

7.7.2 Galaxy-galaxy lensing

As we discussed in Sect. 6.6.2, the tidal component of the gravitational field causes a distortion of the observed shape of distant galaxies. This distortion is such that for axi-symmetric matter distributions, images are stretched in a direction tangent to the center of mass (see Fig. 6.53).

The shear. In the language of gravitation lensing, this distortion—or the tidal components of the deflection—is quantified by the *shear*. It is symbolized by the sticks in Fig. 6.53. The shear is linearly related to the surface mass density κ of the lens.

From the equations of gravitational lensing, one can show for an axi-symmetric mass distribution with dimensionless surface mass density $\kappa(\theta)$ that the shear $\gamma(\theta)$ is

$$\gamma(\theta) = \bar{\kappa}(\theta) - \kappa(\theta) \,, \tag{7.69}$$

where $\bar{\kappa}(\theta)$ is the mean surface mass density inside a circle of radius θ; it is related to the dimensionless mass $m(\theta)$ [see (3.70)] by $\bar{\kappa}(\theta) = m(\theta)/\theta^2$. Remarkably, the relation (7.69) is also valid for arbitrary mass distributions, if we interpret $\gamma(\theta)$ as the tangential component of the shear averaged over the circle of radius θ, and $\kappa(\theta)$ to be the mean surface mass density on that circle.

The principle of galaxy-galaxy lensing. The observed ellipticity of the image of a background source is the sum of the intrinsic ellipticity and the shear. Since the intrinsic orientation of galaxies has no preferred direction, the intrinsic ellipticity will have a mean of zero if we average over enough background galaxies.

Thus, consider a set of (foreground) galaxies, together with a population of (background) galaxies for which their ellipticities have been measured. If one then considers all foreground-background pairs within a small angular separation interval $\Delta\theta$ around θ, and measures the tangential component (relative to the center of the foreground galaxies) of the background ellipticities, the mean of this ellipticity provides an estimate of the mean shear of the foreground galaxies, since the intrinsic ellipticities of the background galaxies will average out to almost zero. In this way, the shear profile $\gamma(\theta)$ can be estimated. Since the shear is directly related to the mass density by (7.69), this shear profile determines the mass profile, up to an overall additive constant (which is related to the mass-sheet degeneracy; see problem 3.5).

If the redshifts of the foreground galaxies are individually known, the angular separation between foreground and background galaxy can be translated into a transverse separation, $R = D_{\rm d}\theta$. Furthermore, if the redshift distribution of the background galaxies are known as well, then the mean of the distance ratio $D_{\rm ds}/D_{\rm s}$ can be calculated. Therefore, with these two pieces of information, the critical surface mass density $\Sigma_{\rm cr}$ [see (3.67)] can be determined. Multiplying (7.69) by $\Sigma_{\rm cr}$ then yields

$$\Sigma_{\rm cr}\,\gamma(\theta) \equiv \Delta\Sigma(R) = \bar{\Sigma}(R) - \Sigma(R)\,, \qquad (7.70)$$

i.e., the observed shear profile can be directly related to the mean physical surface mass density inside the circle of radius R, minus the average surface mass density at radius R.

On small angular scales, the shear profile corresponds to the mass profile of the galaxy halos, and is well-fitted by either an SIS profile, or that of the universal halo density profile as described by the NFW model. Hence, GGL allows us to study the mean mass profiles of galaxy halos.

Selected results. In Fig. 7.25, we show the GGL signal $\gamma(\theta)$ as obtained from the RCS2 (the second Red Cluster Sequence) survey, where the bright foreground galaxies were identified by their spectroscopic redshifts as determined by SDSS. A clear signal is seen out to scales of ~ 1 deg. For the upper axis in this figure, the angular scale was converted to a transverse separation, using the mean distance of the lens galaxy sample. Thus we see that the measured shear signal probes the mean mass profiles of the foreground galaxies out to $\sim$ 20 Mpc, i.e., far larger than the virial radius of galaxy-scale dark matter halos. For such large radii, we do not expect that the mass profile of galaxies is described by the universal mass profile (7.59) whose validity is restricted to within the virial radius.

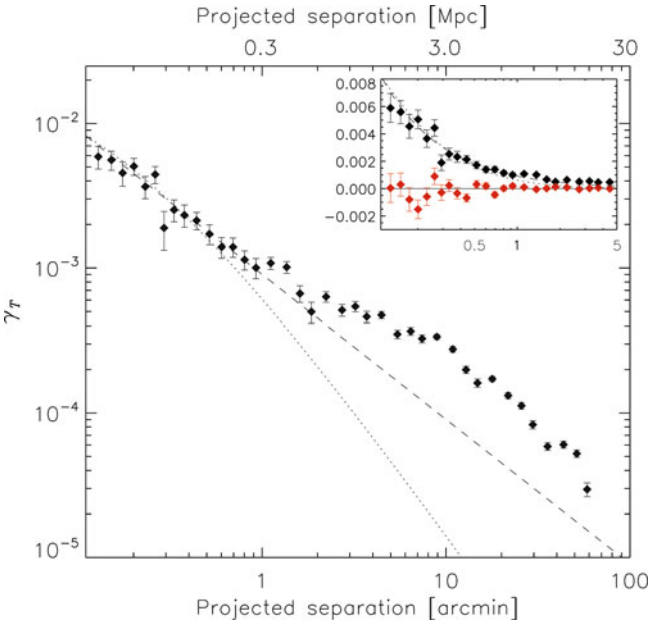

Fig. 7.25 The GGL signal, i.e., the mean tangential shear around foreground galaxies, as measured in the RCS2 survey, is shown by the *symbols* with error bars. The upper axis translates the angular scale into a transverse separation, for the mean distance of the foreground galaxy population. The *insert* displays a zoom of the central region, where the *red points* show the so-called cross component of the shear, whose average should be compatible with zero—which is seen to be the case. The two *curves* show the shear profile of the best-fitting SIS and NFW mass models. Source: E. van Uitert et al. 2011, *Galaxy-galaxy lensing constraints on the relation between baryons and dark matter in galaxies in the Red Sequence Cluster Survey 2*, A&A 534, A14, p. 6, Fig. 5. ©ESO. Reproduced with permission

The inner part of the GGL signal is fitted with an NFW profile (dotted curve) and an SIS model (dashed curve) in Fig. 7.25. In the inner $\sim$ 200 kpc, both mass models yield good fits, but fall short of explaining the shear signal at larger radii. We will see below how the extended shear profile can be interpreted.

Nevertheless, on small scales, the mean shear profile of galaxies probes the mass profile of the galaxy and its dark matter halo. Selecting galaxy samples with different luminosity or stellar mass, the dependence of the parameters describing the NFW profile—in particular the virial mass—on the stellar mass can be studied.

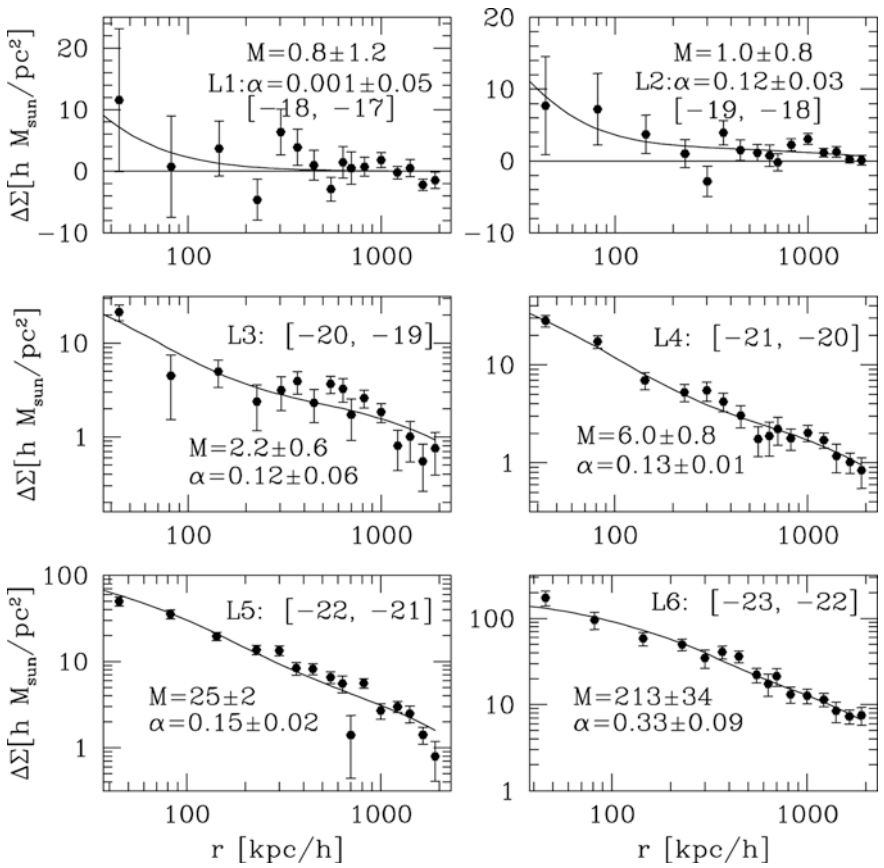

Fig. 7.26 The galaxy-galaxy lensing signal for six luminosity bins of foreground galaxies, as indicated by the absolute magnitude interval in each panel. More than 2.7×10^5 galaxies with spectroscopic redshifts were used as foreground galaxies in this analysis. The *curves* show a two-parameter model fitted to the data, based on the halo model, and the fit parameters are indicated: M is the virial mass of the halo (in units of $10^{11}h^{-1}M_\odot$) in which the galaxies reside, and α is the fraction of the galaxies which are not central inside the halo, but satellite galaxies. Source: U. Seljak et al. 2005, *Cosmological parameter analysis including SDSS Lyα forest and galaxy bias: Constraints on the primordial spectrum of fluctuations, neutrino mass, and dark energy*, Phys. Rev. D 71, 043511, Fig. 1. http://journals.aps.org./prd/abstract/10.1103/PhysRevD.71.043511. Published with kind permission ©APS 2005. All Rights Reserved

The combination of imaging and spectroscopy in the Sloan Digital Sky Survey makes this an ideal data set for studying GGL, since one can select foreground galaxies with known redshifts. Figure 7.26 shows the GGL signal $\Delta\Sigma(R)$ for six different bins of absolute magnitude. The shear signal is clearly detected out to large radii for the more luminous galaxy samples. The signal increases with luminosity, showing that the galaxy+halo mass is a monotonic function of galaxy luminosity, as expected. As was the case for the RCS2 results shown in Fig. 7.25, the GGL signal from the SDSS shown in Fig. 7.26 extends to separations much larger than the expected virial radius for galaxy-mass halos.

7.7.3 Interpretation: The halo model

The shortfall of the NFW mass profile to explain the observed GGL signal on scales larger than the virial radius can be understood from noting that galaxies (and their halos) are not isolated. We have seen in Sect. 7.6.3 that dark matter halos are correlated. Furthermore, many galaxies are members of galaxy groups or clusters.

Therefore, the mean mass profile around galaxies will be a superposition of several components. On small scales, it is dominated by the stellar mass of the galaxy and the dark matter halo in which it is embedded. On intermediate scales, one then starts to see the impact of the groups and clusters in which a certain fraction of the galaxies are embedded. On even larger scales, say $\gtrsim$1 Mpc, which exceed the size of most galaxy clusters, the signal becomes increasingly dominated by the mass from dark matter halos which are correlated with the host halo of the galaxy. Disentangling the various contributions of the GGL signal has to be done in the framework of a model.

Ingredients of the halo model. A very successful framework for the interpretation of the GGL signal is the halo model, which shall be briefly sketched here. It assumes that the all mass in the universe is contained in dark matter halos, so that the density distribution can be written as a sum of the density profiles of these halos,

$$\rho(\boldsymbol{x}) = \sum_i \rho_{\rm h}(|\boldsymbol{x} - \boldsymbol{x}_i|; M_i) \ , \qquad (7.71)$$

where the sum extends over all dark matter halos (in a given volume of space), $\boldsymbol{x}_i$ and M_i is the position and mass of the i-th halo, and $\rho_{\rm h}$ is the halo mass profile, which is assumed to be spherically symmetric (so that the mass contribution of the i-th halo at the location $\boldsymbol{x}$ depends only on the separation $|\boldsymbol{x} - \boldsymbol{x}_i|$ between the point $\boldsymbol{x}$ and the halo center $\boldsymbol{x}_i$). Fur-

thermore, by writing (7.71) we have assumed that the density profile of a halo is fully characterized by its mass M_i—e.g., that the density profile ρ_h is given by the NFW-profile with a concentration determined by its mass [see (7.62)]. One can account for the dispersion of the concentration parameter about its mean (7.62) by using c as a further argument of ρ_h. The halo population is characterized by the halo mass spectrum (Sect. 7.5.2) which, together with the halo correlation function (7.68) is obtained from numerical simulations.

The basic assumption of the halo model, namely that all the mass in the Universe is contained in halos, is fairly well supported by cosmological simulations. Down to a mass scale of $2 \times 10^8\, M_\odot$ (the scale that is resolved by the Millennium-II simulation), about 60 % of the total mass of the present Universe is contained in halos. Extrapolating to even lower halo mass, using analytic models (see Sect. 7.5.2), suggests that this fraction rises to some 80 % down to the smallest halo masses. Whereas the mass fraction contained in halos is lower at higher redshifts, these results provide a good motivation for the halo model.

The mass correlation function. Combining these ingredients, the correlation function of matter can be derived. If we want to calculate the correlator $\langle \rho(\boldsymbol{x})\, \rho(\boldsymbol{x}')\rangle$, we obtain a double sum,

$$\left\langle \rho(\boldsymbol{x})\, \rho(\boldsymbol{x}')\right\rangle = \sum_{ij} \left\langle \rho_\mathrm{h}(|\boldsymbol{x} - \boldsymbol{x}_i|; M_i)\, \rho_\mathrm{h}(|\boldsymbol{x}' - \boldsymbol{x}_j|; M_j)\right\rangle \,. \tag{7.72}$$

This double sum can then be split into a diagonal term (i.e., where $i = j$) and a non-diagonal one. For the former, the density of the halo is correlated with itself—we call this the one-halo term of the mass correlation function. The non-diagonal term correlates the mass in one halo with that of another halo, giving rise to the two-halo term in the correlation function. This latter term arises from the fact that the halo centers are correlated, according to (7.68).

Averaging the result (7.72) over the mass spectrum of halos, as well as over the probability distribution of halo positions $\boldsymbol{x}_i$, accounting for their mass-dependent correlation, one obtains the two-point correlation function of the matter distribution as predicted by the halo model. Indeed, the halo model yields a description of the matter distribution which appears to be an astonishingly good approximation to the more accurate results from simulations. As argued before, the matter correlation function is a sum of two terms, $\xi_\mathrm{m} = \xi_\mathrm{m}^\mathrm{1h} + \xi_\mathrm{m}^\mathrm{2h}$, where the first term describes matter correlations within the same halo, whereas the second is the correlation between two different halos.

Inclusion of galaxies. Galaxies are next introduced into this halo model. This is done by first noting that galaxies form at the center of dark matter halos. Satellite galaxies in halos, such as galaxies in clusters or the dwarf galaxies in the Milky Way, presumably also formed at the center of dark matter halos which subsequently merged with the larger halo.

Hence, one considers a population of galaxies with luminosities (or stellar masses) in a given interval. Let $\langle N|M\rangle$ be the mean number of galaxies (with the prescribed properties) that live in a halo of mass M. This function cannot be obtained from first principles, but can be constrained by observations of galaxies in groups and clusters whose masses are estimated by X-ray or weak lensing observations; additional constraints come from the luminosity function of galaxies. It is usually assumed that the number N of galaxies is Poisson distributed, with mean $\langle N|M\rangle$. Frequently, the mean number of galaxies is prescribed by a power law in mass, $\langle N|M\rangle \propto M^\epsilon$, above some mass threshold which depends on the luminosity (or stellar mass) of the galaxies under consideration. For lower-mass halos, $\langle N|M\rangle$ is assumed to decline rapidly—low-mass halos do not host luminous galaxies (e.g., one does not expect to find any L^* galaxy in a halo of mass $< 10^{11}\, M_\odot$).

If there is only one galaxy in a halo, it is assumed to have formed at the center, whereas if $N > 1$, one assumes that one of them lies at the center (the central galaxy), whereas the rest (satellite galaxies) are distributed according to a specific radial distribution function; typically, they are assumed to also follow an NFW profile, as motivated by the radial distribution of galaxies in clusters (Fig. 7.20).

The halo model as sketched above predicts the two-point correlation functions of matter, galaxies, and the cross-correlation between matter and galaxies. Whereas this prescription of the distribution of matter and galaxies contains a number of free functions—such as $\langle N|M\rangle$—these are well constrained by comparing the predicted abundance of galaxies with observations. Most of the properties of the halo model can be summarized in the fraction of galaxies α which are satellite galaxies, i.e., a fraction $(1-\alpha)$ of all galaxies are central galaxies.

Application to galaxy-galaxy lensing. Since galaxy-galaxy lensing measures the mean shear around galaxies, and since the shear is directly related to the mass distribution [see (7.70)], GGL measures the correlation between galaxy positions and the matter distribution. Therefore, the expected GGL signal can be obtained from the halo model as described above.

In Fig. 7.27, the GGL signal as predicted by the halo model is illustrated, for a halo of fixed virial mass M_{200}. The signal is a superposition of various terms. First, the baryons of the galaxies under consideration produce a lensing signal on small scales, due to their compactness. Second, central galaxies and satellites both contribute to the signal, with their respective abundance fractions $(1-\alpha)$ and α, respectively.

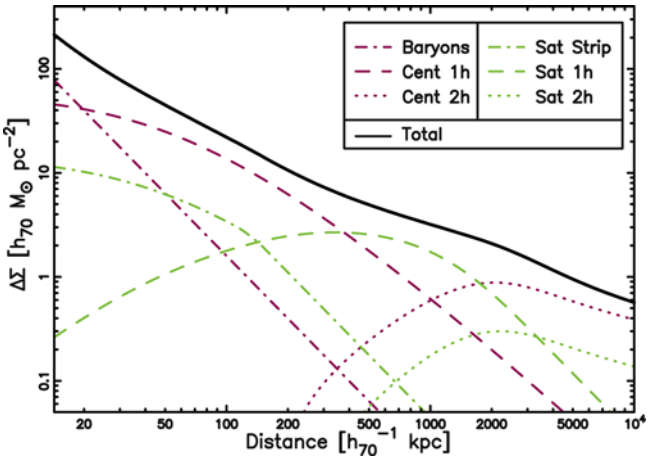

Fig. 7.27 Illustration of the GGL signal in terms of the halo model. Here, a halo of virial mass $M_{200} = 10^{12} M_\odot$ is chosen, with a stellar mass of $M_* = 5 \times 10^{10} M_\odot$. Furthermore, it is assumed that a fraction $\alpha = 0.2$ of all galaxies are satellite galaxies. The GGL signal is then the sum $\Delta\Sigma = \Delta\Sigma_{\rm bar} + (1-\alpha)\Delta\Sigma_{\rm cent} + \alpha\Delta\Sigma_{\rm sat}$ of three terms: That from the stellar mass of the galaxies ($\Delta\Sigma_{\rm bar}$), the signal around central galaxies ($\Delta\Sigma_{\rm cent}$), weighted by their fraction $(1-\alpha)$, and the signal around satellite galaxies ($\Delta\Sigma_{\rm sat}$), weighted by their fraction α. Each of these two latter terms is composed of the signal generated by the matter in the same halo where the galaxies are located (the one-halo term), and that of the neighboring halos (the two-halo term). Finally, one assumes that the satellite galaxies have their own dark matter subhalo, which is a (tidally) stripped version of its original host halo. Hence one writes $\Delta\Sigma_{\rm cent} = \Delta\Sigma^{\rm 1h}_{\rm cent} + \Delta\Sigma^{\rm 2h}_{\rm cent}$ and $\Delta\Sigma_{\rm sat} = \Delta\Sigma^{\rm strip}_{\rm sat} + \Delta\Sigma^{\rm 1h}_{\rm sat} + \Delta\Sigma^{\rm 2h}_{\rm sat}$. The various *curves* in the figure represent these different contributions, as labeled. Source: M. Velander et al. 2013, *CFHTLenS: The relation between galaxy dark matter haloes and baryons from weak gravitational lensing*, arXiv:1304.4265, p. 6, Fig. 3. Reproduced by permission of the author

For both of them, the mass correlated with the galaxy can reside in the same halo as the galaxy, giving rise to the 1-halo term, or in a different halo. Finally, it was assumed for Fig. 7.27 that the satellite galaxies have retained their own dark matter (sub-)halo, though with only half the mass an isolated galaxy of the same luminosity would have; this reduction of halo mass is a natural consequence of tidal stripping.

We see that on small scales, the baryons of the galaxies and the 1-halo term of central galaxies totally dominate the lensing signal. Hence, on scales smaller than about the virial radius of the halo (which is about 200 kpc for the halo mass assumed in the figure), the GGL signal indeed probes the radial mass profile of the galaxy+dark matter halo. Beyond the virial radius, the 1-halo term of satellite galaxies becomes stronger and then starts to dominate the signal. For scales beyond ~ 1 Mpc, the signal becomes increasingly dominated by other halos which are correlated with the host halo, i.e., the 2-halo terms of central galaxies and satellites becomes the dominant signal. The transition between these various regimes gives the total GGL signal its characteristic shape, as seen by the black solid curve in Fig. 7.27, which is also seen in Figs. 7.26 and 7.25.

Studying the GGL results as a function of luminosity or stellar mass, and separately for red and blue galaxies, one finds a number of interesting results. First, for fixed luminosity, the signal is considerably larger for red galaxies than for blue ones. This is particularly true at large separations which reflects the clustering properties of the halos in which the different galaxy types are embedded. The halo model yields a satisfactory fit to the data, as shown in Fig. 7.26. From these fits, the halo mass M_{200} as a function of luminosity or stellar mass M_* can be obtained. It is found that the $M_{200}(M_*)$-relation is steeper for red galaxies. When fitted with a power law, one finds $M_{200} \propto M_*^{\sim 1.5}$ for red galaxies, whereas the slope is flatter than unity for blue galaxies. Furthermore, the satellite fraction is higher for red galaxies, approaching unity for $M_* \lesssim 5 \times 10^9 M_\odot$ (i.e., low-mass red galaxies are almost never central galaxies in halos), and decreasing to $\alpha \sim 0.2$ for higher-mass red galaxies. In contrast, almost all blue galaxies are centrals.

Studying the relation between halo and stellar mass further, one finds that the two are not simply related by a power law: The ratio M_{200}/M_* is not a monotonic function of M_*, but reaches a minimum at $M_* \sim 5 \times 10^{10} M_\odot$, as shown in Fig. 7.28. This result implies that the efficiency with which gas is transformed into stars in a halo is a function of halo mass, and that there is a preferred mass scale for maximum star-formation efficiency. We will discuss this result and its implications in much more detail in Chap. 10.

7.7.4 Masses of groups and clusters

In a similar manner as done for galaxies, one can also superpose the weak lensing signal of galaxy groups and clusters in order to determine their mean mass profiles, as a function of some observable, such as the optical luminosity of the group or its richness (i.e., number of bright group members). The large number of redshifts obtained with the SDSS allowed the construction of group catalogs, based on the spatial (i.e., 3D) overdensity of galaxies; in particular, we described the properties of the maxBCG group catalog in Sect. 6.2.4.

For this group sample, the galaxy-group signal $\Delta\Sigma$ as a function of separation is shown in Fig. 7.29, separately for different richness bins of these groups. The first point to note is that the lensing signal, and thus the mass, increases with increasing richness—more massive groups contain more luminous galaxies on average.

The interpretation of the lensing signal is again performed in the framework of the halo model, and the differently colored curves in Fig. 7.29 correspond to the various contributions. The NFW-profile of the dark matter halo

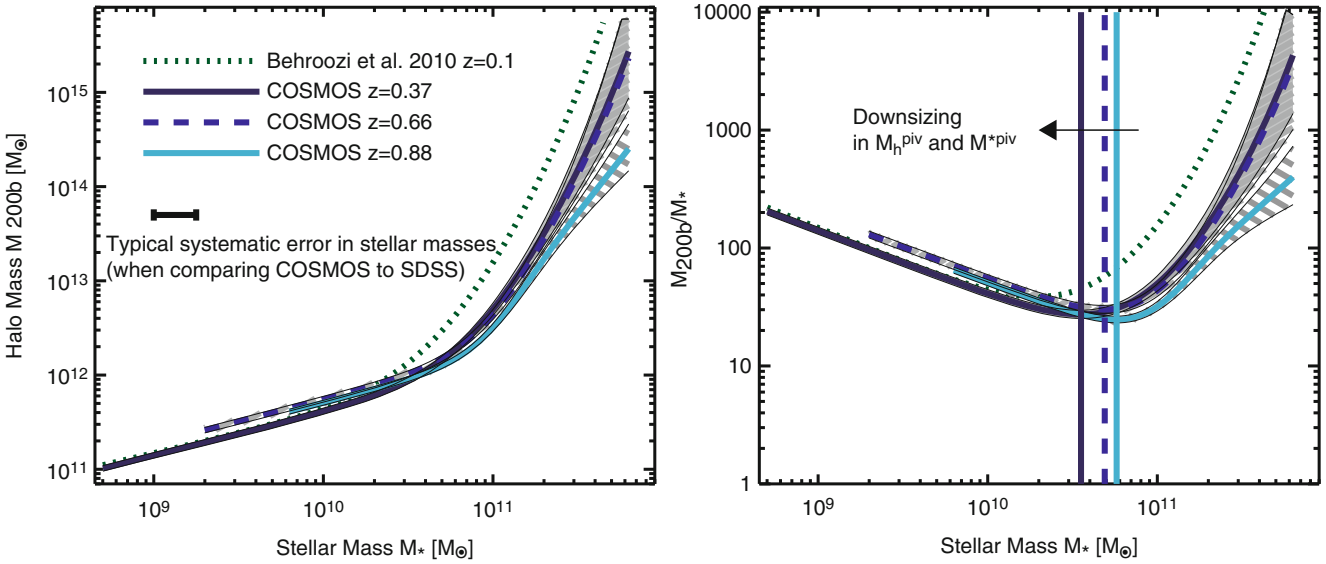

Fig. 7.28 From the analysis of the GGL signal in the COSMOS data field, the relation between stellar and halo mass can be studied over a wide range of galaxy masses and redshifts. In the *left panel*, the halo mass is plotted as a function of the stellar mass, as obtained by fitting the GGL signal with the halo model, for three different redshift intervals (where the mean redshift is indicated by line-type), and the *shaded area* around the curves indicates the estimated uncertainty. For comparison, the *dotted curve* shows the same relation obtained by matching the abundance of low-redshift galaxies with the halo mass function. The relative calibration between the two different methods is uncertain by an amount indicated by the error bar. One finds that the functional form of the $M_{200}(M_*)$-relation exhibits a characteristic change of slope, steepening above $\sim 5 \times 10^{10} M_\odot$. In the *right panel*, the same result is shown, except that now the ratio M_{200}/M_* is plotted. This ratio has a minimum at the mass scale where the $M_{200}(M_*)$-relation steepens. Hence, there is a characteristic mass scale at which the stellar contents of a halo is maximized. As we will discuss in Chap. 10, this corresponds to a mass scale of halos where the conversion of baryons into stars is maximally efficient. This characteristic mass scale seems to decrease with redshift, an effect sometimes called 'downsizing'. Source: A. Leauthau et al. 2012, *New Constraints on the Evolution of the Stellar-to-dark Matter Connection: A Combined Analysis of Galaxy-Galaxy Lensing, Clustering, and Stellar Mass Functions from* $z = 0.2$ *to* $z = 1$, ApJ 744, 159, p. 17, Fig. 11. ©AAS. Reproduced with permission

of the groups is shown as green curves, whereas the red curves show the lensing signal from the baryonic component of the brightest cluster galaxy (BCG), assumed to represent the center of the group halo. However, not in every case is the BCG correctly identified as the group center. For a fraction of groups, the BCG is displaced from the center (or may even be misidentified). This fraction was estimated from simulations, and amounts to some 40 % for groups with small richness, decreasing to $\sim$ 20 % for more massive groups and clusters. The corresponding correction to the lensing signal is shown as orange curves. Finally, the 2-halo term dominates the signal on the largest scales. The sum of these contributions is shown by the violet curves, which provide a very good fit to the lensing data.

From the model fit of the stacked group/cluster lensing signal shown in Fig. 7.29, one can derive the halo mass M_{200} as a function of the richness or the total optical luminosity of the groups. The results are shown in Fig. 7.30. The halo mass increases monotonically with richness and luminosity, approximately as $M_{200} \propto N_{200}^{1.28}$ and $M_{200} \propto L_{200}^{1.22}$. In particular we note that the latter relation implies that the mass-to-light ratio of clusters increases with mass, in agreement with the results obtained from GGL in the COSMOS field (see Fig. 7.28). Furthermore, it is found that the mass of the BCG increases with the halo mass, for low-mass groups, but then saturates for large group masses.

Summarizing this section, the statistical (or stacked) weak lensing signal of galaxies and groups provides the most direct way to study the relation between their observed properties and their mass properties. The NFW mass profile yields satisfactory fits to the lensing data when used in the context of the halo model, which provides a convenient framework of parametrizing the distribution of mass and galaxies in the Universe.

7.8 The substructure of halos

Sub-halos of galaxies and clusters of galaxies. Numerical simulations of structure formation in the CDM model show that the mass density in halos is not smooth; instead, they reveal that halos contain numerous halos of much lower mass, so-called sub-halos. For instance, a halo with the mass of a galaxy cluster contains hundreds or even thousands of halos with masses that are orders of magnitude lower. These sub-halos are indeed observed, since the substructure in clusters is visible—in the form of the cluster galaxies. In

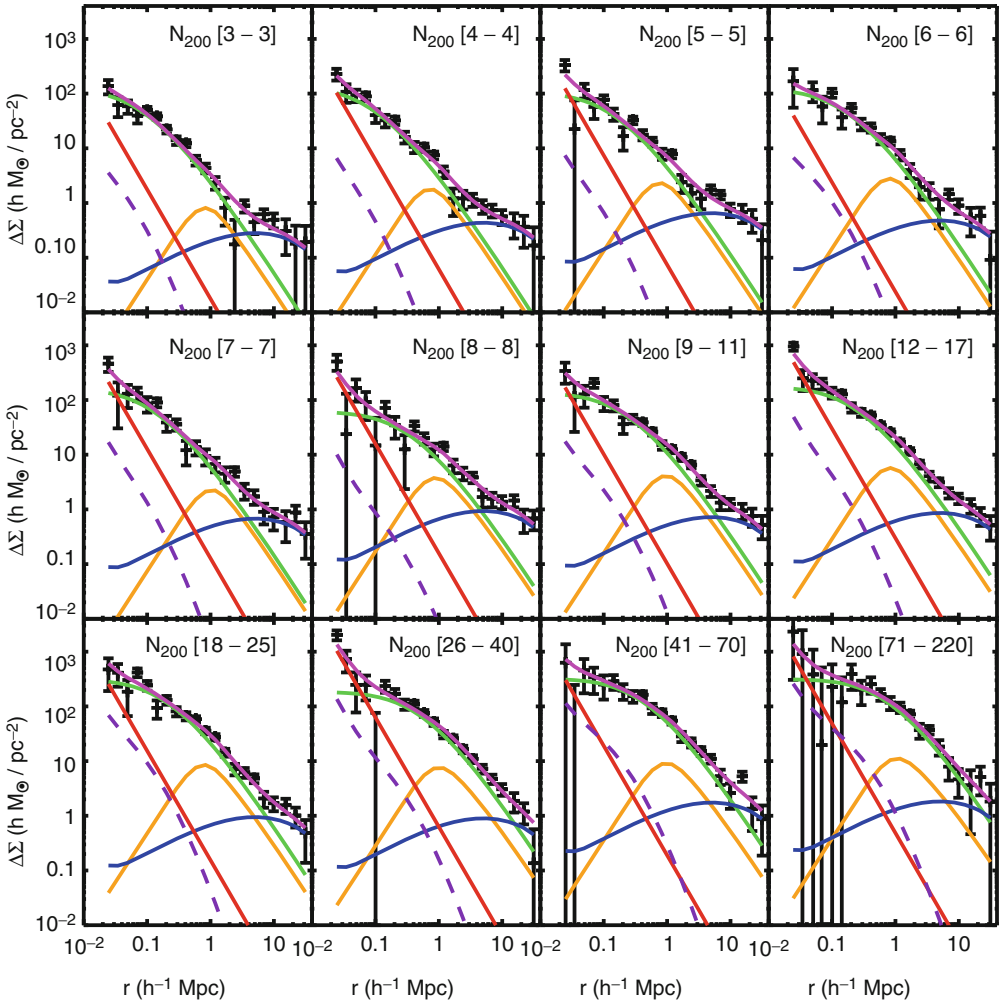

Fig. 7.29 The galaxy-group lensing signal for the maxBCG sample of groups/clusters, where different *panels* correspond to different richness bins. The signal is modeled by a number of different contributions: *Green* shows the NFW signal of the dark matter halo, *orange* is the contribution from miscentering the halo, *red* the signal from the brightest cluster galaxy, and *blue* the contribution from neighboring halos. Source: D.E. Johnston et al. 2007, *Cross-correlation Weak Lensing of SDSS galaxy Clusters II: Cluster Density Profiles and the Mass–Richness Relation*, arXiv:0709.1159, Fig. 8. Reproduced by permission of the author

the upper part of Fig. 7.31, the simulation of a cluster and its substructure is displayed. In fact, this mass distribution looks just like the mass distribution expected in a cluster of galaxies, with the main cluster halo and its distribution of member galaxies. The lower part of Fig. 7.31 shows the simulation of a halo with mass $\sim 2 \times 10^{12} M_\odot$, which corresponds to a massive galaxy. As one can easily see, its mass distribution shows a large number of sub-halos as well. In fact, the two mass distributions are nearly indistinguishable, except for their scaling in the total mass.[12] The presence of substructure over a very wide range in mass is a direct consequence of hierarchical structure formation, in which objects of higher mass each contain smaller structures that have been formed earlier in the cosmic evolution.

Such simulations show that of order $\sim 10\,\%$ of the mass of halos is contained in sub-halos, with a fraction that is slightly smaller for galaxy-mass halos ($\sim 7\,\%$) than for cluster-mass halos. Furthermore, the mass spectrum of the sub-halos follows a simple power law, $n(M) \propto M^{-1.9}$, down to the smallest mass-scale that can be resolved by simulations (which is currently about 10^{-7} times the mass of the parent halo). In fact, the very small velocity dispersion of cold dark matter particles predicts that this mass spectrum should continue down to the smallest halo masses that can be formed by CDM—which is about an Earth mass, or $10^{-6} M_\odot$.

[12] The reason for this is found in the property of the power spectrum of density fluctuations that has been discussed in Sect. 7.5.2, namely that $P(k)$ can be approximated by a power law over a wide range in k. Such a power law features no characteristic scale. For this reason, the properties of halos of high and low mass are scale-invariant, as is clearly visible in Fig. 7.31.

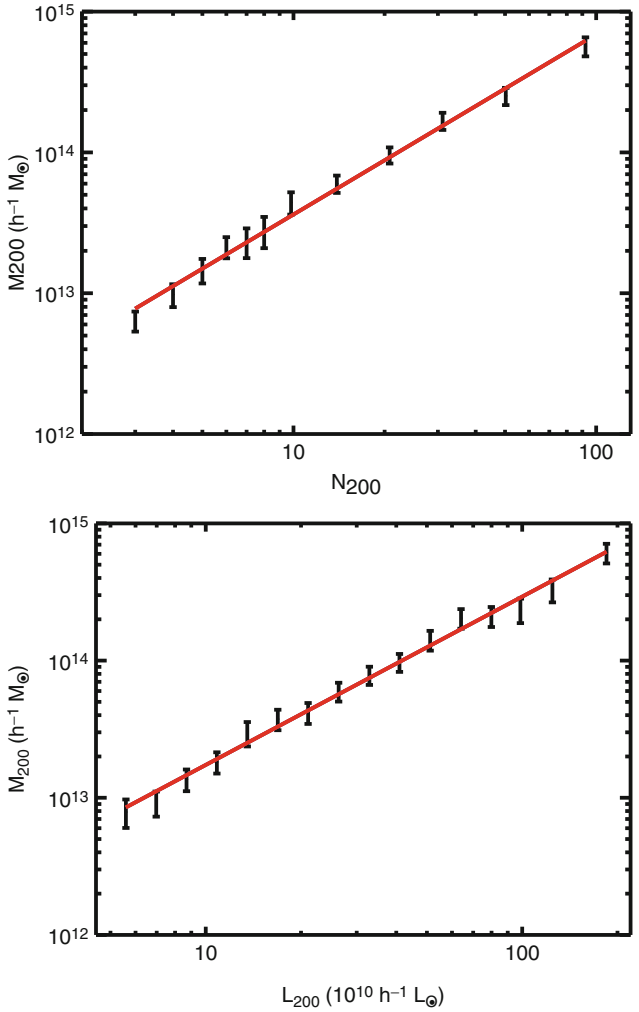

Fig. 7.30 Cluster mass as a function of optical richness (*top*) and cluster luminosity (*bottom*). For the definition of richness and luminosity, the number of red galaxies brighter than $0.4L^*$ and within a projected separation less than $1\,h^{-1}$Mpc from the brightest cluster galaxy (BCG), N_g, is used to define the radius $r_g = 0.156 N_g^{0.6} h^{-1}$ Mpc. This was previously found to be the radius within which the mean luminosity density is 200 times the average cosmic luminosity density of galaxies, which is determined from the galaxy luminosity function. The number of red galaxies within r_g is then defined as N_{200}; likewise, the sum of the luminosities of all red galaxies within r_g is defined as L_{200}. Binning the groups and clusters according to their richness and luminosity, and separately analyzing their weak lensing signal yielded the mass-richness and mass-luminosity relations shown. Note that here also very poor groups are included; some of them consisting of just two or a few galaxies. However, there are still more than 13 000 clusters with $N_{200} \geq 10$. Clusters are restricted to the redshift range $0.1 \leq z \leq 0.3$. Source: D.E. Johnston et al. 2007, *Cross-correlation Weak Lensing of SDSS galaxy Clusters II: Cluster Density Profiles and the Mass–Richness Relation*, arXiv:0709.1159, Fig. 11. Reproduced by permission of the author

Fig. 7.31 Density distribution of two simulated dark matter halos. In the *top image*, the halo has a virial mass of $5 \times 10^{14} M_\odot$, corresponding to a cluster of galaxies. The halo in the *bottom image* has a mass of $2 \times 10^{12} M_\odot$, representing a massive galaxy. In both cases, the presence of substructure in the mass distribution can be seen. It can be identified with individual cluster galaxies in the case of the galaxy cluster. The substructure in a galaxy can not be identified easily with any observable source population; one may expect that these are satellite galaxies, but observations show that these are considerably less abundant than the substructure seen here. Apart from the length-scale (and thus also the mass-scale), both halos appear very similar from a qualitative point of view. Source: B. Moore et al. 1999, *Dark Matter Substructure within Galactic Halos*, ApJ 524, L19, p. L20, Fig. 1. ©AAS. Reproduced with permission

The spatial distribution of the sub-halos is less concentrated towards the halo center than the total mass. The reason for this property lies in the fact that sub-halos whose orbits bring them deep into the potential well of the host halo are subject to strong tidal forces, and they get disrupted in the course of evolution. Simulations which include gas physics (we will describe some of these simulations in Sect. 10.6.1 below) find that the disruption becomes weaker if baryons are included—their dissipational nature leads to more compact, and thus more tightly bound, sub-halos; hence, they can resist the disruptive tidal forces for a longer time.

The 'substructure problem'. As we will discuss in detail in the next chapter, the CDM model of cosmology has proven to be enormously successful in describing and predicting cosmological observations. Because this model has achieved this success and is therefore considered the standard model, results that apparently do not fit into this standard model are of particular interest. The rotation curves of LSB galaxies mentioned above are one such result. Either one finds a good explanation for this apparent discrepancy between observation and the predictions of the CDM model—such as we indicated above—or, otherwise, results of this kind may necessitate to introduce extensions to the CDM model. In the former case, the model would have overcome another hurdle in demonstrating its consistency with observations and would be strengthened even further, whereas in the latter case, new insights would be gained into the physics of cosmology. Besides the rotation curves of dwarf and LSB galaxies, there is another observation that does not seem to fit into the picture of the CDM model at first sight.

Whereas the substructure in clusters is easily identified with the cluster member galaxies, the question arises as to what the sub-halos in galaxy-mass halos can possibly correspond to. The mass spectrum of these halos, as obtained from an early simulation, is displayed in Fig. 7.32. Some of these sub-halos are recognized in our Milky Way, namely the known satellite galaxies like, e.g., the Magellanic Clouds. In a similar way, the satellite galaxies of the Andromeda galaxy may also be identified with sub-halos. However, as we have seen in Sect. 6.1, not more than 40 members of the Local Group were known before 2003—whereas the numerical simulations predict hundreds of satellite galaxies for the Galaxy. This apparent deficit in the number of observed sub-halos, clearly indicated in Fig. 7.32, is considered to be another potential problem of CDM models.

We note that since this discrepancy was first explicitly pointed out, some 25 new dwarf galaxies in the Local Group have been detected from the Sloan Digital Sky Survey. Given that the SDSS observed only a quarter of the sky, one expects to have at least another $\sim$ 70 satellite galaxies in the Local Group which are not contained in the footprint of the SDSS. Therefore, the difference between the number of satellite galaxies and the predictions from numerical simulations has become smaller in recent years. On the other hand, these newly discovered satellites are all of very low mass, and thus do not fully fill in the gap between the curves in Fig. 7.32.

However, one always needs to remember that the dark matter simulations only predict the distribution of mass, and not that of light (which is accessible to observation). One possibility of resolving this apparent discrepancy centers on the interpretation that these sub-halos do in fact exist, but that most of them do not, or only weakly, emit radiation. What appears as a cheap excuse at first sight is indeed already part of the models of the formation and evolution of galaxies. As

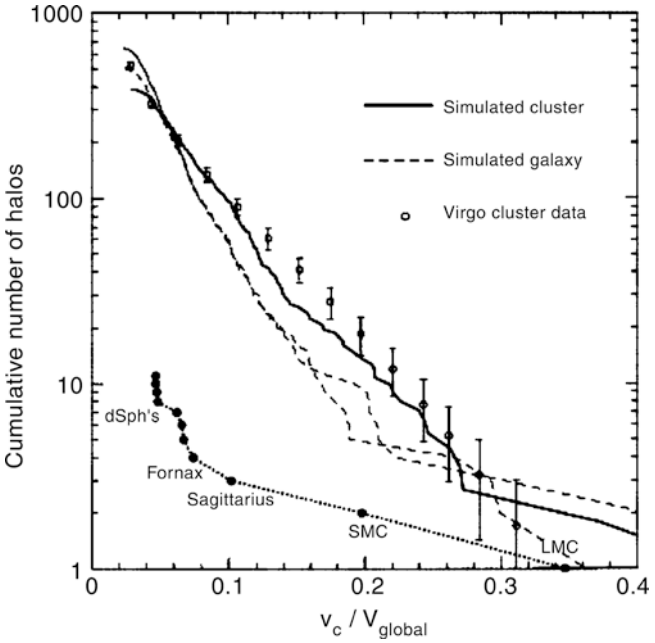

Fig. 7.32 Number density of sub-halos as a function of their mass. The mass is expressed by the corresponding Keplerian rotational velocity v_c, measured in units of the corresponding rotational velocity of the main halo. The *curves* show this number density of sub-halos with rotational velocity $\geq v_c$ for a halo of either cluster mass or galaxy mass. The observed numbers of sub-halos (i.e., of galaxies) in the Virgo cluster are plotted as *open circles* with error bars, and the number of satellite galaxies of the Milky Way as *filled circles*. One can see that the simulations describe the abundance of cluster galaxies quite well, but around the Galaxy significantly fewer satellite galaxies exist than predicted by a CDM model. Source: B. Moore et al. 1999, *Dark Matter Substructure within Galactic Halos*, ApJ 524, L19, p. L20, Fig. 2. ©AAS. Reproduced with permission

will be discussed in Sect. 10.7 in more detail, it is difficult to form a considerable stellar population in halos of masses below $\sim 10^9 M_\odot$. Most halos below this mass threshold will therefore be hardly detectable because of their low luminosity. Thus, in this picture, sub-halos in galaxies are in fact present, as predicted by the CDM models, but most of these would be 'dark'.

The low-mass satellite galaxies in the Local Group that were recently discovered by the Sloan Survey all have a very large mass-to-light ratio, which implies that their stellar mass-to-halo mass ratio is very small. In fact, many of these dwarf galaxies are less luminous than a star cluster, with $L \sim 10^2$–$10^4\, L_\odot$. In contrast to these small luminosities, they all display a rather high stellar velocity dispersion of $\sigma \sim 5\,\mathrm{km/s}$, indicating a fairly high mass. Indeed, these dwarf galaxies are not only the faintest galaxies known, but also those with the largest mass-to-light ratio, $M/L \gtrsim 100$ in Solar units; for some of the newly discovered dwarfs, M/L apparently exceeds 10^4. Their extremely low metallicity argues for a very early epoch of star formation; this is confirmed by the color-magnitude diagrams for some of the

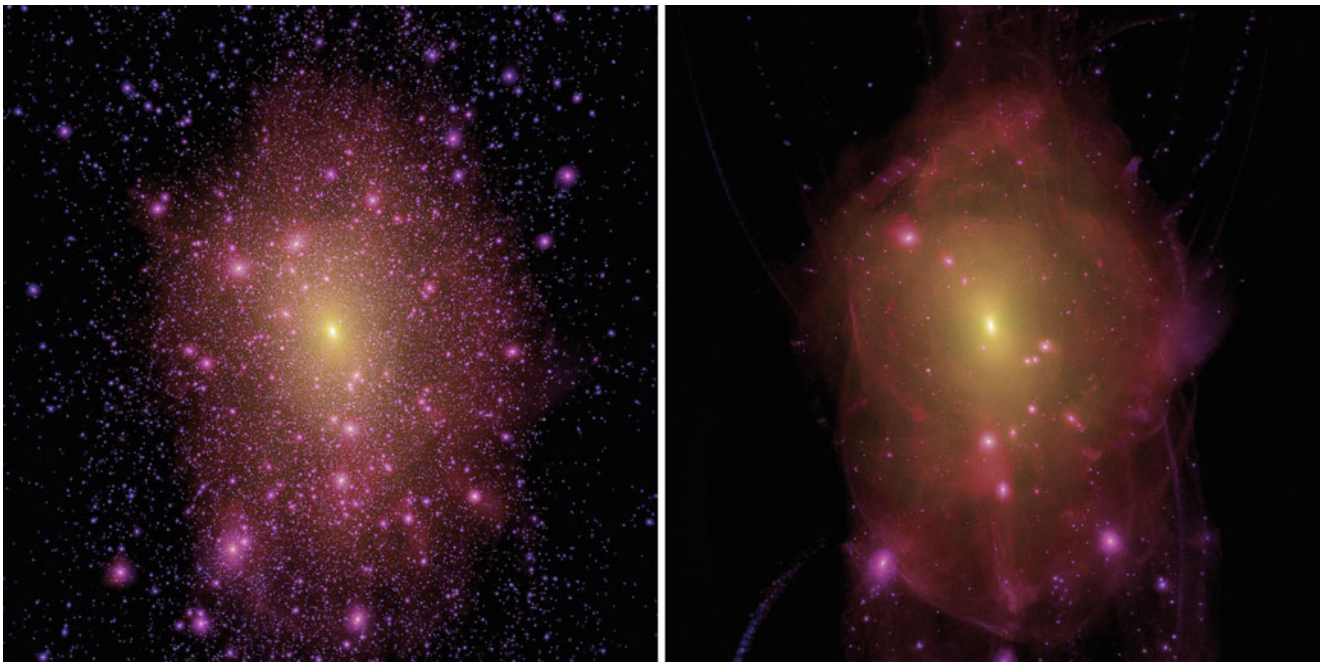

Fig. 7.33 Comparison of a galaxy-mass dark matter halo in the standard CDM model (*left panel*) and in a cosmological model with warm dark matter (*right panel*), within a 1.5 Mpc box. For the simulation, the initial conditions of the density field were chosen to be very similar, except that the power spectrum in the WDM model was truncated, due to the free-streaming of the particles. Obviously, the WDM model halo has far fewer satellite halos; in particular, those of the smallest mass are absent. Source: M. Lovell et al. 2011, *The haloes of bright satellite galaxies in a warm dark matter universe*, arXiv:1104.2929, Fig. 3. Reproduced by permission of the author

dwarfs, which assign them an age of $\sim$ 13 Gys. We shall see in Sect. 10.7 that these observed properties are naturally explained in the framework of galaxy evolution models.

Warm dark matter as alternative. The apparent conflict between the abundance of sub-halos and the observed satellite galaxies in the Milky Way can be potentially avoided if the initial power spectrum of density fluctuations has less power on small spatial scales—corresponding to masses of satellite galaxies. At the same time, the power spectrum at large spatial scales should not be affected, to not endanger the spectacular success of our cosmological model in the description of key cosmological observations (see next chapter). Such a modification of the density fluctuation spectrum would be a consequence of warm dark matter models; as we discussed above, their free streaming would wash out smaller-scale fluctuations. In particular, if the WDM particle has a mass of $\sim$ 2 keV, the cut-off in the power spectrum would correspond to the halos mass of dwarf galaxies.

Figure 7.33 shows the resulting mass distributions of a galaxy-scale halo as predicted by a CDM and a WDM model. In the latter, essentially all small-scale sub-halos are absent, which are found plentiful in the CDM simulation. Hence, in the WDM model, the satellite problem essentially is non-existent.

However, before jumping to conclusions, three points need to be made here. First, a WDM particle appears less natural as seen from the point of view of particle physics, although plausible candidates may exist in some extensions of the Standard Model of particle physics. Second, observations of the Lyman-α forest strongly constrain the allowed mass range of WDM particles (see Sect. 8.5), and lower limits on the mass of the WDM particle obtained from these studies come exceedingly close to the mass needed to substantially reduce the abundance of sub-halos. Third, sub-halos are in fact observed indirectly, as discussed next.

Evidence for the presence of CDM substructure in galaxies. A direct indication of the presence of substructure in the mass distribution of galaxies indeed exists, which originates from gravitational lens systems. As we have seen in Sect. 3.11, the image configuration of multiple quasars can be described by simple mass models for the gravitational lens. Concentrating on those systems with four images of a source, for which the position of the lens is also observed (e.g., with the HST), a simple mass model for the lens has fewer free parameters than the coordinates of the observed quasar images that need to be fitted. Despite of this, it is possible, with very few exceptions, to describe the angular positions of the images with such a model very accurately. This result is not trivial, because for some lens systems which are observed using VLBI techniques, the image positions are known with a precision of better than 10^{-4} arcseconds, with an image separation of the order of $1''$. This result demonstrates that the

mass distribution of lens galaxies is, on scales of the image separation, quite well described by simple mass models.

Besides the image positions, such lens models also predict the magnifications μ of the individual images. Therefore, the ratio of the magnifications of two images should agree with the flux ratio of these images of the background source. The surprising result from the analysis of lens systems is that, although the image positions of (nearly) all quadruply imaged systems are very precisely reproduced by a simple mass model, in not a single one of these systems does the mass model correctly reproduce the flux ratios of the images!

Perhaps the simplest explanation for these results is that the simple mass models used for the lens are not correct and other kinds of lens models should be used. However, this explanation can be excluded for many of the observed systems. Some of these systems contain two or three images of the source that are positioned very closely together, for which one therefore knows that they are located close to a critical curve. In such a case, the magnification ratios can be estimated quite well analytically; in particular, they no longer depend on the exact form of the lens model employed. Hence, the existence of such 'universal properties' of the lens mapping excludes the existence of *simple* (i.e., 'smooth') mass models capable of describing the observed flux ratios. One example of this is presented in Fig. 7.34.

The natural explanation for these flux discrepancies is the fact that a lensing galaxy does not only have a smooth large-scale mass profile, but that there is also small-scale substructure in its density. In the case of spiral galaxies, this may be the spiral arms, which can be seen as a small-scale perturbation in an otherwise smooth mass profile. However, most lens galaxies are ellipticals. The sub-halos that are predicted by the CDM model may then represent the substructure in their mass distribution. For a further discussion of this model, we first should mention that a small-scale perturbation of the mass profile only slightly changes the deflection angle caused by the lens, whereas the magnification μ may be modified much more strongly. As a matter of fact, by means of simulations, it was demonstrated that lens galaxies containing sub-halos of about the same abundance as postulated by the CDM model give rise to a statistical distribution of discrepancies in the flux ratios which is very similar to that found in the observed lens systems. Furthermore, these simulations show that, on average, a particular image of the source is clearly demagnified compared to the predictions by simple, smooth lens models, again in agreement with the observational results. And finally, in the case that a relatively massive sub-halo is located close to one of the images, the image position should also be slightly shifted, compared to the smooth mass model. This effect was in fact directly detected in two lens systems: in these cases, a sub-halo exists in the lens galaxy

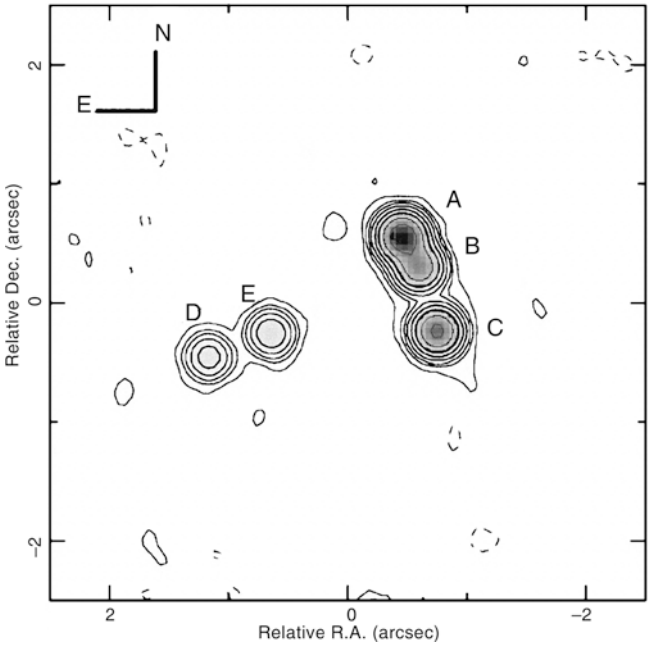

Fig. 7.34 8.5 GHz map of the lens system 2045+265. The source at $z_s = 1.28$ is imaged four-fold (components A–D) by a lens galaxy at $z_d = 0.867$, while component E represents emission from the lens, as is evident from its different radio spectrum. The three images A, B, and C have a separation which is much smaller than the Einstein radius of the lens. From the general properties of the gravitational lens mapping, one can show that any 'smooth' mass model of the lens predicts the flux of the middle one of those (i.e., image B) to be roughly the same as the sum of the fluxes of components A and C. Obviously, this rule is strongly violated in this lens system, because B is weaker than either A or C. This result can only be explained by small-scale structure in the mass distribution of the lens galaxy. Source: C. Fassnacht et al. 1999, *B2045+265: A New Four-Image Gravitational Lens from CLASS*, AJ 117, 658, p. 659, Fig. 1. ©AAS. Reproduced with permission

which is massive enough to form stars, and which therefore can be observed. Its effect on the magnification and the image position can then be inferred from the lens model (see Fig. 7.35).

In strong lensing clusters, one can actually see the impact of mass substructure quite clearly: The three arcs to the left of the center in the cluster Cl 0024+17 (see lower panel of Fig. 6.51) are predicted by any smooth mass model to follow a 'universal' behavior, in that the middle of the arcs should have a length that is the sum of the lengths of the two outer arcs. Clearly, the middle arc is seen to be by far the shortest. This is due to substructure, which is easily identified by the two cluster galaxies located close to the middle arc and thus destroying this universal behavior of the lens mapping.

In addition to these flux-ratio discrepancies, at least two sub-halos have been identified in modeling lens systems with Einstein rings for which very high-quality data were obtained. In one of these cases, where the estimated mass of the sub-halo amounts to $M \sim 3.5 \times 10^9\, M_\odot$, there are

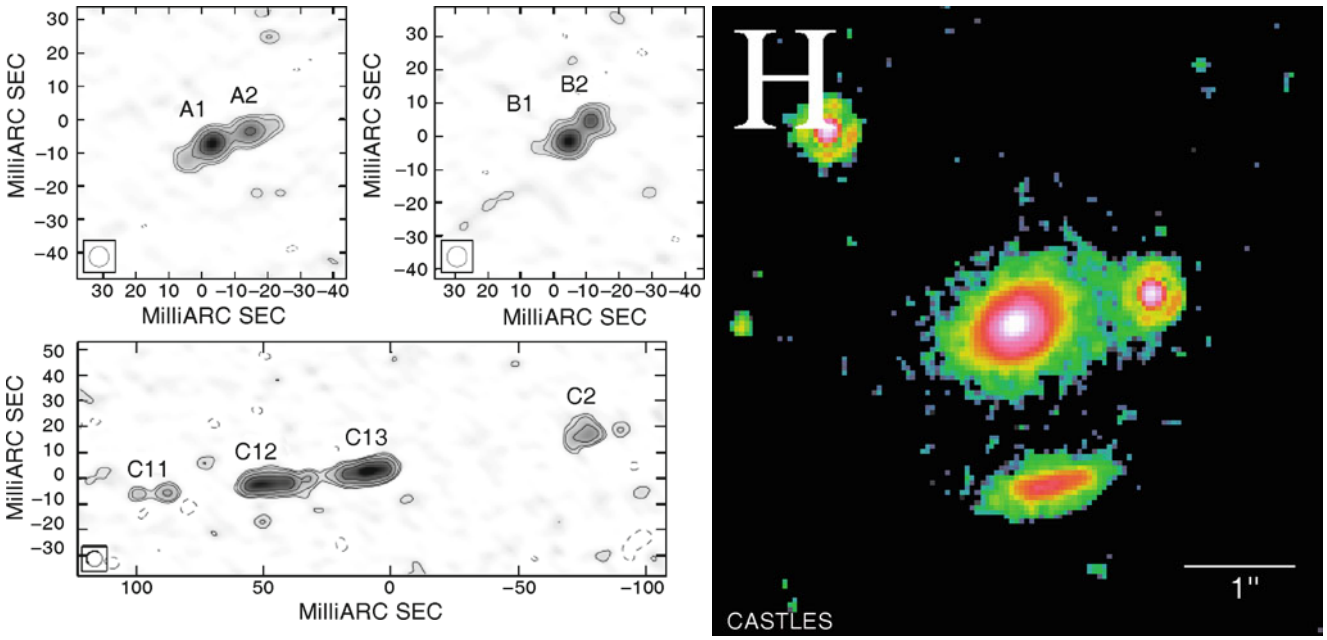

Fig. 7.35 *On the right*, an H-band image of the lens system MG 2016+112 is shown, consisting of a lens galaxy in the center and four images of the background source, the two southernmost of which are nearly merged in this image. *On the left*, VLBI maps of these components are presented; the radio source consists of a compact core and a jet component, clearly visible in images A and B. The VLBI map of component C reveals that is it in fact a double image of the source, in which the core and jet components each are visible twice. Any smooth mass model for the lens galaxy predicts that the separation C12—C11 should roughly be the same as that between C13—C2, which obviously contradicts the observation. In this case, the substructure in the mass distribution is even visible: if one includes the weak emission south of component C, which is visible in the image on the right, into the lens model as a mass component, the separation of the components in image C can be well modeled. Source: *Left*: L.V.E. Koopmans et al. 2002, *2016+112: a gravitationally lensed type II quasar*, MNRAS 334, 39, p. 41, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society. *Right*: Castles Collaboration/C.S. Kochanek, E.E. Falco, C. Impey, J. Lehar, B. McLeod, H.-W. Rix

stringent limits on its luminosity, which translates into a lower limit of its mass-to-light ratio of 120 in Solar units.

For these reasons, it is probable that galaxies contain sub-halos, as predicted by the CDM model, but most sub-halos, in particular those with low mass, contain only few stars and are therefore not visible. One consequence of this explanation is that the low-mass satellite galaxies that are seen in our Local Group should be dominated by dark matter. Given the faintness and low surface brightness of these galaxies, obtaining kinematical information for them is very difficult and requires large telescopes for spectroscopy of individual stars in these objects. The results of such investigations indicate that the dwarf galaxies in the Local Group are indeed dark matter dominated, with a mass-to-light ratio of ~ 100 in Solar units. Whereas some uncertainty remains, e.g., related to the assumption of dynamical equilibrium, it is clear that these faint satellites represent sub-halos which are unusually poor in stars.

The 'disk of satellite galaxies'. Whereas the abundance of dark matter subhalos in galaxies no longer presents a serious problem for CDM models of structure formation, the spatial distribution of satellite galaxies around the Milky Way requires more explanation. As we mentioned in Sect. 6.1.1, the 11 classical satellites of the Galaxy seem to form a planar distribution. Such a distribution would be extremely unlikely if the satellite population was drawn from a near-isotropic probability distribution. Therefore, the planar satellite distribution has been considered as a further potential problem for CDM-like models. However, using semi-analytic models of galaxy formation, combined with simulations of the large-scale structure, a different picture emerges. Since galaxies preferentially form in filaments of the large-scale structure, the accretion of smaller mass halos onto a high-mass halo occurs predominantly in the direction of the filament. The most massive sub-halos therefore tend to form a planar distribution, not unlike the one seen in the Milky Way's satellite distribution. The anisotropy of the distribution of massive satellites may also serve to explain the Holmberg effect.

7.9 Origin of the density fluctuations

We have seen in Sect. 4.5.3 that the horizon and the flatness problem in the normal Friedmann–Lemaître evolution of the Universe can be solved by postulating an early phase of very rapid—exponential—expansion of the cosmos. In this

inflationary phase of the Universe, any initial curvature of space is smoothed away by the tremendous expansion. Furthermore, the exponential expansion enables the complete currently visible Universe to have been in causal contact prior to the inflationary phase. These two aspects of the inflationary model are so attractive that today most cosmologists consider inflation as part of the standard model, even if the physics of inflation is as yet not understood in detail.

Density fluctuations from inflation. The inflationary model has another property that is considered to be essential. Through the huge expansion of the Universe, microscopic scales are blown up to macroscopic dimensions. The large-scale structure in the current Universe corresponds to microscopic scales prior to and during the inflationary phase. From quantum mechanics, we know that the matter distribution cannot be fully homogeneous, but it is subject to quantum fluctuations, expressed, e.g., by Heisenberg's uncertainty relation. By inflation, these small quantum fluctuations are expanded to large-scale density fluctuations. For this reason, the inflationary model also provides a natural explanation for the origin of initial density fluctuations.

Indeed, it is the only mechanism known in which perturbations can be generated which are larger than the horizon. As we discussed in the framework of the 'horizon problem', two points further apart than $\sim 1^\circ$ have not been in causal contact before recombination when considering standard Friedmann expansion. Despite of this, we observe temperature fluctuations in the CMB on larger scales, which implies that density perturbations larger than the horizon scale were present at $z \sim 1100$. In the same manner as inflation provides an explanation for the horizon problem, it explains the possibility to have superhorizon fluctuations—before the inflationary phase, the whole visible Universe had been in causal contact.

The primordial power spectrum. In fact, one can study the generation of macroscopic density perturbations from quantum fluctuations quantitatively and calculate the initial power spectrum of these fluctuations. The result of such investigations depends slightly on the details of the inflationary model they are based on. However, these models agree in their prediction that the initial power spectrum should have a form very similar to the Harrison–Zeldovich fluctuation spectrum, except that the spectral index n_s of the primordial power spectrum should be slightly smaller than the Harrison–Zeldovich value of $n_\mathrm{s} = 1$. Thus, the model of inflation can be directly tested by measuring the power spectrum and, as we shall see in Chap. 8, the power-law slope n_s indeed seems to be slightly, but significantly flatter that unity, as expected from inflation. The deviation of n_s from unity is called the *tilt* of the initial density fluctuation spectrum.

The various inflationary models also differ in their predictions of the relative strength of the fluctuations of space-time, which should be present after inflation. Such fluctuations are not directly linked to density fluctuations, but they are a consequence of General Relativity, according to which space-time itself is also a dynamical quantity. One consequence of this is the existence of gravitational waves. Although no gravitational waves have been directly detected until now, the analysis of the double pulsar PSR J1915+1606 proves the existence of such waves.[13] Primordial gravitational waves provide an opportunity to empirically distinguish between the various models of inflation. These gravitational waves leave a 'footprint' in the polarization of the cosmic microwave background that is measurable in principle. Several experiments are currently searching for this polarization signature in the CMB.

7.10 Problems

7.1. Homogeneous solution of the Euler–Poisson system. The system of equations (7.2), (7.3), (7.4) admits an exact solution, namely that of a homogeneous universe with an expansion law given by $\boldsymbol{v}(\boldsymbol{r}, t) = H(t)\boldsymbol{r}$, as will be shown here.

1. If we require the density to be homogeneous at all times, show that this implies that $\nabla \cdot \boldsymbol{v}$ is independent of $\boldsymbol{r}$.
2. Show that the continuity equation implies for the Hubble velocity field that the density ρ satisfies (4.11).
3. Determine $\nabla\Phi$ from (7.4), and show that the Euler equation for the Hubble velocity field then reduces to the Friedmann equation (4.19), specialized to the case of vanishing pressure.

7.2. The growth equation. A second-order differential equation such as (7.15) has two linearly independent solutions. In general, they are difficult to find, and in the general case, they must be obtained numerically. However,

[13] The binary pulsar PSR J1915+1606 was discovered in 1974. From the orbital motion of the pulsar and its companion star, gravitational waves are emitted, according to General Relativity. Through this, the system loses kinetic (orbital) energy, so that the size of the orbit decreases over time. Since pulsars represent excellent clocks, and we can measure time with extremely high precision, this change in the orbital motion can be observed with very high accuracy and compared with predictions from General Relativity. The fantastic agreement of theory and observation is considered a definite proof of the existence of gravitational waves. For the discovery of the binary pulsar and the detailed analysis of this system, Russell Hulse and Joseph Taylor were awarded the Nobel Prize in Physics in 1993. In 2003, a double neutron star binary was discovered where pulsed radiation from both components can be observed. This fact, together with the small orbital period of 2.4 h implying a small separation of the two stars, makes this an even better laboratory for studying strong-field gravity.

the growth equation (7.15) can be solved explicitly, as will be shown here.

1. Show that the Hubble function $H(t)$ is a solution of the growth equation. Hint: Make use of the second Friedmann equation (4.19).
2. Unfortunately, $H(t)$ decreases with time, and thus is not the growing solution we are interested in. Show that a second solution of (7.15) is given by

$$D_+(a) = CH(a) \int_0^a \frac{\mathrm{d}a'}{[a'\,H(a')]^3} \ ,$$

   where C is a constant, chosen such that $D_+(1) = 1$. Hint: make use of the first part of this problem.
3. Use the expansion law of an Einstein–de Sitter universe, $a(t) = (t/t_0)^{2/3}$, to show that the corresponding Hubble function $H(t)$ and D_+ as calculated from (7.17) solve the growth equation (7.15).

7.3. Bulk properties of dark matter halos. Consider two dark matter halos of mass $10^{12}h^{-1}M_\odot$ and $10^{15}h^{-1}M_\odot$, corresponding to the halos of a massive galaxy and of a massive cluster, respectively. Assume $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$.

1. What is the virial radius r_{200} and the virial velocity V_{200} of these halos at redshift $z = 2$ and today? Hint: make use of the fact that the Schwarzschild radius of the Sun is $\approx 3\,\mathrm{km} = 3 \times 10^5\,\mathrm{cm}$, and $c/H_0 \approx 3h^{-1}\,\mathrm{Gpc} \approx 9 \times 10^{27}h^{-1}\,\mathrm{cm}$
2. In order to assemble this mass into a halo, matter from a large region must have accumulated. Assuming that this region is spherical, determine its comoving radius.
3. This radius can be identified with the typical length-scale of a perturbation out of which such dark matter halos grow. At which redshift do these density fluctuations enter the horizon?

7.4. Behavior of the growth factor. We have seen that, in an Einstein–de Sitter universe, the growth factor equals the scale factor—see (7.19). In universes with curvature and/or a cosmological constant, this is no longer the case, as seen in Fig. 7.3.

1. Show that for sufficiently small values of a during the matter-dominated epoch, the growth factor is proportional to a in all universes.
2. Derive the lowest-order correction to this linear behavior, and estimate from that the epoch when significant deviations from the linear behavior of the growth factor occur. Assume for simplicity that the universe is flat, and compare your estimate with Fig. 7.3.