First light in the universe : Saas-Fee Advanced Course 36, [3 to 8 April 2006 in Les Diablerets]

Schaerer, Daniel Olivier (ed.); Hempel, Angela (ed.); Puy, D. (ed.)

Saas-Fee Advanced Course 36

A. Loeb A. Ferrara R.S. Ellis

First Light in the Universe

Saas-Fee Advanced Course 36

Swiss Society for Astrophysics and Astronomy
Edited by D. Schaerer, A. Hempel and D. Puy

With 160 Figures, 25 in Color

Springer

Abraham Loeb
Department of Astronomy
Harvard University, 60 Garden St.
Cambridge, MA 02138, USA
aloeb@cfa.harvard.edu

Andrea Ferrara
SISSA/International School
Advanced Studies
Via Beirut 2-4
34014 Trieste, Italy
ferrara@sissa.it

Richard S. Ellis
Astronomy Department
California Institute of Technology
Pasadena, CA 91125, USA
rse@astro.caltech.edu

Volume Editors:

Daniel Schaerer
Angela Hempel
Observatoire de Genève
Université de Genève
Chemin des Maillettes 51
1290 Sauverny, Switzerland
daniel.schaerer@obs.unige.ch
angela.hempel@obs.unige.ch

Denis Puy
Université des Sciences Montpellier II
GRAAL CC72
34095 Montpellier cedex
France
denis.puy@graal.univ-montp2.fr

This series is edited on behalf of the Swiss Society for Astrophysics and Astronomy:
Société Suisse d'Astrophysique et d'Astronomie
Observatoire de Genève, ch. des Maillettes 51, 1290 Sauverny, Switzerland

Cover picture: First light – Artists view. Credit: NASA/WMAP Science Team.

Library of Congress Control Number: 2007937500

ISBN 978-3-540-74162-6 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springer.com

Typesetting: by the authors and Integra using a Springer LaTeX macro package
Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper SPIN: 12103294 55/Integra 5 4 3 2 1 0

Preface

The exploration of the first billion year of the history of the Universe, from the so-called dark ages to cosmic reionisation, represents one of the great challenges of contemporary astrophysics. During these phases the first structures start to grow forming the first stars, galaxies, and possibly also soon the first quasars. At the same time the dark, neutral Universe starts to be lit up and ionised by these source, leading to its progressive reionisation ending at redshift $z \sim 6$. Furthermore the first stars and supernovae begin to enrich their surroundings and the intergalactic medium, and to produce the first dust.

All these phenomena represent a rich interplay between various fields of astrophysics, which have seen important developments over recent years. Indeed tremendous progress has been made on the theoretical understanding and on numerical simulations. In addition, observations of signatures of reionisation and even direct observations of galaxies at $z > 6$ are now becoming feasible. Such observations actually provide a main driver for upcoming facilities such as ground-based multi-object spectrographs in the near-IR, extremely large telescopes, and for the James Webb Space Telescope.

Given these important achievements and the increasing developments made in this rapidly expanding field, the members of the Swiss Society for Astronomy and Astrophysics chose this topic for the 36th Saas-Fee advanced course. The course took place from 3 to 8 April 2006 in the Swiss Alps in Les Diablerets. Approximately 70 participants from a great diversity of countries could benefit from the excellent lectures delivered by Abraham Loeb, Andrea Ferrara and Richard Ellis, who kindly accepted this task. We wish to thank them here for all their work, including not only the lectures but also the chapters assembled in this book. Their knowledge, pedagogical talents, and enthusiasm have been essential for the success of this course.

We also thank our colleague Olivier Genevay for his technical help with the projection and computers. A special thanks goes to the course secretary, Ms Myriam Burgerner Frick, for all her help in the practical organisation of this course.

Geneva,
May 2007

Daniel Schaerer
Angela Hempel
Denis Puy

Contents

List of Previous Saas-Fee Advanced Courses

!! 2006 First Light in the Universe
A. Loeb, A. Ferrara. R.S. Ellis

!! 2005 Trans-Neptunian Objects and Comets
D. Jewitt, A. Morbidelli, H. Rauer

!! 2004 The Sun, Solar Analogs and the Climate
J.D. Haigh, M. Lockwood, M.S. Giampapa

!! 2003 Gravitation Lensing: Strong, Weak and Micro
P. Schneider, C. Kochanek, J. Wambsganss

!! 2002 The Cold Universe
A.W. Blain, F. Combes, B.T. Draine

!! 2001 Extrasolar Planets
T. Guillot, P. Cassen, A. Quirrenbach

!! 2000 High-Energy Spectroscopic Astrophysics
S.M. Kahn, P. von Ballmoos, R.A. Sunyaev

!! 1999 Physics of Star Formation in Galaxies
F. Palla, H. Zinnecker

!! 1998 Star Clusters
B.W. Carney, W.E. Harris

!! 1997 Computational Methods for Astrophysical Fluid Flow
R.J. LeVeque, D. Mihalas, E.A. Dorfi, E. Müller

!! 1996 Galaxies Interactions and Induced Star Formation
R.C. Kennicutt, F. Schweizer, J.E. Barnes

!! 1995 Stellar Remnants
S.D. Kawaler, I. Novikov, G. Srinivasan

!! 1994 Plasma Astrophysics
J.G. Kirk, D.B. Melrose, E.R. Priest

!! 1993 The Deep Universe
A.R. Sandage, R.G. Kron, M.S. Longair

!! 1992 Interacting Binaries
S.N. Shore, M. Livio, E.J.P. van den Heuvel

!! 1991 The Galactic Interstellar Medium
W.B. Burton, B.G. Elmegreen, R. Genzel

!! 1990 Active Galactic Nuclei
R. Blandford, H. Netzer, L. Woltjer

* 1989 The Milky Way as a Galaxy
G. Gilmore, I. King, P. van der Kruit

! 1988 Radiation in Moving Gaseous Media
H. Frisch, R.P. Kudritzki, H.W. Yorke

! 1987 Large Scale Structures in the Universe
A.C. Fabian, M. Geller, A. Szalay

! 1986 Nucleosynthesis and Chemical Evolution
J. Audouze, C. Chiosi, S.E. Woosley

! 1985 High Resolution in Astronomy
R.S. Booth, J.W. Brault, A. Labeyrie

! 1984 Planets, Their Origin, Interior and Atmosphere
D. Gautier, W.B. Hubbard, H. Reeves

! 1983 Astrophysical Processes in Upper Main Sequence Stars
A.N. Cox, S. Vauclair, J.P. Zahn

* 1982 Morphology and Dynamics of Galaxies
J. Binney, J. Kormendy, S.D.M. White

! 1981 Activity and Outer Atmospheres of the Sun and Stars
F. Praderie, D.S. Spicer, G.L. Withbroe

* 1980 Star Formation
J. Appenzeller, J. Lequeux, J. Silk

* 1979 Extragalactic High Energy Physics
F. Pacini, C. Ryter, P.A. Strittmatter

* 1978 Observational Cosmology
J.E. Gunn, M.S. Longair, M.J. Rees

* 1977 Advanced Stages in Stellar Evolution
I. Iben Jr., A. Renzini, D.N. Schramm

* 1976 Galaxies
K. Freeman, R.C. Larson, B. Tinsley

* 1975 Atomic and Molecular Processes in Astrophysics
A. Dalgarno, F. Masnou-Seeuws, R.V.P. McWhirter

* 1974 Magnetohydrodynamics
L. Mestel, N.O. Weiss

* 1973 Dynamical Structure and Evolution of Stellar Systems
G. Contopoulos, M. Hénon, D. Lynden-Bell

* 1972 Interstellar Matter
N.C. Wickramasinghe, F.D. Kahn, P.G. Metzger

* 1971 Theory of the Stellar Atmospheres
D. Mihalas, B. Pagel, P. Souffrin

---

* Out of print

! May be ordered from Geneva Observatory
Saas-Fee Courses
Geneva Observatory
CH-1290 Sauverny
Switzerland

!! May be ordered from Springer and/or are available online at springerlink.com.

First Light

A. Loeb

Abstract. The first dwarf galaxies, which constitute the building blocks of the collapsed objects we find today in the Universe, had formed hundreds of millions of years after the big bang. This pedagogical review describes the early growth of their small-amplitude seed fluctuations from the epoch of inflation through dark matter decoupling and matter-radiation equality, to the final collapse and fragmentation of the dark matter on all mass scales above $\sim 10^{-4}$ $M_\odot$. The condensation of baryons into halos in the mass range of $\sim 10^5$–10^{10} $M_\odot$ led to the formation of the first stars and the re-ionization of the cold hydrogen gas, left over from the Big Bang. The production of heavy elements by the first stars started the metal enrichment process that eventually led to the formation of rocky planets and life.

A wide variety of instruments currently under design [including large-aperture infrared telescopes on the ground or in space (*JWST*), and low-frequency arrays for the detection of redshifted 21 cm radiation], will establish better understanding of the first sources of light during an epoch in cosmic history that was largely unexplored so far. Numerical simulations of reionization are computationally challenging, as they require radiative transfer across large cosmological volumes as well as sufficently high resolution to identify the sources of the ionizing radiation. The technological challenges for observations and the computational challenges for numerical simulations, will motivate intense work in this field over the coming decade.

1 Opening Remarks

When I open the daily newspaper as part of my morning routine, I often see lengthy descriptions of conflicts between people on borders, properties, or liberties. Today's news is often forgotten a few days later. But when one opens ancient texts that have appealed to a broad audience over a longer period of time, such as the Bible, what does one often find in the opening chapter?... a discussion of how the constituents of the Universe (including light, stars and life) were created. Although humans are often occupied with mundane problems, they are curious about the big picture. As citizens of the Universe, we cannot help but wonder how the first sources of light formed,

A. Loeb: *First Light*. In: D. Schaerer et al.: First Light in the Universe, Saas-Fee Advanced Courses, pp. 1–159 (2008)
DOI 10.1007/978-3-540-74163-3_1

how life came to existence, and whether we are alone as intelligent beings in this vast space. As astronomers in the twenty first century, we are uniquely positioned to answer these big questions with scientific instruments and a quantitative methodology. In this pedagogical review, intended for students preparing to specialize in cosmology, I will describe current ideas about one of these topics: *the appearance of the first sources of light and their influence on the surrounding Universe*. This topic is one of the most active frontiers in present-day cosmology. As such it is an excellent area for a PhD thesis of a graduate student interested in cosmology. I will therefore highlight the unsolved questions in this field as much as the bits we understand.

2 Excavating the Universe for Clues About Its History

When we look at our image reflected off a mirror at a distance of 1 m, we see the way we looked 6 nano-seconds ago, the light travel time to the mirror and back. If the mirror is spaced 10^{19} cm = 3 pc away, we will see the way we looked 21 years ago. Light propagates at a finite speed, and so by observing distant regions, we are able to see how the Universe looked like in the past, a light travel time ago (Fig. 1). The statistical homogeneity of the Universe on large scales guarantees that what we see far away is a fair statistical representation of the conditions that were present in our region of the Universe a long time ago.

This fortunate situation makes cosmology an empirical science. We do not need to guess how the Universe evolved. Using telescopes we can simply see

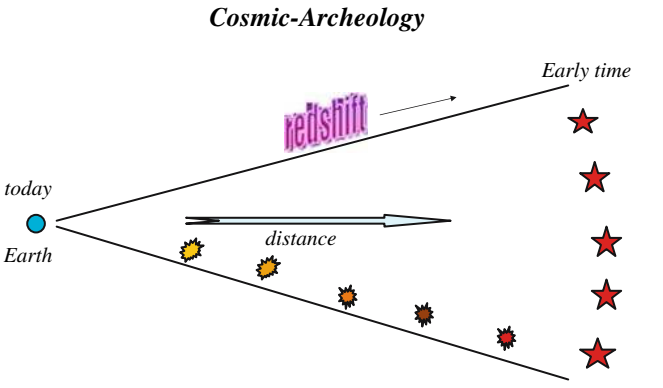

Fig. 1. Cosmology is like archeology. The deeper one looks, the older is the layer that one is revealing, owing to the finite propagation speed of light. (Figure from Loeb 2007 [218].)

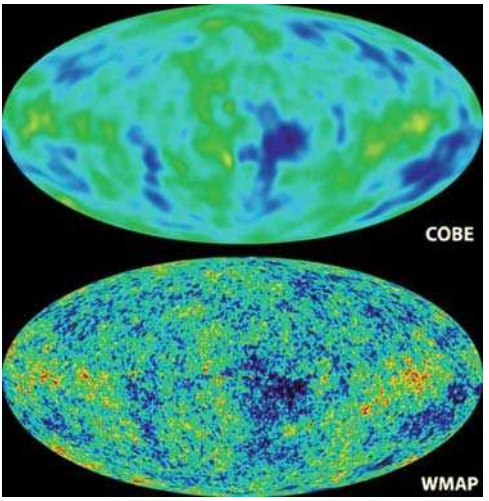

Fig. 2. Images of the Universe shortly after it became transparent, taken by the *COBE* and *WMAP* satellites (see http://map.gsfc.nasa.gov/ for details). The slight density inhomogeneties in the otherwise uniform Universe are imprinted in the hot and cold brightness map of the cosmic microwave background. The existence of these anisotropies was predicted three decades before the technology for taking this image became available in a number of theoretical papers, including (Sunyaev & Zelclovich 1970 [354], Sachs & Wolfe 1967 [307], Roos & Sciama 1968 [296], Silk 1968 [337], Peebles & Yu 1970 [281])

the way it appeared at earlier cosmic times. Since a greater distance means a fainter flux from a source of a fixed luminosity, the observation of the earliest sources of light requires the development of sensitive instruments and poses challenges to observers.

We can in principle image the Universe only if it is transparent. Earlier than 0.4 million years after the Big Bang, the cosmic plasma was ionized and the Universe was opaque to Thomson scattering by the dense gas of free electrons that filled it. Thus, telescopes cannot be used to image the infant Universe at earlier times (or redshifts $\gtrsim 10^3$). The earliest possible image of the Universe was recorded by COBE and WMAP (see Fig. 2).

3 Background Cosmological Model

3.1 The Expanding Universe

The modern physical description of the Universe as a whole can be traced back to Einstein, who argued theoretically for the so-called "cosmological principle": that the distribution of matter and energy must be homogeneous

and isotropic on the largest scales. Today isotropy is well established (see the review by Wu et al. 1999 [388]) for the distribution of faint radio sources, optically-selected galaxies, the X-ray background, and most importantly the cosmic microwave background (hereafter, CMB; see, e.g., Bennett et al. 1996 [36]). The constraints on homogeneity are less strict, but a cosmological model in which the Universe is isotropic but significantly inhomogeneous in spherical shells around our special location, is also excluded [155].

In General Relativity, the metric for a space which is spatially homogeneous and isotropic is the Friedman-Robertson-Walker metric, which can be written in the form

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)\left[\frac{\mathrm{d}R^2}{1-k\,R^2} + R^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right] , \tag{1}$$

where $a(t)$ is the cosmic scale factor which describes expansion in time, and (R, θ, ϕ) are spherical comoving coordinates. The constant k determines the geometry of the metric; it is positive in a closed Universe, zero in a flat Universe, and negative in an open Universe. Observers at rest remain at rest, at fixed (R, θ, ϕ), with their physical separation increasing with time in proportion to $a(t)$. A given observer sees a nearby observer at physical distance D receding at the Hubble velocity $H(t)D$, where the Hubble constant at time t is $H(t) = \mathrm{d}\,a(t)/\mathrm{d}t$. Light emitted by a source at time t is observed at $t = 0$ with a redshift $z = 1/a(t) - 1$, where we set $a(t = 0) \equiv 1$ for convenience (but note that old textbooks may use a different convention).

The Einstein field equations of General Relativity yield the Friedmann equation (e.g., Weinberg 1972 [375]; Kolb & Turner 1990 [203])

$$H^2(t) = \frac{8\pi G}{3}\rho - \frac{k}{a^2} , \tag{2}$$

which relates the expansion of the Universe to its matter-energy content. For each component of the energy density ρ, with an equation of state $p = p(\rho)$, the density ρ varies with $a(t)$ according to the equation of energy conservation

$$\mathrm{d}(\rho R^3) = -p\mathrm{d}(R^3) . \tag{3}$$

With the critical density

$$\rho_{\mathrm{C}}(t) \equiv \frac{3H^2(t)}{8\pi G} \tag{4}$$

defined as the density needed for $k = 0$, we define the ratio of the total density to the critical density as

$$\Omega \equiv \frac{\rho}{\rho_{\mathrm{C}}} . \tag{5}$$

With Ω_{m}, Ω_Λ, and Ω_{r} denoting the present contributions to Ω from matter (including cold dark matter as well as a contribution Ω_{b} from baryons), vacuum density (cosmological constant), and radiation, respectively, the Friedmann equation becomes

$$\frac{H(t)}{H_0} = \left[\frac{\Omega_{\rm m}}{a^3} + \Omega_\Lambda + \frac{\Omega_{\rm r}}{a^4} + \frac{\Omega_{\rm k}}{a^2} \right] , \tag{6}$$

where we define H_0 and $\Omega_0 = \Omega_{\rm m} + \Omega_\Lambda + \Omega_{\rm r}$ to be the present values of H and Ω, respectively, and we let

$$\Omega_k \equiv -\frac{k}{H_0^2} = 1 - \Omega_{\rm m}. \tag{7}$$

In the particularly simple Einstein-de Sitter model ($\Omega_{\rm m} = 1$, $\Omega_\Lambda = \Omega_{\rm r} = \Omega_{\rm k} = 0$), the scale factor varies as $a(t) \propto t^{2/3}$. Even models with non-zero Ω_Λ or Ω_k approach the Einstein-de Sitter behavior at high redshift, i.e. when $(1+z) \gg |{\Omega_{\rm m}}^{-1} - 1|$ (as long as $\Omega_{\rm r}$ can be neglected). In this high-z regime the age of the Universe is,

$$t \approx \frac{2}{3H_0\sqrt{\Omega_{\rm m}}} (1+z)^{-3/2} . \tag{8}$$

The Friedmann equation implies that models with $\Omega_{\rm k} = 0$ converge to the Einstein-de Sitter limit faster than do open models.

In the standard hot Big Bang model, the Universe is initially hot and the energy density is dominated by radiation. The transition to matter domination occurs at $z \sim 3500$, but the Universe remains hot enough that the gas is ionized, and electron-photon scattering effectively couples the matter and radiation. At $z \sim 1100$ the temperature drops below $\sim 3000\,{\rm K}$ and protons and electrons recombine to form neutral hydrogen (Fig. 3). The photons then decouple and travel freely until the present, when they are observed as the CMB [347].

3.2 Composition of the Universe

According to the standard cosmological model, the Universe started at the Big Bang about 14 billion years ago. During an early epoch of accelerated superluminal expansion, called inflation, a region of microscopic size was stretched to a scale much bigger than the visible Universe and our local geometry became flat. At the same time, primordial density fluctuations were generated out of quantum mechanical fluctuations of the vacuum. These inhomogeneities seeded the formation of present-day structure through the process of gravitational instability. The mass density of ordinary (baryonic) matter makes up only a fifth of the matter that led to the emergence of structure and the rest is the form of an unknown dark matter component. Recently, the Universe entered a new phase of accelerated expansion due to the dominance of some dark vacuum energy density over the ever rarefying matter density.
The basic question that cosmology attempts to answer is:
What are the ingredients (composition and initial conditions) of the Universe and what processes generated the observed structures in it?

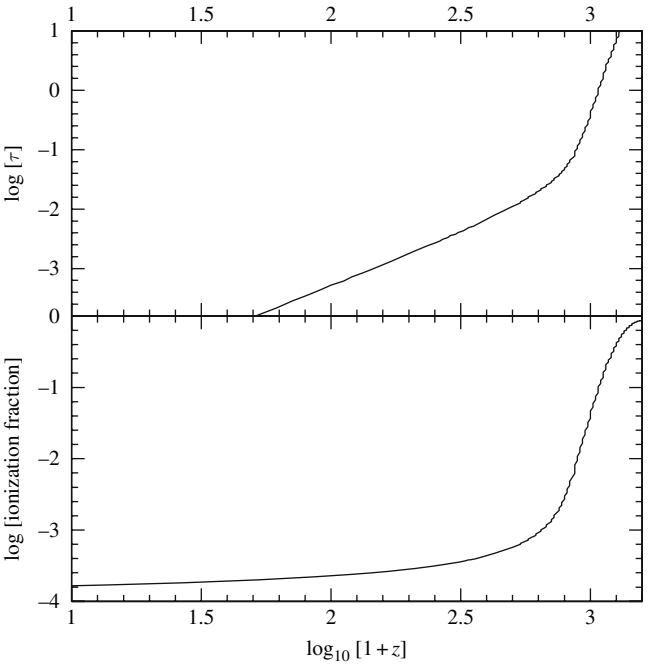

Fig. 3. The optical depth of the Universe to electron scattering (**upper panel**) and the ionization fraction (**lower panel**) as a function of redshift before reionization. Observatories of electromagnetic radiation cannot image the opaque Universe beyond a redshift of $z \sim 1100$

In detail, we would like to know:

(a) Did inflation occur and when? If so, what drove it and how did it end?
(b) What is the nature of the dark energy and how does it change over time and space?
(c) What is the nature of the dark matter and how did it regulate the evolution of structure in the Universe?

Before hydrogen recombined, the Universe was opaque to electromagnetic radiation, precluding any possibility for direct imaging of its evolution. The only way to probe inflation is through the fossil record that it left behind in the form of density perturbations and gravitational waves. Following inflation, the Universe went through several other milestones which left a detectable record. These include: baryogenesis (which resulted in the observed asymmetry between matter and anti-matter), the electroweak phase transition (during which the symmetry between electromagnetic and weak interactions was broken), the QCD (quantum chromodynamics) phase transition (during which protons and neutrons were assembled out of quarks and gluons), the dark matter freeze-out epoch (during which the dark matter decoupled from the cosmic plasma), neutrino decoupling, electron-positron annihilation, and

light-element nucleosynthesis (during which helium, deuterium and lithium were synthesized). The signatures that these processes left in the Universe can be used to constrain its parameters and answer the above questions.

Half a million years after the Big Bang, hydrogen recombined and the Universe became transparent. The ultimate goal of observational cosmology is to image the entire history of the Universe since then. Currently, we have a snapshot of the Universe at recombination from the CMB, and detailed images of its evolution starting from an age of a billion years until the present time. The evolution between a million and a billion years has not been imaged as of yet.

Within the next decade, NASA plans to launch an infrared space telescope (JWST) that will image the very first sources of light (stars and black holes) in the Universe, which are predicted theoretically to have formed in the first hundreds of millions of years. In parallel, there are several initiatives to construct large-aperture infrared telescopes on the ground with the same goal in mind[1,2,3]. The neutral hydrogen, relic from cosmological recombination, can be mapped in three-dimensions through its 21 cm line even before the first galaxies formed [225]. Several groups are currently constructing low-frequency radio arrays in an attempt to map the initial inhomogeneities as well as the process by which the hydrogen was re-ionized by the first galaxies.

The next generation of ground-based telescopes will have a diameter of 20–30 m (cf. Fig. 5). Together with JWST (that will not be affected by the atmospheric backgound, Fig. 4) they will be able to image the first galaxies. Given that these galaxies also created the ionized bubbles around them, the same galaxy locations should correlate with bubbles in the neutral hydrogen (created by their UV emission). Within a decade it would be possible to explore the environmental influence of individual galaxies by using the two sets of instruments in concert [389].

The dark ingredients of the Universe can only be probed indirectly through a variety of luminous tracers. The distribution and nature of the dark matter are constrained by detailed X-ray and optical observations of galaxies and galaxy clusters. The evolution of the dark energy with cosmic time will be constrained over the coming decade by surveys of Type Ia supernovae, as well as surveys of X-ray clusters, up to a redshift of two.

On large scales ($\gtrsim 10$ Mpc) the power-spectrum of primordial density perturbations is already known from the measured microwave background anisotropies, galaxy surveys, weak lensing, and the Lyα forest. Future programs will refine current knowledge, and will search for additional trademarks of inflation, such as gravitational waves (through CMB polarization), small-scale structure (through high-redshift galaxy surveys and 21 cm studies), or the Gaussian statistics of the initial perturbations.

[1] http://www.eso.org/projects/owl/
[2] http://celt.ucolick.org/
[3] http://www.gmto.org/

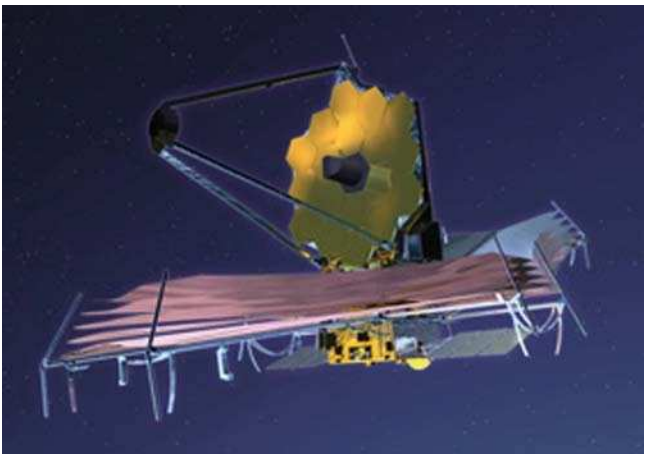

Fig. 4. A sketch of the current design for the *James Webb Space Telescope*, the successor to the *Hubble Space Telescope* to be launched in 2011 (http://www.jwst.nasa.gov/). The current design includes a primary mirror made of beryllium which is 6.5 m in diameter as well as an instrument sensitivity that spans the full range of infrared wavelengths of 0.6–28 μm and will allow detection of the first galaxies in the infant Universe. The telescope will orbit 1.5 million km from Earth at the Lagrange L2 point

Fig. 5. Artist conception of the design for one of the future giant telescopes that could probe the first generation of galaxies from the ground. The *Giant Magellan Telescope* (GMT) will contain seven mirrors (each 8.4 m in diameter) and will have the resolving power equivalent to a 24.5 m (80 ft) primary mirror. For more details see http://www.gmto.org/

The Big Bang is the only known event where particles with energies approaching the Planck scale $[(\hbar c^5/G)^{1/2} \sim 10^{19}\,\mathrm{GeV}]$ interacted. It therefore offers prospects for probing the unification physics between quantum mechanics and general relativity (to which string theory is the most-popular candidate). Unfortunately, the exponential expansion of the Universe during inflation erases memory of earlier cosmic epochs, such as the Planck time.

3.3 Linear Gravitational Growth

Observations of the CMB (e.g., Bennett et al. 1996 [36]) show that the Universe at recombination was extremely uniform, but with spatial fluctuations in the energy density and gravitational potential of roughly one part in 10^5. Such small fluctuations, generated in the early Universe, grow over time due to gravitational instability, and eventually lead to the formation of galaxies and the large-scale structure observed in the present Universe.

As before, we distinguish between fixed and comoving coordinates. Using vector notation, the fixed coordinate $\mathbf{r}$ corresponds to a comoving position $\mathbf{x} = \mathbf{r}/a$. In a homogeneous Universe with density ρ, we describe the cosmological expansion in terms of an ideal pressureless fluid of particles each of which is at fixed $\mathbf{x}$, expanding with the Hubble flow $\mathbf{v} = H(t)\mathbf{r}$ where $\mathbf{v} = \mathrm{d}\mathbf{r}/\mathrm{d}t$. Onto this uniform expansion we impose small perturbations, given by a relative density perturbation

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{r})}{\bar{\rho}} - 1 \ , \tag{9}$$

where the mean fluid density is $\bar{\rho}$, with a corresponding peculiar velocity $\mathbf{u} \equiv \mathbf{v} - H\mathbf{r}$. Then the fluid is described by the continuity and Euler equations in comoving coordinates [282, 283]:

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla \cdot [(1+\delta)\mathbf{u}] = 0 \tag{10}$$

$$\frac{\partial \mathbf{u}}{\partial t} + H\,\mathbf{u} + \frac{1}{a}(\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{a}\nabla\phi \ . \tag{11}$$

The potential ϕ is given by the Poisson equation, in terms of the density perturbation:

$$\nabla^2\phi = 4\pi G\bar{\rho}a^2\delta \ . \tag{12}$$

This fluid description is valid for describing the evolution of collisionless cold dark matter particles until different particle streams cross. This "shell-crossing" typically occurs only after perturbations have grown to become non-linear, and at that point the individual particle trajectories must in general be followed. Similarly, baryons can be described as a pressureless fluid as long as their temperature is negligibly small, but non-linear collapse leads to the formation of shocks in the gas.

For small perturbations $\delta \ll 1$, the fluid equations can be linearized and combined to yield

$$\frac{\partial^2\delta}{\partial t^2}+2H\frac{\partial\delta}{\partial t}=4\pi G\bar{\rho}\delta\ . \tag{13}$$

This linear equation has in general two independent solutions, only one of which grows with time. Starting with random initial conditions, this "growing mode" comes to dominate the density evolution. Thus, until it becomes nonlinear, the density perturbation maintains its shape in comoving coordinates and grows in proportion to a growth factor $D(t)$. The growth factor in the matter-dominated era is given by [282]

$$D(t)\propto\frac{\left(\Omega_\Lambda a^3+\Omega_{\rm k}a+\Omega_{\rm m}\right)^{1/2}}{a^{3/2}}\int_0^a\frac{a'^{3/2}\,{\rm d}a'}{\left(\Omega_\Lambda a'^3+\Omega_{\rm k}a'+\Omega_{\rm m}\right)^{3/2}}\ , \tag{14}$$

where we neglect $\Omega_{\rm r}$ when considering halos forming in the matter-dominated regime at $z\ll 10^4$. In the Einstein-de Sitter model (or, at high redshift, in other models as well) the growth factor is simply proportional to $a(t)$.

The spatial form of the initial density fluctuations can be described in Fourier space, in terms of Fourier components

$$\delta_{\mathbf{k}}=\int{\rm d}^3x\,\delta(x){\rm e}^{-{\rm i}\mathbf{k}\cdot\mathbf{x}}\ . \tag{15}$$

Here we use the comoving wavevector $\mathbf{k}$, whose magnitude k is the comoving wavenumber which is equal to 2π divided by the wavelength. The Fourier description is particularly simple for fluctuations generated by inflation (e.g., Kolb & Turner 1990 [203]). Inflation generates perturbations given by a Gaussian random field, in which different $\mathbf{k}$-modes are statistically independent, each with a random phase. The statistical properties of the fluctuations are determined by the variance of the different $\mathbf{k}$-modes, and the variance is described in terms of the power spectrum $P(k)$ as follows:

$$\langle\delta_{\mathbf{k}}\delta^*_{\mathbf{k}'}\rangle=(2\pi)^3\,P(k)\,\delta^{(3)}\left(\mathbf{k}-\mathbf{k}'\right)\ , \tag{16}$$

where $\delta^{(3)}$ is the three-dimensional Dirac delta function. The gravitational potential fluctuations are sourced by the density fluctuations through Poisson's equation.

In standard models, inflation produces a primordial power-law spectrum $P(k)\propto k^n$ with $n\sim 1$. Perturbation growth in the radiation-dominated and then matter-dominated Universe results in a modified final power spectrum, characterized by a turnover at a scale of order the horizon cH^{-1} at matter-radiation equality, and a small-scale asymptotic shape of $P(k)\propto k^{n-4}$. The overall amplitude of the power spectrum is not specified by current models of inflation, and it is usually set by comparing to the observed CMB temperature fluctuations or to local measures of large-scale structure.

Since density fluctuations may exist on all scales, in order to determine the formation of objects of a given size or mass it is useful to consider the statistical distribution of the smoothed density field. Using a window function $W(\mathbf{r})$

normalized so that $\int \mathrm{d}^3 r\, W(\mathbf{r}) = 1$, the smoothed density perturbation field, $\int \mathrm{d}^3 r \delta(\mathbf{x}) W(\mathbf{r})$, itself follows a Gaussian distribution with zero mean. For the particular choice of a spherical top-hat, in which $W = 1$ in a sphere of radius R and is zero outside, the smoothed perturbation field measures the fluctuations in the mass in spheres of radius R. The normalization of the present power spectrum is often specified by the value of $\sigma_8 \equiv \sigma(R = 8\,h^{-1}\,\mathrm{Mpc})$. For the top-hat, the smoothed perturbation field is denoted δ_R or δ_M, where the mass M is related to the comoving radius R by $M = 4\pi\rho_\mathrm{m} R^3/3$, in terms of the current mean density of matter ρ_m. The variance $\langle \delta_M \rangle^2$ is

$$\sigma^2(M) = \sigma^2(R) = \int_0^\infty \frac{\mathrm{d}k}{2\pi^2}\, k^2 P(k) \left[\frac{3 j_1(kR)}{kR}\right]^2 \,, \tag{17}$$

where $j_1(x) = (\sin x - x \cos x)/x^2$. The function $\sigma(M)$ plays a crucial role in estimates of the abundance of collapsed objects, as we describe later.

Species that decouple from the cosmic plasma (like the dark matter or the baryons) would show fossil evidence for acoustic oscillations in their power spectrum of inhomogeneities due to sound waves in the radiation fluid to which they were coupled at early times. This phenomenon can be understood as follows. Imagine a localized point-like perturbation from inflation at $t = 0$. The small perturbation in density or pressure will send out a sound wave that will reach the sound horizon $c_s t$ at any later time t. The perturbation will therefore correlate with its surroundings up to the sound horizon and all k-modes with wavelengths equal to this scale or its harmonics will be correlated. The scales of the perturbations that grow to become the first collapsed objects at $z < 100$ cross the horizon in the radiation dominated era after the dark matter decouples from the cosmic plasma. Next we consider the imprint of this decoupling on the smallest-scale structure of the dark matter.

3.4 The Smallest-Scale Power Spectrum of Cold Dark Matter

A broad range of observational data involving the dynamics of galaxies, the growth of large-scale structure, and the dynamics and nucleosynthesis of the Universe as a whole, indicate the existence of dark matter with a mean cosmic mass density that is $\sim$5 times larger than the density of the baryonic matter [187, 347]. The data is consistent with a dark matter composed of weakly-interacting, massive particles, that decoupled early and adiabatically cooled to an extremely low temperature by the present time [187]. The Cold Dark Matter (CDM) has not been observed directly as of yet, although laboratory searches for particles from the dark halo of our own Milky-Way galaxy have been able to restrict the allowed parameter space for these particles. Since an alternative more-radical interpretation of the dark matter phenomenology involves a modification of gravity [252], it is of prime importance to find direct fingerprints of the CDM particles. One such fingerprint involves the small-scale structure in the Universe [158], on which we focus in this section.

The most popular candidate for the CDM particle is a Weakly Interacting Massive Particle (WIMP). The lightest supersymmetric particle (LSP) could be a WIMP (for a review see [187]). The CDM particle mass depends on free parameters in the particle physics model but typical values cover a range around $M \sim 100\,\mathrm{GeV}$ (up to values close to a TeV). In many cases the LSP hypothesis will be tested at the Large Hadron Collider (e.g. [33]) or in direct detection experiments (e.g. [16]).

The properties of the CDM particles affect their response to the small-scale primordial inhomogeneities produced during cosmic inflation. The particle cross-section for scattering off standard model fermions sets the epoch of their thermal and kinematic decoupling from the cosmic plasma (which is significantly later than the time when their abundance freezes-out at a temperature $T \sim M$). Thermal decoupling is defined as the time when the temperature of the CDM stops following that of the cosmic plasma while kinematic decoupling is defined as the time when the bulk motion of the two species start to differ. For CDM the epochs of thermal and kinetic decoupling coincide. They occur when the time it takes for collisions to change the momentum of the CDM particles equals the Hubble time. The particle mass determines the thermal spread in the speeds of CDM particles, which tends to smooth-out fluctuations on very small scales due to the free-streaming of particles after kinematic decoupling [158, 159]. Viscosity has a similar effect before the CDM fluid decouples from the cosmic radiation fluid [180]. An important effect involves the memory the CDM fluid has of the acoustic oscillations of the cosmic radiation fluid out of which it decoupled. Here we consider the imprint of these acoustic oscillations on the small-scale power spectrum of density fluctuations in the Universe. Analogous imprints of acoustic oscillations of the baryons were identified recently in maps of the CMB [347], and the distribution of nearby galaxies [119]; these signatures appear on much larger scales, since the baryons decouple much later when the scale of the horizon is larger. The discussion in this section follows Loeb & Zaldarriaga (2005) [227].

Formalism

Kinematic decoupling of CDM occurs during the radiation-dominated era. For example, if the CDM is made of neutralinos with a particle mass of $\sim$100 GeV, then kinematic decoupling occurs at a cosmic temperature of $T_{\rm d} \sim 10\,\mathrm{MeV}$ [180, 87]. As long as $T_{\rm d} \ll 100\,\mathrm{MeV}$, we may ignore the imprint of the QCD phase transition (which transformed the cosmic quark-gluon soup into protons and neutrons) on the CDM power spectrum [320]. Over a short period of time during this transition, the pressure does not depend on density and the sound speed of the plasma vanishes, resulting in a significant growth for perturbations with periods shorter than the length of time over which the sound speed vanishes. The transition occurs when the temperature of the cosmic plasma is $\sim$ 100–200 MeV and lasts for a small fraction of the Hubble time. As a result, the induced modifications are on scales smaller than those we

are considering here and the imprint of the QCD phase transition is washed-out by the effects we calculate.

At early times the contribution of the dark matter to the energy density is negligible. Only at relatively late times when the cosmic temperature drops to values as low as $\sim 1\,\mathrm{eV}$, matter and radiation have comparable energy densities. As a result, the dynamics of the plasma at earlier times is virtually unaffected by the presence of the dark matter particles. In this limit, the dynamics of the radiation determines the gravitational potential and the dark matter just responds to that potential. We will use this simplification to obtain analytic estimates for the behavior of the dark matter transfer function.

The primordial inflationary fluctuations lead to acoustic modes in the radiation fluid during this era. The interaction rate of the particles in the plasma is so high that we can consider the plasma as a perfect fluid down to a comoving scale,

$$\lambda_{\mathrm{f}} \sim \eta_{\mathrm{d}}/\sqrt{N} \quad ; \quad N \sim n\sigma t_{\mathrm{d}}, \tag{18}$$

where $\eta_{\mathrm{d}} = \int_0^{t_{\mathrm{d}}} \mathrm{d}t/a(t)$ is the conformal time (i.e. the comoving size of the horizon) at the time of CDM decoupling, t_{d}; σ is the scattering cross section and n is the relevant particle density. (Throughout this section we set the speed of light and Planck's constant to unity for brevity.) The damping scale depends on the species being considered and its contribution to the energy density, and is the largest for neutrinos which are only coupled through weak interactions. In that case $N \sim (T/T_{\mathrm{d}}^{\nu})^3$ where $T_{\mathrm{d}}^{\nu} \sim 1\,\mathrm{MeV}$ is the temperature of neutrino decoupling. At the time of CDM decoupling $N \sim M/T_{\mathrm{d}} \sim 10^4$ for the rest of the plasma, where M is the mass of the CDM particle. Here we will consider modes of wavelength larger than λ_{f}, and so we neglect the effect of radiation diffusion damping and treat the plasma (without the CDM) as a perfect fluid.

The equations of motion for a perfect fluid during the radiation era can be solved analytically. We will use that solution here, following the notation of Dodelson [109]. As usual we Fourier decompose fluctuations and study the behavior of each Fourier component separately. For a mode of comoving wavenumber k in Newtonian gauge, the gravitational potential fluctuations are given by:

$$\Phi = 3\Phi_{\mathrm{p}} \left[\frac{\sin(\omega\eta) - \omega\eta\cos(\omega\eta)}{(\omega\eta)^3} \right], \tag{19}$$

where $\omega = k/\sqrt{3}$ is the frequency of a mode and Φ_{p} is its primordial amplitude in the limit $\eta \to 0$. In this section we use conformal time $\eta = \int \mathrm{d}t/a(t)$ with $a(t) \propto t^{1/2}$ during the radiation-dominated era. Expanding the temperature anisotropy in multipole moments and using the Boltzmann equation to describe their evolution, the monopole Θ_0 and dipole Θ_1 of the photon distribution can be written in terms of the gravitational potential as [109]:

$$\Theta_0 = \Phi\left(\frac{x^2}{6}+\frac{1}{2}\right)+\frac{x}{2}\Phi'$$
$$\Theta_1 = -\frac{x^2}{6}\left(\Phi'+\frac{1}{x}\Phi\right) \tag{20}$$

where $x \equiv k\eta$ and a prime denotes a derivative with respect to x.

The solutions in (19) and (20) assume that both the sound speed and the number of relativistic degrees of freedom are constant over time. As a result of the QCD phase transition and of various particles becoming non-relativistic, both of these assumptions are not strictly correct. The vanishing sound speed during the QCD phase transition provides the most dramatic effect, but its imprint is on scales smaller than the ones we consider here because the transition occurs at a significantly higher temperature and only lasts for a fraction of the Hubble time [320].

Before the dark matter decouples kinematically, we will treat it as a fluid which can exchange momentum with the plasma through particle collisions. At early times, the CDM fluid follows the motion of the plasma and is involved in its acoustic oscillations. The continuity and momentum equations for the CDM can be written as:

$$\dot{\delta}_{\rm c}+\theta_{\rm c} = 3\dot{\Phi}$$
$$\dot{\theta}_{\rm c}+\frac{\dot{a}}{a}\theta_{\rm c} = k^2 c_{\rm s}^2\delta_{\rm c}-k^2\sigma_{\rm c}-k^2\Phi+\tau_{\rm c}^{-1}(\Theta_1-\theta_{\rm c}) \tag{21}$$

where a dot denotes an η-derivative, $\delta_{\rm c}$ is the dark matter density perturbation, $\theta_{\rm c}$ is the divergence of the dark matter velocity field and $\sigma_{\rm c}$ denotes the anisotropic stress. In writing these equations we have followed [229]. The term $\tau_{\rm c}^{-1}(\Theta_1-\theta_{\rm c})$ encodes the transfer of momentun between the radiation and CDM fluids and $\tau_{\rm c}^{-1}$ provides the collisional rate of momentum transfer,

$$\tau_{\rm c}^{-1} = n\sigma\frac{T}{M}a, \tag{22}$$

with n being the number density of particles with which the dark matter is interacting, $\sigma(T)$ the average cross section for interaction and M the mass of the dark matter particle. The relevant scattering partners are the standard model leptons which have thermal abundances. For detailed expressions of the cross section in the case of supersymmetric (SUSY) dark matter, see [87, 159]. For our purpose, it is sufficient to specify the typical size of the cross section and its scaling with cosmic time,

$$\sigma \approx \frac{T^2}{M_\sigma^4}, \tag{23}$$

where the coupling mass M_σ is of the order of the weak-interaction scale ($\sim 100\,{\rm GeV}$) for SUSY dark matter. This equation should be taken as the definition of M_σ, as it encodes all the uncertainties in the details of the particle

physics model into a single parameter. The temperature dependance of the averaged cross section is a result of the available phase space. Our results are quite insensitive to the details other than through the decoupling time. Equating $\tau_{\rm c}^{-1}/a$ to the Hubble expansion rate gives the temperature of kinematic decoupling:

$$T_{\rm d} = \left(\frac{M_\sigma^4 M}{M_{\rm pl}}\right)^{1/4} \approx 10\ {\rm MeV} \left(\frac{M_\sigma}{100\ {\rm GeV}}\right) \left(\frac{M}{100\ {\rm GeV}}\right)^{1/4}. \qquad (24)$$

The term $k^2 c_{\rm s}^2 \delta_{\rm c}$ in (21) results from the pressure gradient force and c_s is the dark matter sound speed. In the tight coupling limit, $\tau_{\rm c} \ll H^{-1}$ we find that $c_{\rm s}^2 \approx f_{\rm c} T/M$ and that the shear term is $k^2\sigma_{\rm c} \approx f_{\rm v} c_{\rm s}^2 \tau_{\rm c} \theta_{\rm c}$. Here $f_{\rm v}$ and $f_{\rm c}$ are constant factors of order unity. We will find that both these terms make a small difference on the scales of interest, so their precise value is unimportant.

By combining both in (21) into a single equation for δ_c we get

$$\begin{aligned} &\delta_{\rm c}'' + \frac{1}{x}\left[1 + F_{\rm v}(x)\right]\delta_{\rm c}' + c_{\rm s}^2(x)\delta_{\rm c} \\ &= S(x) - 3F_{\rm v}(x)\Phi' + \frac{x_{\rm d}^4}{x^5}\left(3\theta_0' - \delta_{\rm c}'\right), \end{aligned} \qquad (25)$$

where $x_{\rm d} = k\eta_{\rm d}$ and $\eta_{\rm d}$ denotes the time of kinematic decoupling which can be expressed in terms of the decoupling temperature as,

$$\begin{aligned} \eta_{\rm d} = 2t_{\rm d}(1+z_{\rm d}) \approx \frac{M_{\rm pl}}{T_0 T_{\rm d}} &\approx 10\ {\rm pc}\ \left(\frac{T_{\rm d}}{10\ {\rm MeV}}\right)^{-1} \\ &\propto M_\sigma^{-1} M^{-1/4}, \end{aligned} \qquad (26)$$

with $T_0 = 2.7\,{\rm K}$ being the present-day CMB temperature and $z_{\rm d}$ being the redshift at kinematic decoupling. We have also introduced the source function,

$$S(x) \equiv -3\Phi'' + \Phi - \frac{3}{x}\Phi'. \qquad (27)$$

For $x \ll x_{\rm d}$, the dark matter sound speed is given by

$$c_{\rm s}^2(x) = c_{\rm s}^2(x_{\rm d})\frac{x_{\rm d}}{x}, \qquad (28)$$

where $c_{\rm s}^2(x_{\rm d})$ is the dark matter sound speed at kinematic decoupling (in units of the speed of light),

$$c_{\rm s}(x_{\rm d}) \approx 10^{-2} f_{\rm c}^{1/2} \left(\frac{T_{\rm d}}{10\,{\rm MeV}}\right)^{1/2} \left(\frac{M}{100\,{\rm GeV}}\right)^{-1/2}. \qquad (29)$$

In writing (28) we have assumed that prior to decoupling the temperature of the dark matter follows that of the plasma. For the viscosity term we have,

$$F_{\rm v}(x) = f_{\rm v} c_{\rm s}^2(x_{\rm d}) x_{\rm d}^2 \left(\frac{x_{\rm d}}{x}\right)^5. \qquad (30)$$

Free Streaming After Kinematic Decoupling

In the limit of the collision rate being much slower than the Hubble expansion, the CDM is decoupled and the evolution of its perturbations is obtained by solving a Boltzman equation:

$$\frac{\partial f}{\partial \eta} + \frac{\mathrm{d}r_\mathrm{i}}{\mathrm{d}\eta}\frac{\partial f}{\partial r_\mathrm{i}} + \frac{\mathrm{d}q_\mathrm{i}}{\mathrm{d}\eta}\frac{\partial f}{\partial q_\mathrm{i}} = 0, \tag{31}$$

where $f(\mathbf{r}, \mathbf{q}, \eta)$ is the distribution function which depends on position, comoving momentum $\mathbf{q}$, and time. The comoving momentum 3-components are $\mathrm{d}x_\mathrm{i}/\mathrm{d}\eta = q_\mathrm{i}/a$. We use the Boltzman equation to find the evolution of modes that are well inside the horizon with $x \gg 1$. In the radiation era, the gravitational potential decays after horizon crossing (see 19). In this limit the comoving momentum remains constant, $\mathrm{d}q_\mathrm{i}/\mathrm{d}\eta = 0$ and the Boltzman equation becomes,

$$\frac{\partial f}{\partial \eta} + \frac{q_\mathrm{i}}{a}\frac{\partial f}{\partial r_\mathrm{i}} = 0. \tag{32}$$

We consider a single Fourier mode and write f as,

$$f(\mathbf{r}, \mathbf{q}, \eta) = f_0(q)[1 + \delta_F(\mathbf{q}, \eta)\mathrm{e}^{\mathbf{i}\mathbf{k}\cdot\mathbf{r}}], \tag{33}$$

where $f_0(q)$ is the unperturbed distribution,

$$f_0(q) = n_\mathrm{CDM}\left(\frac{M}{2\pi T_\mathrm{CDM}}\right)^{3/2}\exp\left[-\frac{1}{2}\frac{Mq^2}{T_\mathrm{CDM}}\right] \tag{34}$$

where n_CDM and T_CDM are the present-day density and temperature of the dark matter.

Our approach is to solve the Boltzman equation with initial conditions given by the fluid solution at a time η_* (which will depend on k). The simplified Boltzman equation can be easily solved to give $\delta_\mathrm{F}(\mathbf{q}, \eta)$ as a function of the initial conditions $\delta_\mathrm{F}(\mathbf{q}, \eta_*)$,

$$\delta_\mathrm{F}(\mathbf{q}, \eta) = \delta_\mathrm{F}(\mathbf{q}, \eta_*)\exp[-\mathrm{i}\mathbf{q}\cdot\mathbf{k}\frac{\eta_*}{a(\eta_*)}\ln(\eta/\eta_*)]. \tag{35}$$

The CDM overdensity δ_c can then be expressed in terms of the perturbation in the distribution function as,

$$\delta_\mathrm{c}(\eta) = \frac{1}{n_\mathrm{CDM}}\int \mathrm{d}^3q\ f_0(q)\ \delta_\mathrm{F}(\mathbf{q}, \eta). \tag{36}$$

We can use (35) to obtain the evolution of δ_c in terms of its value at η_*,

$$\delta_\mathrm{c}(\eta) = \exp\left[-\frac{1}{2}\frac{k^2}{k_\mathrm{f}^2}\ln^2\left(\frac{\eta}{\eta_*}\right)\right]\ \left[\delta|_{\eta_*} + \frac{\mathrm{d}\delta}{\mathrm{d}\eta}|_{\eta_*}\eta_*\ln\left(\frac{\eta}{\eta_*}\right)\right], \tag{37}$$

where $k_{\rm f}^{-2} = \sqrt{(T_{\rm d}/M)}\eta_{\rm d}$. The exponential term is responsible for the damping of perturbations as a result of free streaming and the dispersion of the CDM particles after they decouple from the plasma. The above expression is only valid during the radiation era. The free streaming scale is simply given by $\int {\rm d}t(v/a) \propto \int {\rm d}t a^{-2}$ which grows logarithmically during the radiation era as in (37) but stops growing in the matter era when $a \propto t^{2/3}$.

Equation (37) can be used to show that even during the free streaming epoch, $\delta_{\rm c}$ satisfies (25) but with a modified sound speed and viscous term. For $x \gg x_{\rm d}$ one should use,

$$c_{\rm s}^2(x) = c_{\rm s}^2(x_{\rm d}) \left(\frac{x_{\rm d}}{x}\right)^2 \left[1 + x_{\rm d}^2 c_{\rm s}^2(x_{\rm d}) \ln^2\left(\frac{x}{x_{\rm d}}\right)\right]$$
$$F_{\rm v}(x) = 2c_{\rm s}^2(x_{\rm d}) x_{\rm d}^2 \ln\left(\frac{x_{\rm d}}{x}\right). \tag{38}$$

The differences between the above scalings and those during the tight coupling regime are a result of the fact that the dark matter temperature stops following the plasma temperature but rather scales as a^{-2} after thermal decoupling, which coincides with the kinematic decoupling. We ignore the effects of heat transfer during the fluid stage of the CDM because its temperature is controlled by the much larger heat reservoir of the radiation-dominated plasma at that stage.

To obtain the transfer function we solve the dark matter fluid equation until decoupling and then evolve the overdensity using (37) up to the time of matter–radiation equality. In practice, we use the fluid equations up to $x_* = 10\,\max(x_{\rm d}, 10)$ so as to switch into the free streaming solution well after the gravitational potential has decayed. In the fluid equations, we smoothly match the sound speed and viscosity terms at $x = x_{\rm d}$. As mentioned earlier, because $c_{\rm s}(x_{\rm d})$ is so small and we are interested in modes that are comparable to the size of the horizon at decoupling, i.e. $x_{\rm d} \sim$ few, both the dark matter sound speed and the associated viscosity play only a minor role, and our simplified treatment is adequate.

In Fig. 6 we illustrate the time evolution of modes during decoupling for a variety of k values. The situation is clear. Modes that enter the horizon before kinematic decoupling oscillate with the radiation fluid. This behavior has two important effects. In the absence of the coupling, modes receive a "kick" by the source term $S(x)$ as they cross the horizon. After that they grow logarithmically. In our case, modes that entered the horizon before kinematic decoupling follow the plasma oscillations and thus miss out on both the horizon "kick" and the beginning of the logarithmic growth. Second, the decoupling from the radiation fluid is not instantaneous and this acts to further damp the amplitude of modes with $x_{\rm d} \gg 1$. This effect can be understood as follows. Once the oscillation frequency of the mode becomes high compared to the scattering rate, the coupling to the plasma effectively damps the mode. In that limit one can replace the forcing term Θ_0' by its average value, which is close to zero. Thus in this regime, the scattering is forcing

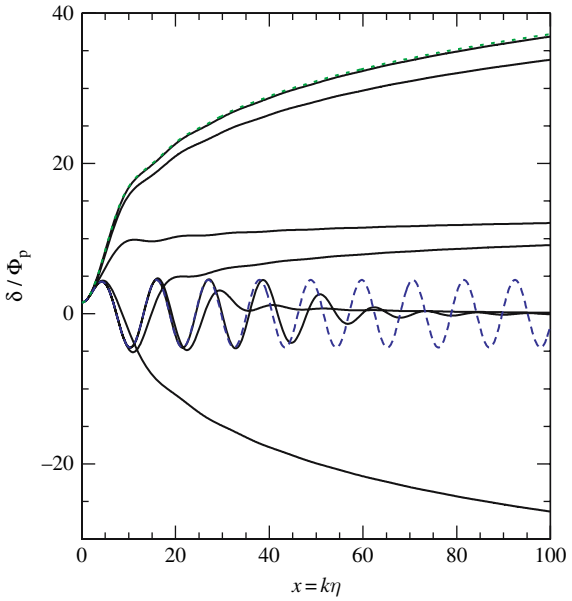

Fig. 6. The normalized amplitude of CDM fluctuations $\delta/\Phi_{\rm P}$ for a variety of modes with comoving wavenumbers $\log(k\eta_{\rm d}) = (0, 1/3, 2/3, 1, 4/3, 5/3, 2)$ as a function of $x \equiv k\eta$, where $\eta = \int_0^t {\rm d}t/a(t)$ is the conformal time coordinate. The dashed line shows the temperature monopole $3\theta_0$ and the uppermost (*dotted*) curve shows the evolution of a mode that is uncoupled to the cosmic plasma. (Figure from Loeb & Zalclarriaga 2005 [227])

the amplitude of the dark matter oscillations to zero. After kinematic decoupling the modes again grow logarithmically but from a very reduced amplitude. *The coupling with the plasma induces both oscillations and damping of modes that entered the horizon before kinematic decoupling. This damping is different from the free streaming damping that occurs after kinematic decoupling.*

In Fig. 7 we show the resulting transfer function of the CDM overdensity. The transfer function is defined as the ratio between the CDM density perturbation amplitude δ_c when the effect of the coupling to the plasma is included and the same quantity in a model where the CDM is a perfect fluid down to arbitrarily small scales (thus, the power spectrum is obtained by multiplying the standard result by the square of the transfer function). This function shows both the oscillations and the damping signature mentioned above. The peaks occur at multipoles of the horizon scale at decoupling,

$$k_{\rm peak} = (8, 15.7, 24.7, ..)\eta_{\rm d}^{-1} \propto \frac{M_{\rm pl}}{T_0 T_{\rm d}}. \tag{39}$$

This same scale determines the "oscillation" damping. The free streaming damping scale is,

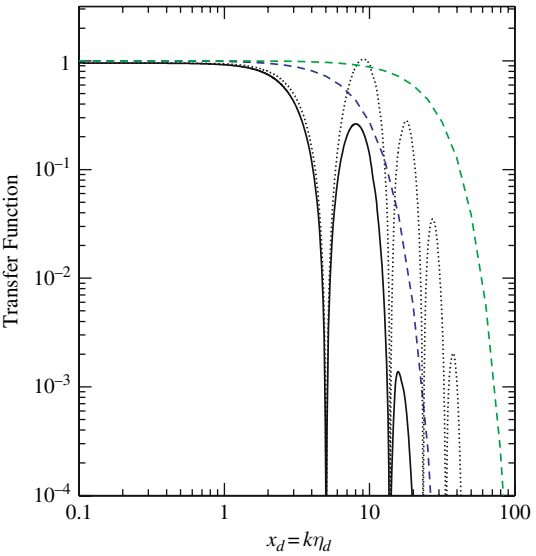

Fig. 7. Transfer function of the CDM density perturbation amplitude (normalized by the primordial amplitude from inflation). We show two cases: (i) $T_{\rm d}/M = 10^{-4}$ and $T_{\rm d}/T_{\rm eq} = 10^{7}$; (ii) $T_{\rm d}/M = 10^{-5}$ and $T_{\rm d}/T_{\rm eq} = 10^{7}$. In each case the oscillatory curve is our result and the other curve is the free-streaming only result that was derived previously in the literature Abel et al. (2000) [4], Adelberger et al. (2003) [7], Aghanim et al. (1996) [8]. (Figure from Loeb & Zaldarriaga 2005 [227].)

$$\eta_{\rm d} c_{\rm d}(\eta_{\rm d}) \ln(\eta_{\rm eq}/\eta_{\rm d}) \propto \frac{M_{\rm pl} M^{1/2}}{T_0 T_{\rm d}^{3/2}} \ln(T_{\rm d}/T_{\rm eq}), \tag{40}$$

where $T_{\rm eq}$ is the temperature at matter radiation equality, $T_{\rm eq} \approx 1\,{\rm eV}$. The free streaming scale is parametrically different from the "oscillation" damping scale. However for our fiducial choice of parameters for the CDM particle they roughly coincide.

The CDM damping scale is significantly smaller than the scales observed directly in the Cosmic Microwave Background or through large scale structure surveys. For example, the ratio of the damping scale to the scale that entered the horizon at Matter-radiation equality is $\eta_{\rm d}/\eta_{\rm eq} \sim T_{\rm eq}/T_{\rm d} \sim 10^{-7}$ and to our present horizon $\eta_{\rm d}/\eta_0 \sim (T_{\rm eq} T_0)^{1/2}/T_{\rm d} \sim 10^{-9}$. In the context of inflation, these scales were created 16 and 20 e–folds apart. Given the large extrapolation, one could certainly imagine that a change in the spectrum could alter the shape of the power spectrum around the damping scale. However, for smooth inflation potentials with small departures from scale invariance this is not likely to be the case. On scales much smaller than the horizon at matter radiation equality, the spectrum of perturbations density before the effects of the damping are included is approximately,

$$\Delta^2(k) \propto \exp\left[(n-1)\ln(k\eta_{\rm eq}) + \frac{1}{2}\alpha^2\ln(k\eta_{\rm eq})^2 + \cdots\right] \times \ln^2(k\eta_{\rm eq}/8) \quad (41)$$

where the first term encodes the shape of the primordial spectrum and the second the transfer function. Primordial departures from scale invariance are encoded in the slope n and its running α. The effective slope at scale k is then,

$$\frac{\partial \ln \Delta^2}{\partial \ln k} = (n-1) + \alpha \ln(k\eta_{\rm eq}) + \frac{2}{\ln(k\eta_{\rm eq}/8)}. \quad (42)$$

For typical values of $(n-1) \sim 1/60$ and $\alpha \sim 1/60^2$ the slope is still positive at $k \sim \eta_{\rm d}^{-1}$, so the cut-off in the power will come from the effects we calculate rather than from the shape of the primordial spectrum. However given the large extrapolation in scale, one should keep in mind the possibility of significant effects resulting from the mechanisms that generates the density perturbations.

Implications

We have found that acoustic oscillations, a relic from the epoch when the dark matter coupled to the cosmic radiation fluid, truncate the CDM power spectrum on a comoving scale larger than effects considered before, such as free-streaming and viscosity [158, 159, 180]. For SUSY dark matter, the minimum mass of dark matter clumps that form in the Universe is therefore increased by more than an order of magnitude to a value of [4]

$$M_{\rm cut} = \frac{4\pi}{3}\left(\frac{\pi}{k_{\rm cut}}\right)^3 \Omega_M \rho_{\rm crit} \simeq 10^{-4}\left(\frac{T_{\rm d}}{10\,{\rm MeV}}\right)^{-3} {\rm M}_\odot, \quad (43)$$

where $\rho_{\rm crit} = (H_0^2/8\pi G) = 9 \times 10^{-30}\,{\rm g\,cm^{-3}}$ is the critical density today, and Ω_M is the matter density for the concordance cosmological model [347]. We define the cut-off wavenumber $k_{\rm cut}$ as the point where the transfer function first drops to a fraction $1/e$ of its value at $k \to 0$. This corresponds to $k_{\rm cut} \approx 3.3\ \eta_{\rm d}^{-1}$.

Recent numerical simulations [105, 146] of the earliest and smallest objects to have formed in the Universe (see Fig. 8), need to be redone for the modified power spectrum that we calculated in this section. Although it is difficult to

[4] Our definition of the cut-off mass follows the convention of the Jeans mass, which is defined as the mass enclosed within a sphere of radius $\lambda_{\rm J}/2$ where $\lambda_{\rm J} \equiv 2\pi/k_{\rm J}$ is the Jeans wavelength [167].

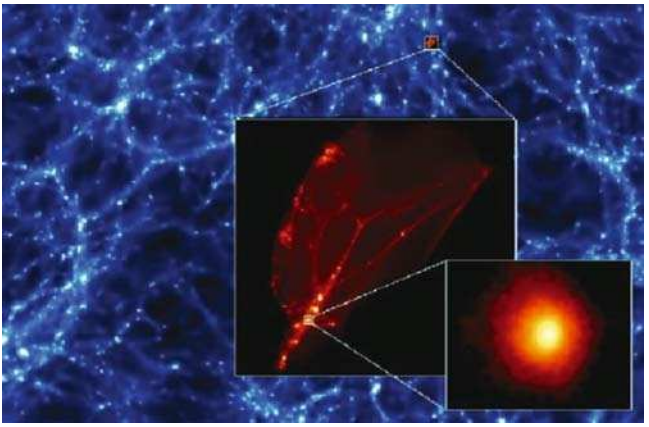

Fig. 8. A slice through a numerical simulation of the first dark matter condensations to form in the Universe. Colors represent the dark matter density at $z = 26$. The simulated volume is 60 comoving pc on a side, simulated with 64 million particles each weighing 1.2×10^{-10} $M_\odot$ (!) (From Diemand et al. 2005 [105].)

forecast the effects of the acoustic oscillations through the standard Press-Schechter formalism [290], it is likely that the results of such simulations will be qualitatively the same as before except that the smallest clumps would have a mass larger than before (as given by 43).

Potentially, there are several observational signatures of the smallest CDM clumps. As pointed out in the literature [105, 352], the smallest CDM clumps could produce γ-rays through dark-matter annihilation in their inner density cusps, with a flux in excess of that from nearby dwarf galaxies. If a substantial fraction of the Milky Way halo is composed of CDM clumps with a mass $\sim 10^{-4}$ $M_\odot$, the nearest clump is expected to be at a distance of $\sim 4 \times 10^{17}$ cm. Given that the characteristic speed of such clumps is a few hundred $\mathrm{km\,s^{-1}}$, the γ-ray flux would therefore show temporal variations on the relatively long timescale of a 1000 years. Passage of clumps through the solar system should also induce fluctuations in the detection rate of CDM particles in direct search experiments. Other observational effects have rather limited prospects for detectability. Because of their relatively low-mass and large size ($\sim 10^{17}$ cm), the CDM clumps are too diffuse to produce any gravitational lensing signatures (including *femto*-lensing [161]), even at cosmological distances.

The smallest CDM clumps should not affect the intergalactic baryons which have a much larger Jeans mass. However, once objects above $\sim 10^6$ $M_\odot$ start to collapse at redshifts $z < 30$, the baryons would be able to cool inside of them via molecular hydrogen transitions and the interior baryonic Jeans mass would drop. The existence of dark matter clumps could then seed the formation of the first stars inside these objects [66].

3.5 Structure of the Baryons

Early Evolution of Baryonic Perturbations on Large Scales

The baryons are coupled through Thomson scattering to the radiation fluid until they become neutral and decouple. After cosmic recombination, they start to fall into the potential wells of the dark matter and their early evolution was derived by Barkana & Loeb (2005) [29].

On large scales, the dark matter (dm) and the baryons (b) are affected only by their combined gravity and gas pressure can be ignored. The evolution of sub-horizon linear perturbations is described in the matter-dominated regime by two coupled second-order differential equations [283]:

$$\begin{aligned} \ddot{\delta}_{\rm dm} + 2H\dot{\delta}_{\rm dm} &= 4\pi G\bar{\rho}_m \left(f_{\rm b}\delta_{\rm b} + f_{\rm dm}\delta_{\rm dm}\right) \ , \\ \ddot{\delta}_{\rm b} + 2H\dot{\delta}_{\rm b} &= 4\pi G\bar{\rho}_m \left(f_{\rm b}\delta_{\rm b} + f_{\rm dm}\delta_{\rm dm}\right) \ , \end{aligned} \tag{44}$$

where $\delta_{\rm dm}(t)$ and $\delta_{\rm b}(t)$ are the perturbations in the dark matter and baryons, respectively, the derivatives are with respect to cosmic time t, $H(t) = \dot{a}/a$ is the Hubble constant with $a = (1+z)^{-1}$, and we assume that the mean mass density $\bar{\rho}_m(t)$ is made up of respective mass fractions $f_{\rm dm}$ and $f_{\rm b} = 1 - f_{\rm dm}$. Since these linear equations contain no spatial gradients, they can be solved spatially for $\delta_{\rm dm}(\mathbf{x},\mathbf{t})$ and $\delta_{\rm b}(\mathbf{x},\mathbf{t})$ or in Fourier space for $\tilde{\delta}_{\rm dm}(\mathbf{k},\mathbf{t})$ and $\tilde{\delta}_{\rm b}(\mathbf{k},\mathbf{t})$.

Defining $\delta_{\rm tot} \equiv f_{\rm b}\delta_{\rm b} + f_{\rm dm}\delta_{\rm dm}$ and $\delta_{\rm b-} \equiv \delta_{\rm b} - \delta_{\rm tot}$, we find

$$\begin{aligned} \ddot{\delta}_{\rm tot} + 2H\dot{\delta}_{\rm tot} &= 4\pi G\bar{\rho}_m\delta_{\rm tot} \ , \\ \ddot{\delta}_{\rm b-} + 2H\dot{\delta}_{\rm b-} &= 0 \ . \end{aligned} \tag{45}$$

Each of these equations has two independent solutions. The equation for $\delta_{\rm tot}$ has the usual growing and decaying solutions, which we denote $D_1(t)$ and $D_4(t)$, respectively, while the $\delta_{\rm b-}$ equation has solutions $D_2(t)$ and $D_3(t)$; we number the solutions in order of declining growth rate (or increasing decay rate). We assume an Einstein-de Sitter, matter-dominated Universe in the redshift range $z = 20$–150, since the radiation contributes less than a few percent at $z < 150$, while the cosmological constant and the curvature contribute to the energy density less than a few percent at $z > 3$. In this regime $a \propto t^{2/3}$ and the solutions are $D_1(t) = a/a_i$ and $D_4(t) = (a/a_i)^{-3/2}$ for $\delta_{\rm tot}$, and $D_2(t) = 1$ and $D_3(t) = (a/a_i)^{-1/2}$ for $\delta_{\rm b-}$, where we have normalized each solution to unity at the starting scale factor a_i, which we set at a redshift $z_i = 150$. The observable baryon perturbation can then be written as

$$\tilde{\delta}_{\rm b}(\mathbf{k},\mathbf{t}) = \tilde{\delta}_{\rm b-} + \tilde{\delta}_{\rm tot} = \sum_{\mathbf{m=1}}^{\mathbf{4}} \tilde{\delta}_{\mathbf{m}}(\mathbf{k})\,\mathbf{D_m}(\mathbf{t}) \ , \tag{46}$$

and similarly for the dark matter perturbation,

$$\tilde{\delta}_{\rm dm} = \frac{1}{f_{\rm dm}} \left(\tilde{\delta}_{\rm tot} - f_b \tilde{\delta}_b \right) = \sum_{m=1}^{4} \tilde{\delta}_m(\mathbf{k})\, \mathbf{C_m(t)} \ , \tag{47}$$

where $C_i = D_i$ for $i = 1, 4$ and $C_i = -(f_{\rm b}/f_{\rm dm}) D_i$ for $i = 2, 3$. We may establish the values of $\tilde{\delta}_{\rm m}(\mathbf{k})$ by inverting the 4×4 matrix $\mathbf{A}$ that relates the 4-vector $(\tilde{\delta}_1, \tilde{\delta}_2, \tilde{\delta}_3, \tilde{\delta}_4)$ to the 4-vector that represents the initial conditions $(\tilde{\delta}_{\rm b}, \tilde{\delta}_{\rm dm}, \dot{\tilde{\delta}}_{\rm b}, \dot{\tilde{\delta}}_{\rm dm})$ at the initial time.

Next we describe the fluctuations in the sound speed of the cosmic gas caused by Compton heating of the gas, which is due to scattering of the residual electrons with the CMB photons. The evolution of the temperature T of a gas element of density $\rho_{\rm b}$ is given by the first law of thermodynamics:

$$\mathrm{d}Q = \frac{3}{2} \mathbf{k} \mathrm{d}T - \mathbf{k} T \mathrm{d} \log \rho_{\rm b} \ , \tag{48}$$

where $\mathrm{d}Q$ is the heating rate per particle. Before the first galaxies formed,

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = 4 \frac{\sigma_{\rm T}\, c}{m_{\rm e}}\, \mathbf{k}_{\rm B} (T_\gamma - T) \rho_\gamma x_{\rm e}(t) \ , \tag{49}$$

where $\sigma_{\rm T}$ is the Thomson cross-section, $x_{\rm e}(t)$ is the electron fraction out of the total number density of gas particles, and ρ_γ is the CMB energy density at a temperature T_γ. In the redshift range of interest, we assume that the photon temperature ($T_\gamma = T_\gamma^0/a$) is spatially uniform, since the high sound speed of the photons (i.e., $c/\sqrt{3}$) suppresses fluctuations on the sub-horizon scales that we consider, and the horizon-scale $\sim 10^{-5}$ fluctuations imprinted at cosmic recombination are also negligible compared to the smaller-scale fluctuations in the gas density and temperature. Fluctuations in the residual electron fraction $x_{\rm e}(t)$ are even smaller. Thus,

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{2}{3} T \frac{\mathrm{d} \log \rho_{\rm b}}{\mathrm{d}t} + \frac{x_{\rm e}(t)}{t_\gamma} \left(T_\gamma - T \right) a^{-4} \ , \tag{50}$$

where $t_\gamma^{-1} \equiv \bar{\rho}_\gamma^0 (8 \sigma_{\rm T}\, c / 3 m_{\rm e}) = 8.55 \times 10^{-13}\, {\rm yr}^{-1}$. After cosmic recombination, $x_{\rm e}(t)$ changes due to the slow recombination rate of the residual ions:

$$\frac{\mathrm{d}x_{\rm e}(t)}{\mathrm{d}t} = -\alpha_{\rm B}(T) x_{\rm e}^2(t) \bar{n}_{\rm H} (1 + y) \ , \tag{51}$$

where $\alpha_{\rm B}(T)$ is the case-B recombination coefficient of hydrogen, $\bar{n}_{\rm H}$ is the mean number density of hydrogen at time t, and $y = 0.079$ is the helium to hydrogen number density ratio. This yields the evolution of the mean temperature, $\mathrm{d}\bar{T}/\mathrm{d}t = -2H\bar{T} + x_{\rm e}(t) t_\gamma^{-1} \left(T_\gamma - \bar{T} \right) a^{-4}$. In prior analyses [283, 229] a spatially uniform speed of sound was assumed for the gas at each redshift. Note that we refer to $\delta p / \delta \rho$ as the square of the sound speed of the fluid, where δp is the pressure perturbation, although we are analyzing perturbations driven by gravity rather than sound waves driven by pressure gradients.

Instead of assuming a uniform sound speed, we find the first-order perturbation equation,

$$\frac{\mathrm{d}\delta_T}{\mathrm{d}t} = \frac{2}{3}\frac{\mathrm{d}\delta_\mathrm{b}}{\mathrm{d}t} - \frac{x_\mathrm{e}(t)}{t_\gamma}\frac{T_\gamma}{\bar{T}}a^{-4}\delta_\mathrm{T} \ , \tag{52}$$

where we defined the fractional temperature perturbation δ_T. Like the density perturbation equations, this equation can be solved separately at each $\mathbf{x}$ or at each $\mathbf{k}$. Furthermore, the solution $\delta_\mathrm{T}(t)$ is a linear functional of $\delta_\mathrm{b}(t)$ [for a fixed function $x_\mathrm{e}(t)$]. Thus, if we choose an initial time t_i then using (46) we can write the solution in Fourier space as

$$\tilde{\delta}_T(\mathbf{k},\mathbf{t}) = \sum_{\mathbf{m=1}}^{\mathbf{4}} \tilde{\delta}_\mathbf{m}(\mathbf{k})\,\mathbf{D^T_m}(\mathbf{t}) + \tilde{\delta}_\mathrm{T}(\mathbf{k},\mathbf{t_i})\,\mathbf{D^T_0}(\mathbf{t}) \ , \tag{53}$$

where $D^T_m(t)$ is the solution of (52) with $\delta_\mathrm{T} = 0$ at t_i and with the perturbation mode $D_m(t)$ substituted for $\delta_\mathrm{b}(t)$, while $D^T_0(t)$ is the solution with no perturbation $\delta_\mathrm{b}(t)$ and with $\delta_\mathrm{T} = 1$ at t_i. By modifying the CMBFAST code (http://www.cmbfast.org/), we can numerically solve (52) along with the density perturbation equations for each $\mathbf{k}$ down to $z_i = 150$, and then match the solution to the form of (53).

Figure 9 shows the time evolution of the various independent modes that make up the perturbations of density and temperature, starting at the time t_i corresponding to $z_\mathrm{i} = 150$. $D^T_2(t)$ is identically zero since $D_2(t) = 1$ is constant, while $D^T_3(t)$ and $D^T_4(t)$ are negative. Figure 10 shows the amplitudes of the various components of the initial perturbations. We consider comoving wavevectors k in the range 0.01 – $40\,\mathrm{Mpc}^{-1}$, where the lower limit is set by considering sub-horizon scales at $z = 150$ for which photon perturbations are negligible compared to δ_dm and δ_b, and the upper limit is set by requiring baryonic pressure to be negligible compared to gravity. $\tilde{\delta}_2$ and $\tilde{\delta}_3$ clearly show a strong signature of the large-scale baryonic oscillations, left over from the era of the photon-baryon fluid before recombination, while $\tilde{\delta}_1$, $\tilde{\delta}_4$, and $\tilde{\delta}_\mathrm{T}$ carry only a weak sign of the oscillations. For each quantity, the plot shows $[k^3P(k)/(2\pi^2)]^{1/2}$, where $P(k)$ is the corresponding power spectrum of fluctuations. $\tilde{\delta}_4$ is already a very small correction at $z = 150$ and declines quickly at lower redshift, but the other three modes all contribute significantly to $\tilde{\delta}_\mathrm{b}$, and the $\tilde{\delta}_\mathrm{T}(t_\mathrm{i})$ term remains significant in $\tilde{\delta}_\mathrm{T}(t)$ even at $z \lesssim 100$. Note that at $z = 150$ the temperature perturbation $\tilde{\delta}_\mathrm{T}$ has a different shape with respect to k than the baryon perturbation $\tilde{\delta}_\mathrm{b}$, showing that their ratio cannot be described by a scale-independent speed of sound.

The power spectra of the various perturbation modes and of $\tilde{\delta}_\mathrm{T}(t_\mathrm{i})$ depend on the initial power spectrum of density fluctuations from inflation and on the values of the fundamental cosmological parameters (Ω_dm, Ω_b, Ω_Λ, and h). If these independent power spectra can be measured through 21 cm fluctuations, this will probe the basic cosmological parameters through multiple combinations, allowing consistency checks that can be used to verify the adiabatic nature and the expected history of the perturbations. Figure 11 illustrates

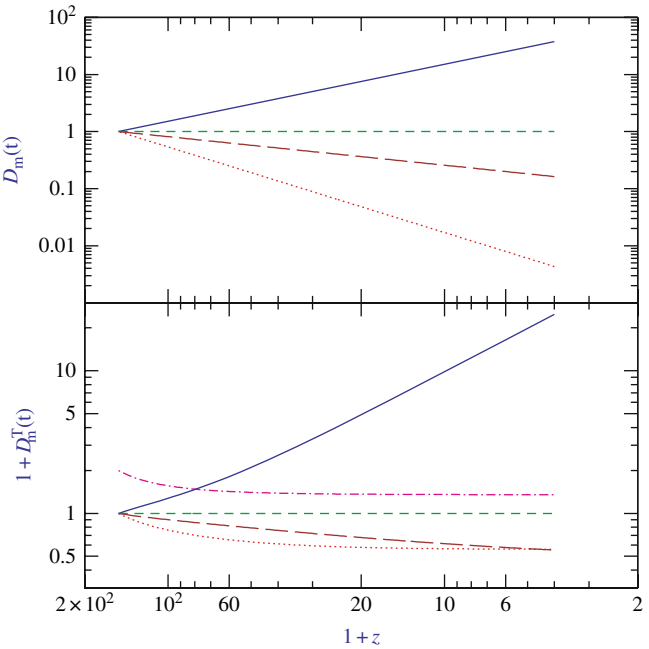

Fig. 9. Redshift evolution of the amplitudes of the independent modes of the density perturbations (**upper panel**) and the temperature perturbations (**lower panel**), starting at redshift 150 (from Barkana & Loeb 2005 [29]). We show $m = 1$ (*solid curves*), $m = 2$ (*short-dashed curves*), $m = 3$ (*long-dashed curves*), $m = 4$ (*dotted curves*), and $m = 0$ (*dot-dashed curve*). Note that the lower panel shows one plus the mode amplitude

the relative sensitivity of $\sqrt{P(k)}$ to variations in $\Omega_{\rm dm}h^2$, $\Omega_{\rm b}h^2$, and h, for the quantities $\tilde{\delta}_1$, $\tilde{\delta}_2$, $\tilde{\delta}_3$, and $\tilde{\delta}_{\rm T}(t_{\rm i})$. Not shown is $\tilde{\delta}_4$, which although it is more sensitive (changing by order unity due to 10% variations in the parameters), its magnitude always remains much smaller than the other modes, making it much harder to detect. Note that although the angular scale of the baryon oscillations constrains also the history of dark energy through the angular diameter distance, we have focused here on other cosmological parameters, since the contribution of dark energy relative to matter becomes negligible at high redshift.

Cosmological Jeans Mass

The Jeans length $\lambda_{\rm J}$ was originally defined (Jeans 1928 [185]) in Newtonian gravity as the critical wavelength that separates oscillatory and exponentially-growing density perturbations in an infinite, uniform, and stationary distribution of gas. On scales ℓ smaller than $\lambda_{\rm J}$, the sound crossing time, $\ell/c_{\rm s}$ is shorter than the gravitational free-fall time, $(G\rho)^{-1/2}$, allowing the build-up

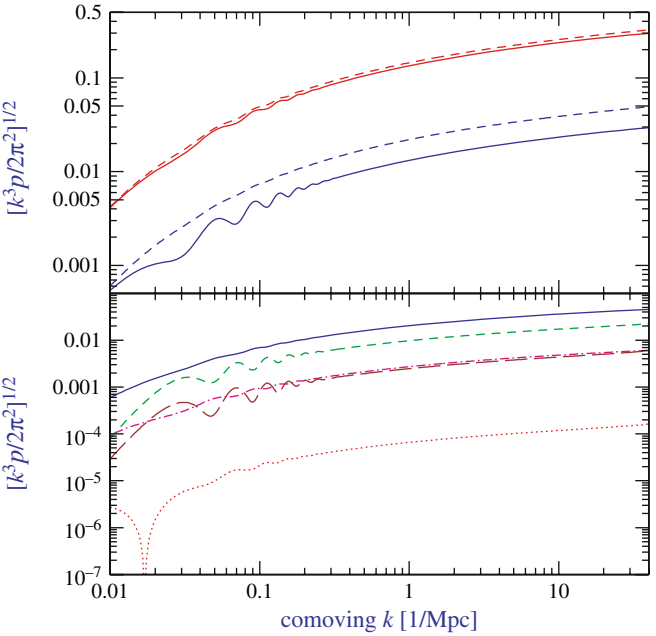

Fig. 10. Power spectra and initial perturbation amplitudes versus wavenumber (from Burkara & Loeb 2005 [29]). The upper panel shows $\tilde{\delta}_{\rm b}$ (*solid curves*) and $\tilde{\delta}_{\rm dm}$ (*dashed curves*) at $z = 150$ and 20 (**from bottom to top**). The lower panel shows the initial ($z = 150$) amplitudes of $\tilde{\delta}_1$ (*solid curve*), $\tilde{\delta}_2$ (*short-dashed curve*), $\tilde{\delta}_3$ (*long-dashed curve*), $\tilde{\delta}_4$ (*dotted curve*), and $\tilde{\delta}_{\rm T}(t_{\rm i})$ (*dot-dashed curve*). Note that if $\tilde{\delta}_1$ is positive then so are $\tilde{\delta}_3$ and $\tilde{\delta}_{\rm T}(t_{\rm i})$, while $\tilde{\delta}_2$ is negative at all k, and $\tilde{\delta}_4$ is negative at the lowest k but is positive at $k > 0.017\,{\rm Mpc}^{-1}$

of a pressure force that counteracts gravity. On larger scales, the pressure gradient force is too slow to react to a build-up of the attractive gravitational force. The Jeans mass is defined as the mass within a sphere of radius $\lambda_{\rm J}/2$, $M_{\rm J} = (4\pi/3)\rho(\lambda_{\rm J}/2)^3$. In a perturbation with a mass greater than $M_{\rm J}$, the self-gravity cannot be supported by the pressure gradient, and so the gas is unstable to gravitational collapse. The Newtonian derivation of the Jeans instability suffers from a conceptual inconsistency, as the unperturbed gravitational force of the uniform background must induce bulk motions (compare to Binney & Tremaine 1987 [43]). However, this inconsistency is remedied when the analysis is done in an expanding Universe.

The perturbative derivation of the Jeans instability criterion can be carried out in a cosmological setting by considering a sinusoidal perturbation superposed on a uniformly expanding background. Here, as in the Newtonian limit, there is a critical wavelength $\lambda_{\rm J}$ that separates oscillatory and growing modes. Although the expansion of the background slows down the exponential growth of the amplitude to a power-law growth, the fundamental concept of

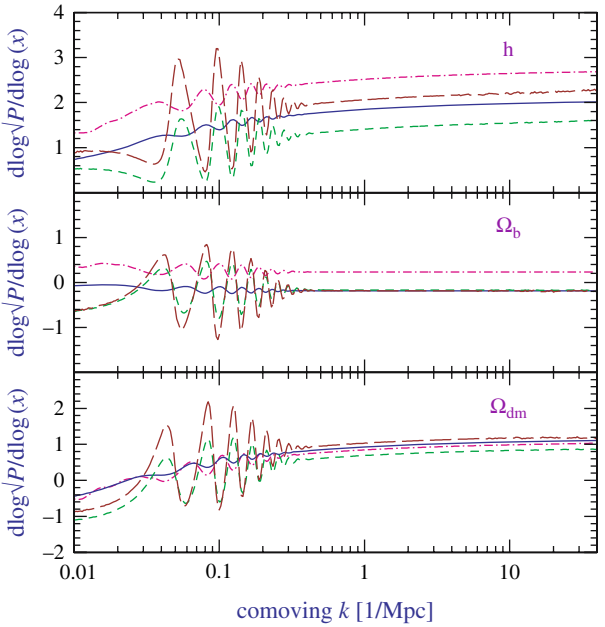

Fig. 11. Relative sensitivity of perturbation amplitudes at $z = 150$ to cosmological parameters (from Barkana & Loeb 2005 [29]). For variations in a parameter x, we show $\mathrm{d}\log\sqrt{P(k)}/\mathrm{d}\log(x)$. We consider variations in $\Omega_{\rm dm}h^2$ (**upper panel**), in $\Omega_{\rm b}h^2$ (**middle panel**), and in the Hubble constant h (**lower panel**). When we vary each parameter we fix the other two, and the variations are all carried out in a flat $\Omega_{\rm total} = 1$ universe. We show the sensitivity of $\tilde{\delta}_1$ (*solid curves*), $\tilde{\delta}_2$ (*short-dashed curves*), $\tilde{\delta}_3$ (*long-dashed curves*), and $\tilde{\delta}_{\rm T}(t_{\rm i})$ (*dot-dashed curves*)

a minimum mass that can collapse at any given time remains the same (see, e.g. Kolb & Turner 1990 [203]; Peebles 1993 [283]).

We consider a mixture of dark matter and baryons with density parameters $\Omega_{\rm dm}^z = \bar{\rho}_{\rm dm}/\rho_{\rm c}$ and $\Omega_{\rm b}^z = \bar{\rho}_{\rm b}/\rho_{\rm c}$, where $\bar{\rho}_{\rm dm}$ is the average dark matter density, $\bar{\rho}_{\rm b}$ is the average baryonic density, $\rho_{\rm c}$ is the critical density, and $\Omega_{\rm dm}^z + \Omega_{\rm b}^z = \Omega_{\rm m}^z$ is given by (83). We also assume spatial fluctuations in the gas and dark matter densities with the form of a single spherical Fourier mode on a scale much smaller than the horizon,

$$\frac{\rho_{\rm dm}(r,t) - \bar{\rho}_{\rm dm}(t)}{\bar{\rho}_{\rm dm}(t)} = \delta_{\rm dm}(t)\frac{\sin(kr)}{kr} \ , \tag{54}$$

$$\frac{\rho_{\rm b}(r,t) - \bar{\rho}_{\rm b}(t)}{\bar{\rho}_{\rm b}(t)} = \delta_{\rm b}(t)\frac{\sin(kr)}{kr} \ , \tag{55}$$

where $\bar{\rho}_{\rm dm}(t)$ and $\bar{\rho}_{\rm b}(t)$ are the background densities of the dark matter and baryons, $\delta_{\rm dm}(t)$ and $\delta_{\rm b}(t)$ are the dark matter and baryon overdensity amplitudes, r is the comoving radial coordinate, and k is the comoving perturbation

wavenumber. We adopt an ideal gas equation-of-state for the baryons with a specific heat ratio γ=5/3. Initially, at time $t = t_{\rm i}$, the gas temperature is uniform $T_{\rm b}(r, t_{\rm i})$=$T_{\rm i}$, and the perturbation amplitudes are small $\delta_{\rm dm,i}, \delta_{\rm b,i} \ll 1$. We define the region inside the first zero of $\sin(kr)/(kr)$, namely $0 < kr < \pi$, as the collapsing "object".

The evolution of the temperature of the baryons $T_{\rm b}(r, t)$ in the linear regime is determined by the coupling of their free electrons to the CMB through Compton scattering, and by the adiabatic expansion of the gas. Hence, $T_{\rm b}(r, t)$ is generally somewhere between the CMB temperature, $T_\gamma \propto (1 + z)^{-1}$ and the adiabatically-scaled temperature $T_{\rm ad} \propto (1 + z)^{-2}$. In the limit of tight coupling to T_γ, the gas temperature remains uniform (Fig. 12). On the other hand, in the adiabatic limit, the temperature develops a gradient according to the relation

$$T_{\rm b} \propto \rho_{\rm b}^{(\gamma-1)}. \tag{56}$$

The evolution of a cold dark matter overdensity, $\delta_{\rm dm}(t)$, in the linear regime is described by the (44),

$$\ddot{\delta}_{\rm dm} + 2H\dot{\delta}_{\rm dm} = \frac{3}{2}H^2\left(\Omega_{\rm b}\delta_{\rm b} + \Omega_{\rm dm}\delta_{\rm dm}\right) \tag{57}$$

whereas the evolution of the overdensity of the baryons, $\delta_{\rm b}(t)$, with the inclusion of their pressure force is described by (see Sect. 9.3.2 of [203]),

$$\ddot{\delta}_{\rm b} + 2H\dot{\delta}_{\rm b} = \frac{3}{2}H^2\left(\Omega_{\rm b}\delta_{\rm b} + \Omega_{\rm dm}\delta_{\rm dm}\right) - \frac{kT_{\rm i}}{\mu m_p}\left(\frac{k}{a}\right)^2\left(\frac{a_{\rm i}}{a}\right)^{(1+\beta)} \times \left(\delta_{\rm b} + \frac{2}{3}\beta[\delta_{\rm b} - \delta_{\rm b,i}]\right). \tag{58}$$

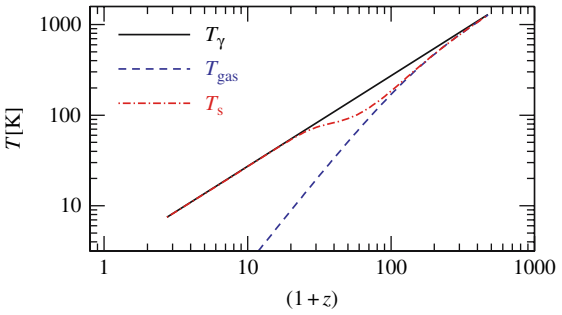

Fig. 12. Thermal history of the baryons, left over from the Big Bang, before the first galaxies formed Loeb & Zaldarriaga (2004) [225]. The residual fraction of free electrons couple the gas temperture $T_{\rm gas}$ to the cosmic microwave background temperature $[T_\gamma \propto (1 + z)]$ until a redshift $z \sim 200$. Subsequently the gas temperature cools adiabatically at a faster rate $[T_{\rm gas} \propto (1 + z)^2]$. Also shown is the spin temperature $T_{\rm s}$ of the 21 cm transition of hydrogen which interpolates between the gas and radiation temperature and will be discussed in detail later in this review

Here, $H(t) = \dot{a}/a$ is the Hubble parameter at a cosmological time t, and $\mu = 1.22$ is the mean molecular weight of the neutral primordial gas in atomic units. The parameter β distinguishes between the two limits for the evolution of the gas temperature. In the adiabatic limit $\beta = 1$, and when the baryon temperature is uniform and locked to the background radiation, $\beta = 0$. The last term on the right hand side (in square brackets) takes into account the extra pressure gradient force in $\nabla(\rho_{\rm b}T) = (T\nabla\rho_{\rm b} + \rho_{\rm b}\nabla T)$, arising from the temperature gradient which develops in the adiabatic limit. The Jeans wavelength $\lambda_{\rm J} = 2\pi/k_{\rm J}$ is obtained by setting the right-hand side of (58) to zero, and solving for the critical wavenumber $k_{\rm J}$. As can be seen from (58), the critical wavelength $\lambda_{\rm J}$ (and therefore the mass $M_{\rm J}$) is in general time-dependent. We infer from (58) that as time proceeds, perturbations with increasingly smaller initial wavelengths stop oscillating and start to grow.

To estimate the Jeans wavelength, we equate the right-hand-side of (58) to zero. We further approximate $\delta_{\rm b} \sim \delta_{\rm dm}$, and consider sufficiently high redshifts at which the Universe is matter dominated and flat, $(1+z) \gg \max[(1-\Omega_{\rm m}-\Omega_\Lambda)/\Omega_{\rm m}, (\Omega_\Lambda/\Omega_{\rm m})^{1/3}]$. In this regime, $\Omega_{\rm b} \ll \Omega_{\rm m} \approx 1$, $H \approx 2/(3t)$, and $a = (1+z)^{-1} \approx (3H_0\sqrt{\Omega_{\rm m}}/2)^{2/3}t^{2/3}$, where $\Omega_{\rm m} = \Omega_{\rm dm} + \Omega_{\rm b}$ is the total matter density parameter. Following cosmological recombination at $z \approx 10^3$, the residual ionization of the cosmic gas keeps its temperature locked to the CMB temperature (via Compton scattering) down to a redshift of [283]:

$$1 + z_{\rm t} \approx 160(\Omega_{\rm b}h^2/0.022)^{2/5} \ . \tag{59}$$

In the redshift range between recombination and z_t, $\beta = 0$ and

$$k_{\rm J} \equiv (2\pi/\lambda_{\rm J}) = [2kT_\gamma(0)/3\mu m_{\rm p}]^{-1/2}\sqrt{\Omega_{\rm m}}H_0 \ , \tag{60}$$

so that the Jeans mass is therefore redshift independent and obtains the value (for the total mass of baryons and dark matter)

$$M_{\rm J} \equiv \frac{4\pi}{3}\left(\frac{\lambda_{\rm J}}{2}\right)^3\bar{\rho}(0) = 1.35\times10^5\left(\frac{\Omega_{\rm m}h^2}{0.15}\right)^{-1/2} {\rm M}_\odot \ . \tag{61}$$

Based on the similarity of $M_{\rm J}$ to the mass of a globular cluster, Peebles & Dicke (1968) [280] suggested that globular clusters form as the first generation of baryonic objects shortly after cosmological recombination. Peebles & Dicke assumed a baryonic Universe, with a nonlinear fluctuation amplitude on small scales at $z \sim 10^3$, a model which has by now been ruled out. The lack of a dominant mass of dark matter inside globular clusters makes it unlikely that they formed through direct cosmological collapse, and more likely that they resulted from fragmentation during the process of galaxy formation.

At $z \lesssim z_{\rm t}$, the gas temperature declines adiabatically as $[(1+z)/(1+z_{\rm t})]^2$ (i.e., $\beta = 1$) and the total Jeans mass obtains the value,

$$M_{\rm J} = 4.54\times10^3\left(\frac{\Omega_{\rm m}h^2}{0.15}\right)^{-1/2}\left(\frac{\Omega_{\rm b}h^2}{0.022}\right)^{-3/5}\left(\frac{1+z}{10}\right)^{3/2} {\rm M}_\odot. \tag{62}$$

It is not clear how the value of the Jeans mass derived above relates to the mass of collapsed, bound objects. The above analysis is perturbative (Equations (57) and (58) are valid only as long as $\delta_{\rm b}$ and $\delta_{\rm dm}$ are much smaller than unity), and thus can only describe the initial phase of the collapse. As $\delta_{\rm b}$ and $\delta_{\rm dm}$ grow and become larger than unity, the density profiles start to evolve and dark matter shells may cross baryonic shells [**?**] due to their different dynamics. Hence the amount of mass enclosed within a given baryonic shell may increase with time, until eventually the dark matter pulls the baryons with it and causes their collapse even for objects below the Jeans mass.

Even within linear theory, the Jeans mass is related only to the evolution of perturbations at a given time. When the Jeans mass itself varies with time, the overall suppression of the growth of perturbations depends on a time-weighted Jeans mass. Gnedin & Hui (1998) [150] showed that the correct time-weighted mass is the filtering mass $M_{\rm F} = (4\pi/3)\,\bar{\rho}\,(2\pi a/k_{\rm F})^3$, in terms of the comoving wavenumber $k_{\rm F}$ associated with the "filtering scale" (note the change in convention from $\pi/k_{\rm J}$ to $2\pi/k_{\rm F}$). The wavenumber $k_{\rm F}$ is related to the Jeans wavenumber $k_{\rm J}$ by

$$\frac{1}{k_{\rm F}^2(t)} = \frac{1}{D(t)} \int_0^t {\rm d}t'\, a^2(t') \frac{\ddot{D}(t') + 2H(t')\dot{D}(t')}{k_J^2(t')} \int_{t'}^t \frac{{\rm d}t''}{a^2(t'')} \ , \qquad (63)$$

where $D(t)$ is the linear growth factor. At high redshift (where $\Omega_{\rm m}^z \to 1$), this relation simplifies to [153]

$$\frac{1}{k_{\rm F}^2(t)} = \frac{3}{a} \int_0^a \frac{{\rm d}a'}{k_{\rm J}^2(a')} \left(1 - \sqrt{\frac{a'}{a}}\right) \ . \qquad (64)$$

Then the relationship between the linear overdensity of the dark matter $\delta_{\rm dm}$ and the linear overdensity of the baryons $\delta_{\rm b}$, in the limit of small k, can be written as [150]

$$\frac{\delta_{\rm b}}{\delta_{\rm dm}} = 1 - \frac{k^2}{k_{\rm F}^2} + O(k^4) \ . \qquad (65)$$

Linear theory specifies whether an initial perturbation, characterized by the parameters k, $\delta_{\rm dm,i}$, $\delta_{\rm b,i}$ and $t_{\rm i}$, begins to grow. To determine the minimum mass of nonlinear baryonic objects resulting from the shell-crossing and virialization of the dark matter, we must use a different model which examines the response of the gas to the gravitational potential of a virialized dark matter halo.

3.6 Formation of Nonlinear Objects

Spherical Collapse

Let us consider a spherically symmetric density or velocity perturbation of the smooth cosmological background, and examine the dynamics of a test particle

at a radius r relative to the center of symmetry. Birkhoff's (1923) [44] theorem implies that we may ignore the mass outside this radius in computing the motion of our particle. We further find that the relativistic equations of motion describing the system reduce to the usual Friedmann equation for the evolution of the scale factor of a homogeneous Universe, but with a density parameter Ω that now takes account of the additional mass or peculiar velocity. In particular, despite the arbitrary density and velocity profiles given to the perturbation, only the total mass interior to the particle's radius and the peculiar velocity at the particle's radius contribute to the effective value of Ω. We thus find a solution to the particle's motion which describes its departure from the background Hubble flow and its subsequent collapse or expansion. This solution holds until our particle crosses paths with one from a different radius, which happens rather late for most initial profiles.

As with the Friedmann equation for a smooth Universe, it is possible to reinterpret the problem into a Newtonian form. Here we work in an inertial (i.e. non-comoving) coordinate system and consider the force on the particle as that resulting from a point mass at the origin (ignoring the possible presence of a vacuum energy density):

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = -\frac{GM}{r^2}, \tag{66}$$

where G is Newton's constant, r is the distance of the particle from the center of the spherical perturbation, and M is the total mass within that radius. As long as the radial shells do not cross each other, the mass M is constant in time. The initial density profile determines M, while the initial velocity profile determines $\mathrm{d}r/\mathrm{d}t$ at the initial time. As is well-known, there are three branches of solutions: one in which the particle turns around and collapses, another in which it reaches an infinite radius with some asymptotically positive velocity, and a third intermediate case in which it reachs an infinite radius but with a velocity that approaches zero. These cases may be written as [164]:

$$\left.\begin{aligned} r &= A(\cos\eta - 1) \\ t &= B(\eta - \sin\,\eta) \end{aligned}\right\} \qquad \text{Closed} \qquad (0 \leqslant \eta \leqslant 2\pi) \tag{67}$$

$$\left.\begin{aligned} r &= A\eta^2/2 \\ t &= B\eta^3/6 \end{aligned}\right\} \qquad \text{Flat} \qquad (0 \leqslant \eta \leqslant \infty) \tag{68}$$

$$\left.\begin{aligned} r &= A(\cosh\eta - 1) \\ t &= B(\sinh\eta - \eta) \end{aligned}\right\} \qquad \text{Open} \qquad (0 \leqslant \eta \leqslant \infty) \tag{69}$$

where $A^3 = GMB^2$ applies in all cases. All three solutions have $r^3 = 9GMt^2/2$ as t goes to zero, which matches the linear theory expectation that the perturbation amplitude get smaller as one goes back in time. In the closed case,

the shell turns around at time πB and radius $2A$ and collapses to zero radius at time $2\pi B$.

We are now faced with the problem of relating the spherical collapse parameters A, B, and M to the linear theory density perturbation δ [282]. We do this by returning to the equation of motion. Consider that at an early epoch (i.e. scale factor $a_{\rm i} \ll 1$), we are given a spherical patch of uniform overdensity $\delta_{\rm i}$ (the so-called "top-hat" perturbation). If Ω is essentially unity at this time and if the perturbation is pure growing mode, then the initial velocity is radially inward with magnitude $\delta_{\rm i} H(t_{\rm i}) r/3$, where $H(t_{\rm i})$ is the Hubble constant at the initial time and r is the radius from the center of the sphere. This can be easily seen from the continuity equation in spherical coordinates. The equation of motion (in noncomoving coordinates) for a particle beginning at radius $r_{\rm i}$ is simply

$$\frac{{\rm d}^2 r}{{\rm d}t^2} = -\frac{GM}{r^2} + \frac{\Lambda r}{3}, \tag{70}$$

where $M = (4\pi/3) r_{\rm i}^3 \rho_{\rm i} (1+\delta_{\rm i})$ and $\rho_{\rm i}$ is the background density of the Universe at time $t_{\rm i}$. We next define the dimensionless radius $x = r a_{\rm i}/r_{\rm i}$ and rewrite (70) as

$$\frac{l}{H_0^2} \frac{{\rm d}^2 x}{{\rm d}t^2} = -\frac{\Omega_m}{2x^2}(1+\delta_{\rm i}) + \Omega_\Lambda x. \tag{71}$$

Our initial conditions for the integration of this orbit are

$$x(t_{\rm i}) = a_{\rm i} \tag{72}$$

$$\frac{{\rm d}x}{{\rm d}t}(t_{\rm i}) = H(t_1) x \left(1 - \frac{\delta_{\rm i}}{3}\right) = H_0 a_{\rm i} \left(1 - \frac{\delta_{\rm i}}{3}\right) \sqrt{\frac{\Omega_{\rm m}}{a_{\rm i}^3} + \frac{\Omega_{\rm k}}{a_{\rm i}^2} + \Omega_\Lambda}, \tag{73}$$

where $H(t_1) = H_0 [\Omega_{\rm m}/a^3(t_1) + (1-\Omega_{\rm m})]^{1/2}$ is the Hubble parameter for a flat Universe at a a cosmic time t_1. Integrating (71) yields

$$\frac{1}{H_0^2} \left(\frac{{\rm d}x}{{\rm d}t}\right)^2 = \frac{\Omega_{\rm m}}{x}(1+\delta_{\rm i}) + \Omega_\Lambda x^2 + K, \tag{74}$$

where K is a constant of integration. Evaluating this at the initial time and dropping terms of $O(a_{\rm i})$ (but $\delta_{\rm i} \sim a_{\rm i}$, so we keep ratios of order unity), we find

$$K = -\frac{5\delta_{\rm i}}{3a_{\rm i}} \Omega_{\rm m} + \Omega_{\rm k}. \tag{75}$$

If K is sufficiently negative, the particle will turn-around and the sphere will collapse at a time

$$H_0 t_{\rm coll} = 2 \int_0^{a_{\rm max}} {\rm d}a \left(\Omega_{\rm m}/a + K + \Omega_\Lambda a^2\right)^{-1/2}, \tag{76}$$

where $a_{\rm max}$ is the value of a which sets the denominator of the integral to zero.

For the case of $\Lambda = 0$, we can determine the spherical collapse parameters A and B. $K > 0$ ($K < 0$) produces an open (closed) model. Comparing coefficients in the energy equations [(74) and the integration of (66)], one finds

$$A = \frac{\Omega_m r_\mathrm{i}}{2a_\mathrm{i}} \left| \frac{5\delta_\mathrm{i}}{3a_\mathrm{i}} \Omega_m - \Omega_k \right|^{-1} \tag{77}$$

$$B = \frac{\Omega_m}{2H_0} \left| \frac{5\delta_\mathrm{i}}{3a_\mathrm{i}} \Omega_m - \Omega_k \right|^{-3/2}, \tag{78}$$

where $\Omega_\mathrm{k} = 1 - \Omega_\mathrm{m}$. In particular, in an $\Omega = 1$ Universe, where $1 + z = (3H_0 t/2)^{-2/3}$, we find that a shell collapses at redshift $1 + z_\mathrm{c} = 0.5929\delta_\mathrm{i}/a_\mathrm{i}$, or in other words a shell collapsing at redshift z_c had a linear overdensity extrapolated to the present day of $\delta_0 = 1.686(1 + z_\mathrm{c})$.

While this derivation has been for spheres of constant density, we may treat a general spherical density profile $\delta_\mathrm{i}(r)$ up until shell crossing [164]. A particular radial shell evolves according to the mass interior to it; therefore, we define the average overdensity $\overline{\delta_\mathrm{i}}$

$$\overline{\delta_\mathrm{i}}(R) = \frac{3}{4\pi R^3} \int_0^R \mathrm{d}^3 r \delta_\mathrm{i}(r), \tag{79}$$

so that we may use $\overline{\delta_\mathrm{i}}$ in place of δ_i in the above formulae. If $\overline{\delta_\mathrm{i}}$ is not monotonically decreasing with R, then the spherical top-hat evolution of two different radii will predict that they cross each other at some late time; this is known as shell crossing and signals the breakdown of the solution. Even well-behaved $\overline{\delta_\mathrm{i}}$ profiles will produce shell crossing if shells are allowed to collapse to $r = 0$ and then reexpand, since these expanding shells will cross infalling shells. In such a case, first-time infalling shells will never be affected prior to their turn-around; the more complicated behavior after turn-around is a manifestation of virialization. While the end state for general initial conditions cannot be predicted, various results are known for a self-similar collapse, in which $\delta(r)$ is a power-law [132, 40], as well as for the case of secondary infall models [156, 165, 179].

Halo Properties

The small density fluctuations evidenced in the CMB grow over time as described in the previous subsection, until the perturbation δ becomes of order unity, and the full non-linear gravitational problem must be considered. The dynamical collapse of a dark matter halo can be solved analytically only in cases of particular symmetry. If we consider a region which is much smaller than the horizon cH^{-1}, then the formation of a halo can be formulated as a problem in Newtonian gravity, in some cases with minor corrections coming from General Relativity. The simplest case is that of spherical symmetry, with an initial ($t = t_\mathrm{i} \ll t_0$) top-hat of uniform overdensity δ_i inside a sphere

of radius R. Although this model is restricted in its direct applicability, the results of spherical collapse have turned out to be surprisingly useful in understanding the properties and distribution of halos in models based on cold dark matter.

The collapse of a spherical top-hat perturbation is described by the Newtonian equation (with a correction for the cosmological constant)

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = H_0^2 \Omega_\Lambda \, r - \frac{GM}{r^2} \ , \tag{80}$$

where r is the radius in a fixed (not comoving) coordinate frame, H_0 is the present-day Hubble constant, M is the total mass enclosed within radius r, and the initial velocity field is given by the Hubble flow $\mathrm{d}r/\mathrm{d}t = H(t)r$. The enclosed δ grows initially as $\delta_\mathrm{L} = \delta_\mathrm{i} D(t)/D(t_\mathrm{i})$, in accordance with linear theory, but eventually δ grows above δ_L. If the mass shell at radius r is bound (i.e., if its total Newtonian energy is negative) then it reaches a radius of maximum expansion and subsequently collapses. As demonstrated in the previous section, at the moment when the top-hat collapses to a point, the overdensity predicted by linear theory is $\delta_\mathrm{L} = 1.686$ in the Einstein-de Sitter model, with only a weak dependence on Ω_m and Ω_Λ. Thus a top-hat collapses at redshift z if its linear overdensity extrapolated to the present day (also termed the critical density of collapse) is

$$\delta_\mathrm{crit}(z) = \frac{1.686}{D(z)} \ , \tag{81}$$

where we set $D(z = 0) = 1$.

Even a slight violation of the exact symmetry of the initial perturbation can prevent the top-hat from collapsing to a point. Instead, the halo reaches a state of virial equilibrium by violent relaxation (phase mixing). Using the virial theorem $U = -2K$ to relate the potential energy U to the kinetic energy K in the final state (implying that the virial radius is half the turnaround radius – where the kinetic energy vanishes), the final overdensity relative to the critical density at the collapse redshift is $\Delta_\mathrm{c} = 18\pi^2 \simeq 178$ in the Einstein-de Sitter model, modified in a Universe with $\Omega_\mathrm{m} + \Omega_\Lambda = 1$ to the fitting formula (Bryan & Norman 1998 [71])

$$\Delta_\mathrm{c} = 18\pi^2 + 82d - 39d^2 \ , \tag{82}$$

where $d \equiv \Omega_\mathrm{m}^z - 1$ is evaluated at the collapse redshift, so that

$$\Omega_\mathrm{m}^z = \frac{\Omega_\mathrm{m}(1+z)^3}{\Omega_\mathrm{m}(1+z)^3 + \Omega_\Lambda + \Omega_\mathrm{k}(1+z)^2} \ . \tag{83}$$

A halo of mass M collapsing at redshift z thus has a virial radius

$$r_\mathrm{vir} = 0.784 \left(\frac{M}{10^8 \ h^{-1} \ \mathrm{M}_\odot} \right)^{1/3} \left[\frac{\Omega_\mathrm{m}}{\Omega_\mathrm{m}^z} \ \frac{\Delta_c}{18\pi^2} \right]^{-1/3} \left(\frac{1+z}{10} \right)^{-1} \ h^{-1} \ \mathrm{kpc} \ , \tag{84}$$

and a corresponding circular velocity,

$$V_c = \left(\frac{GM}{r_{\rm vir}}\right)^{1/2} = 23.4 \left(\frac{M}{10^8\ h^{-1}\ {\rm M}_\odot}\right)^{1/3} \left[\frac{\Omega_{\rm m}}{\Omega_{\rm m}^z}\ \frac{\Delta_c}{18\pi^2}\right]^{1/6} \left(\frac{1+z}{10}\right)^{1/2}\ {\rm km\,s^{-1}}\ . \tag{85}$$

In these expressions we have assumed a present Hubble constant written in the form $H_0 = 100\,{\rm h\,km\,s^{-1}\,Mpc^{-1}}$. We may also define a virial temperature

$$T_{\rm vir} = \frac{\mu m_{\rm p} V_{\rm c}^2}{2k} = 1.98 \times 10^4\ \left(\frac{\mu}{0.6}\right) \left(\frac{M}{10^8\ h^{-1}\ {\rm M}_\odot}\right)^{2/3} \left[\frac{\Omega_{\rm m}}{\Omega_{\rm m}^z}\ \frac{\Delta_c}{18\pi^2}\right]^{1/3} \left(\frac{1+z}{10}\right)\ {\rm K}\ , \tag{86}$$

where μ is the mean molecular weight and $m_{\rm p}$ is the proton mass. Note that the value of μ depends on the ionization fraction of the gas; for a fully ionized primordial gas $\mu = 0.59$, while a gas with ionized hydrogen but only singly-ionized helium has $\mu = 0.61$. The binding energy of the halo is approximately[5]

$$E_{\rm b} = \frac{1}{2}\frac{GM^2}{r_{\rm vir}} = 5.45 \times 10^{53} \left(\frac{M}{10^8\ h^{-1}\ {\rm M}_\odot}\right)^{5/3} \left[\frac{\Omega_{\rm m}}{\Omega_{\rm m}^z}\ \frac{\Delta_c}{18\pi^2}\right]^{1/3} \times \left(\frac{1+z}{10}\right) h^{-1}\,{\rm erg}\ . \tag{87}$$

Note that the binding energy of the baryons is smaller by a factor equal to the baryon fraction $\Omega_{\rm b}/\Omega_{\rm m}$.

Although spherical collapse captures some of the physics governing the formation of halos, structure formation in cold dark matter models procedes hierarchically. At early times, most of the dark matter is in low-mass halos, and these halos continuously accrete and merge to form high-mass halos. Numerical simulations of hierarchical halo formation indicate a roughly universal spherically-averaged density profile for the resulting halos (Navarro, Frenk, & White 1997, hereafter NFW [265]), though with considerable scatter among different halos (e.g., [72]). The NFW profile has the form

$$\rho(r) = \frac{3H_0^2}{8\pi G}(1+z)^3 \frac{\Omega_{\rm m}}{\Omega_{\rm m}^z} \frac{\delta_c}{c_{\rm N} x (1 + c_{\rm N} x)^2}\ , \tag{88}$$

where $x = r/r_{\rm vir}$, and the characteristic density $\delta_{\rm c}$ is related to the concentration parameter $c_{\rm N}$ by

$$\delta_{\rm c} = \frac{\Delta_{\rm c}}{3} \frac{c_{\rm N}^3}{\ln(1 + c_{\rm N}) - c_{\rm N}/(1 + c_{\rm N})}\ . \tag{89}$$

[5] The coefficient of 1/2 in (87) would be exact for a singular isothermal sphere, $\rho(r) \propto 1/r^2$.

The concentration parameter itself depends on the halo mass M, at a given redshift z [376].

More recent N-body simulations indicate deviations from the original NFW profile; for details and refined fitting formula see [267].

4 Nonlinear Growth

4.1 The Abundance of Dark Matter Halos

In addition to characterizing the properties of individual halos, a critical prediction of any theory of structure formation is the abundance of halos, i.e. the number density of halos as a function of mass, at any redshift. This prediction is an important step toward inferring the abundances of galaxies and galaxy clusters. While the number density of halos can be measured for particular cosmologies in numerical simulations, an analytic model helps us gain physical understanding and can be used to explore the dependence of abundances on all the cosmological parameters.

A simple analytic model which successfully matches most of the numerical simulations was developed by Press & Schechter (1974) [290]. The model is based on the ideas of a Gaussian random field of density perturbations, linear gravitational growth, and spherical collapse. To determine the abundance of halos at a redshift z, we use $\delta_{\rm M}$, the density field smoothed on a mass scale M, as defined in Sect. 3.3. Since $\delta_{\rm M}$ is distributed as a Gaussian variable with zero mean and standard deviation $\sigma(M)$ [which depends only on the present linear power spectrum, see (17)], the probability that $\delta_{\rm M}$ is greater than some δ equals

$$\int_{\delta}^{\infty} {\rm d}\delta_{\rm M} \frac{1}{\sqrt{2\pi}\,\sigma(M)} \exp\left[-\frac{\delta_{\rm M}^2}{2\,\sigma^2(M)}\right] = \frac{1}{2}{\rm erfc}\left(\frac{\delta}{\sqrt{2}\,\sigma(M)}\right) \ . \qquad (90)$$

The fundamental ansatz is to identify this probability with the fraction of dark matter particles which are part of collapsed halos of mass greater than M, at redshift z. There are two additional ingredients: First, the value used for δ is $\delta_{\rm crit}(z)$ given in (81), which is the critical density of collapse found for a spherical top-hat (extrapolated to the present since $\sigma(M)$ is calculated using the present power spectrum); and second, the fraction of dark matter in halos above M is multiplied by an additional factor of 2 in order to ensure that every particle ends up as part of some halo with $M > 0$. Thus, the final formula for the mass fraction in halos above M at redshift z is

$$F(>M|z) = {\rm erfc}\left(\frac{\delta_{\rm crit}(z)}{\sqrt{2}\,\sigma(M)}\right) \ . \qquad (91)$$

This ad-hoc factor of 2 is necessary, since otherwise only positive fluctuations of δ_M would be included. Bond et al. (1991) [52] found an alternate

derivation of this correction factor, using a different ansatz. In their derivation, the factor of 2 has a more satisfactory origin, namely the so-called "cloud-in-cloud" problem: For a given mass M, even if $\delta_{\rm M}$ is smaller than $\delta_{\rm crit}(z)$, it is possible that the corresponding region lies inside a region of some larger mass $M_{\rm L} > M$, with $\delta_{M_{\rm L}} > \delta_{\rm crit}(z)$. In this case the original region should be counted as belonging to a halo of mass M_L. Thus, the fraction of particles which are part of collapsed halos of mass greater than M is larger than the expression given in (90). Bond et al. showed that, under certain assumptions, the additional contribution results precisely in a factor of 2 correction.

Differentiating the fraction of dark matter in halos above M yields the mass distribution. Letting dn be the comoving number density of halos of mass between M and $M + {\rm d}M$, we have

$$\frac{{\rm d}n}{{\rm d}M} = \sqrt{\frac{2}{\pi}}\,\frac{\rho_{\rm m}}{M}\,\frac{-{\rm d}(\ln\sigma)}{{\rm d}M}\,\nu_{\rm c}\,{\rm e}^{-\nu_{\rm c}^2/2}\ , \tag{92}$$

where $\nu_{\rm c} = \delta_{\rm crit}(z)/\sigma(M)$ is the number of standard deviations which the critical collapse overdensity represents on mass scale M. Thus, the abundance of halos depends on the two functions $\sigma(M)$ and $\delta_{\rm crit}(z)$, each of which depends on the energy content of the Universe and the values of the other cosmological parameters. Since recent observations confine the standard set of parameters to a relatively narrow range, we illustrate the abundance of halos and other results for a single set of parameters: $\Omega_{\rm m} = 0.3$, $\Omega_\Lambda = 0.7$, $\Omega_{\rm b} = 0.045$, $\sigma_8 = 0.9$, a primordial power spectrum index $n = 1$ and a Hubble constant $h = 0.7$.

Figure 13 shows $\sigma(M)$ and $\delta_{\rm crit}(z)$, with the input power spectrum computed from Eisenstein & Hu (1999) [118]. The solid line is $\sigma(M)$ for the cold dark matter model with the parameters specified above. The horizontal dotted lines show the value of $\delta_{\rm crit}(z)$ at $z = 0, 2, 5, 10, 20$ and 30, as indicated in the figure. From the intersection of these horizontal lines with the solid line we infer, e.g., that at $z = 5$ a $1-\sigma$ fluctuation on a mass scale of 2×10^7 M$_\odot$ will collapse. On the other hand, at $z = 5$ collapsing halos require a $2-\sigma$ fluctuation on a mass scale of 3×10^{10} M$_\odot$, since $\sigma(M)$ on this mass scale equals about half of $\delta_{\rm crit}(z = 5)$. Since at each redshift a fixed fraction (31.7%) of the total dark matter mass lies in halos above the $1-\sigma$ mass, Fig. 13 shows that most of the mass is in small halos at high redshift, but it continuously shifts toward higher characteristic halo masses at lower redshift. Note also that $\sigma(M)$ flattens at low masses because of the changing shape of the power spectrum. Since $\sigma \to \infty$ as $M \to 0$, in the cold dark matter model all the dark matter is tied up in halos at all redshifts, if sufficiently low-mass halos are considered.

Also shown in Fig. 13 is the effect of cutting off the power spectrum on small scales. The short-dashed curve corresponds to the case where the power spectrum is set to zero above a comoving wavenumber $k = 10\,{\rm Mpc}^{-1}$, which corresponds to a mass $M = 1.7 \times 10^8$ M$_\odot$. The long-dashed curve corresponds to a more radical cutoff above $k = 1\,{\rm Mpc}^{-1}$, or below $M = 1.7 \times 10^{11}$ M$_\odot$.

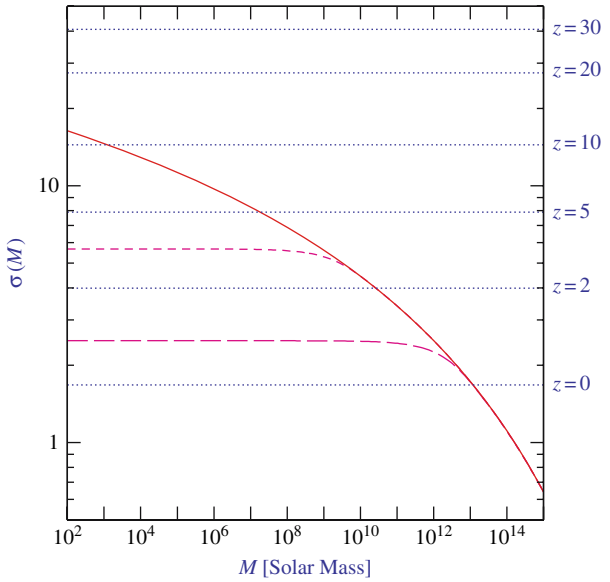

Fig. 13. Mass fluctuations and collapse thresholds in cold dark matter models (from Barkana & Loeb 2001 [23]). The horizontal dotted lines show the value of the extrapolated collapse overdensity $\delta_{\rm crit}(z)$ at the indicated redshifts. Also shown is the value of $\sigma(M)$ for the cosmological parameters given in the text (*solid curve*), as well as $\sigma(M)$ for a power spectrum with a cutoff below a mass $M = 1.7 \times 10^8$ $M_\odot$ (*short-dashed curve*), or $M = 1.7 \times 10^{11}$ $M_\odot$ (*long-dashed curve*). The intersection of the horizontal lines with the other curves indicate, at each redshift z, the mass scale (for each model) at which a $1-\sigma$ fluctuation is just collapsing at z (see the discussion in the text)

A cutoff severely reduces the abundance of low-mass halos, and the finite value of $\sigma(M = 0)$ implies that at all redshifts some fraction of the dark matter does not fall into halos. At high redshifts where $\delta_{\rm crit}(z) \gg \sigma(M = 0)$, all halos are rare and only a small fraction of the dark matter lies in halos. In particular, this can affect the abundance of halos at the time of reionization, and thus the observed limits on reionization constrain scenarios which include a small-scale cutoff in the power spectrum [21].

In Figs. 14–17 we show explicitly the properties of collapsing halos which represent $1-\sigma$, $2-\sigma$, and $3-\sigma$ fluctuations (corresponding in all cases to the curves in order from bottom to top), as a function of redshift. No cutoff is applied to the power spectrum. Figure 14 shows the halo mass, Fig. 15 the virial radius, Fig. 16 the virial temperature (with μ in (86) set equal to 0.6, although low temperature halos contain neutral gas) as well as circular velocity, and Fig. 17 shows the total binding energy of these halos. In Figs. 14 and 16, the dotted curves indicate the minimum virial temperature required for efficient cooling with primordial atomic species only (upper curve) or with

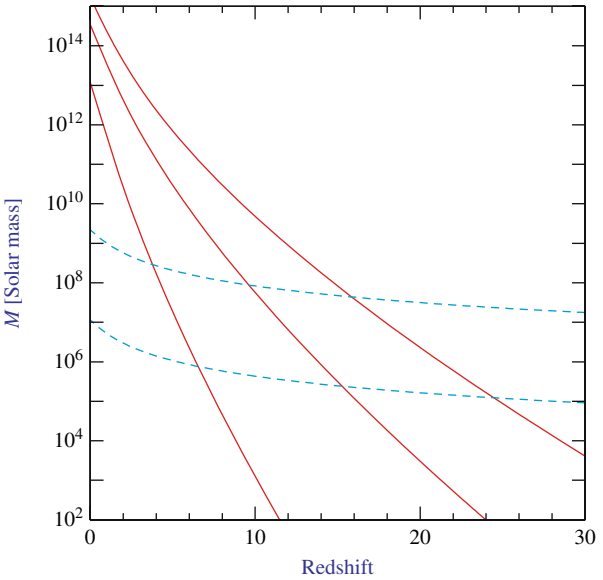

Fig. 14. Characteristic properties of collapsing halos: Halo mass (from Barkana & Loeb 2001 [23]). The solid curves show the mass of collapsing halos which correspond to $1-\sigma$, $2-\sigma$, and $3-\sigma$ fluctuations (**in order from bottom to top**). The dotted curves show the mass corresponding to the minimum temperature required for efficient cooling with primordial atomic species only (*upper curve*) or with the addition of molecular hydrogen (*lower curve*)

the addition of molecular hydrogen (lower curve). Figure 17 shows the binding energy of dark matter halos. The binding energy of the baryons is a factor $\sim \Omega_{\rm b}/\Omega_{\rm m} \sim 15\%$ smaller, if they follow the dark matter. Except for this constant factor, the figure shows the minimum amount of energy that needs to be deposited into the gas in order to unbind it from the potential well of the dark matter. For example, the hydrodynamic energy released by a single supernovae, $\sim 10^{51}$ erg, is sufficient to unbind the gas in all $1-\sigma$ halos at $z \gtrsim 5$ and in all $2-\sigma$ halos at $z \gtrsim 12$.

At $z = 5$, the halo masses which correspond to $1-\sigma$, $2-\sigma$, and $3-\sigma$ fluctuations are 1.8×10^7 M$_\odot$, 3.0×10^{10} M$_\odot$, and 7.0×10^{11} M$_\odot$, respectively. The corresponding virial temperatures are 2.0×10^3 K, 2.8×10^5 K, and 2.3×10^6 K. The equivalent circular velocities are $7.5\,{\rm km\,s^{-1}}$, $88\,{\rm km\,s^{-1}}$, and $250\,{\rm km\,s^{-1}}$. At $z = 10$, the $1-\sigma$, $2-\sigma$, and $3-\sigma$ fluctuations correspond to halo masses of 1.3×10^3 M$_\odot$, 5.7×10^7 M$_\odot$, and 4.8×10^9 M$_\odot$, respectively. The corresponding virial temperatures are 6.2 K, 7.9×10^3 K, and 1.5×10^5 K. The equivalent circular velocities are $0.41\,{\rm km\,s^{-1}}$, $15\,{\rm km\,s^{-1}}$, and $65\,{\rm km\,s^{-1}}$. Atomic cooling is efficient at $T_{\rm vir} \gtrsim 10^4$ K, or a circular velocity $V_{\rm c} \gtrsim 17\,{\rm km\,s^{-1}}$. This corresponds to a $1.2-\sigma$ fluctuation and a halo mass of 2.1×10^8 M$_\odot$ at $z = 5$, and a $2.1-\sigma$ fluctuation and a halo mass of 8.3×10^7 M$_\odot$ at $z = 10$. Molecular

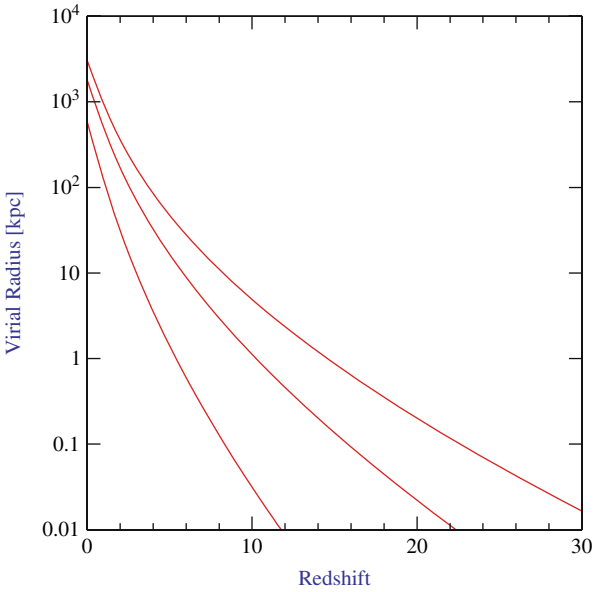

Fig. 15. Characteristic properties of collapsing halos: Halo virial radius (from Barkana & Loeb 2001 [23]). The curves show the virial radius of collapsing halos which correspond to $1-\sigma$, $2-\sigma$, and $3-\sigma$ fluctuations (**in order from bottom to top**)

hydrogen provides efficient cooling down to $T_{\rm vir} \sim 300\,{\rm K}$, or a circular velocity $V_{\rm c} \sim 2.9\,{\rm km\,s^{-1}}$. This corresponds to a $0.81-\sigma$ fluctuation and a halo mass of $1.1 \times 10^6\ {\rm M}_\odot$ at $z = 5$, and a $1.4-\sigma$ fluctuation and a halo mass of $4.3 \times 10^5\ {\rm M}_\odot$ at $z = 10$.

In Fig. 18 we show the halo mass function ${\rm d}n/{\rm d}\ln(M)$ at several different redshifts: $z = 0$ (solid curve), $z = 5$ (dotted curve), $z = 10$ (short-dashed curve), $z = 20$ (long-dashed curve), and $z = 30$ (dot-dashed curve). Note that the mass function does not decrease monotonically with redshift at all masses. At the lowest masses, the abundance of halos is higher at $z > 0$ than at $z = 0$.

4.2 The Excursion-Set (Extended Press-Schechter) Formalism

The usual Press-Schechter formalism makes no attempt to deal with the correlations between halos or between different mass scales. In particular, this means that while it can generate a distribution of halos at two different epochs, it says nothing about how particular halos in one epoch are related to those in the second. We therefore would like some method to predict, at least statistically, the growth of individual halos via accretion and mergers. Even restricting ourselves to spherical collapse, such a model must utilize the full spherically-averaged density profile around a particular point. The

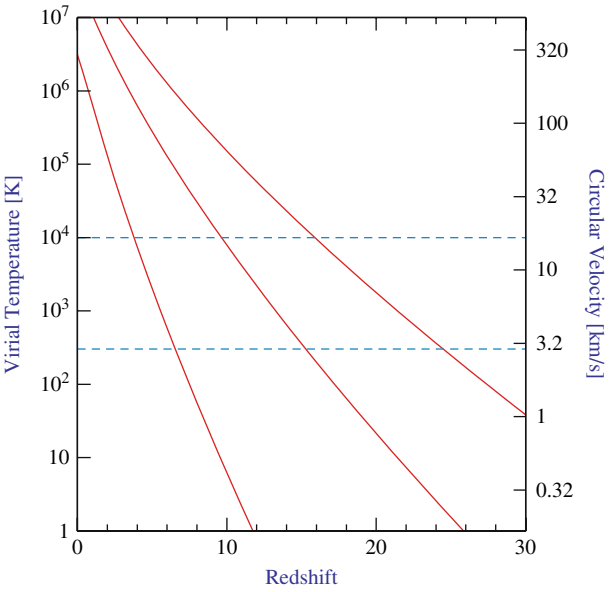

Fig. 16. Characteristic properties of collapsing halos: Halo virial temperature and circular velocity (from Barkana & Loeb 2001 [23]). The solid curves show the virial temperature (or, equivalently, the circular velocity) of collapsing halos which correspond to $1-\sigma$, $2-\sigma$, and $3-\sigma$ fluctuations (**in order from bottom to top**). The dotted curves show the minimum temperature required for efficient cooling with primordial atomic species only (*upper curve*) or with the addition of molecular hydrogen (*lower curve*)

potential correlations between the mean overdensities at different radii make the statistical description substantially more difficult.

The excursion set formalism (Bond et al. 1991 [52]) seeks to describe the statistics of halos by considering the statistical properties of $\overline{\delta}(R)$, the average overdensity within some spherical window of characteristic radius R, as a function of R. While the Press-Schechter model depends only on the Gaussian distribution of $\overline{\delta}$ for one particular R, the excursion set considers all R. Again the connection between a value of the linear regime δ and the final state is made via the spherical collapse solution, so that there is a critical value $\delta_{\rm c}(z)$ of $\overline{\delta}$ which is required for collapse at a redshift z.

For most choices of the window function, the functions $\overline{\delta}(R)$ are correlated from one R to another such that it is prohibitively difficult to calculate the desired statistics directly (although Monte Carlo realizations are possible [52]). However, for one particular choice of a window function, the correlations between different R greatly simplify and many interesting quantities may be calculated [52, 210]. The key is to use a k-space top-hat window function, namely $W_{\rm k} = 1$ for all k less than some critical $k_{\rm c}$ and $W_{\rm k} = 0$ for $k > k_{\rm c}$.

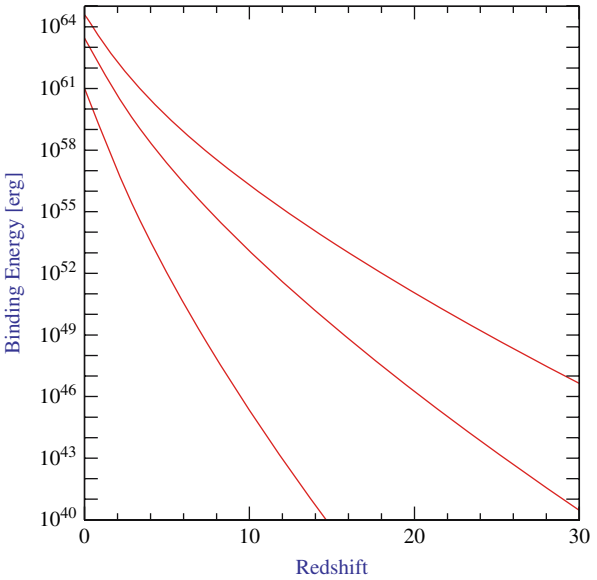

Fig. 17. Characteristic properties of collapsing halos: Halo binding energy (from Barkana & Loeb 2001 [23]). The curves show the total binding energy of collapsing halos which correspond to $1-\sigma$, $2-\sigma$, and $3-\sigma$ fluctuations (**in order from bottom to top**)

This filter has a spatial form of $W(r) \propto j_1(k_{\rm c}r)/k_{\rm c}r$, which implies a volume $6\pi^2/k_{\rm c}^3$ or mass $6\pi^2\rho_{\rm b}/k_{\rm c}^3$. The characteristic radius of the filter is $\sim k_{\rm c}^{-1}$, as expected. Note that in real space, this window function converges very slowly, due only to a sinusoidal oscillation, so the region under study is rather poorly localized.

The great advantage of the sharp k-space filter is that the difference at a given point between $\overline{\delta}$ on one mass scale and that on another mass scale is statistically independent from the value on the larger mass scale. With a Gaussian random field, each $\delta_{\rm k}$ is Gaussian distributed independently from the others. For this filter,

$$\overline{\delta}(M) = \int_{k<k_{\rm c}(M)} \frac{{\rm d}^3k}{(2\pi)^3}\delta_k, \tag{93}$$

meaning that the overdensity on a particular scale is simply the sum of the random variables $\delta_{\rm k}$ interior to the chosen $k_{\rm c}$. Consequently, the difference between the $\overline{\delta}(M)$ on two mass scales is just the sum of the $\delta_{\rm k}$ in the spherical shell between the two $k_{\rm c}$, which is independent from the sum of the $\delta_{\rm k}$ interior to the smaller $k_{\rm c}$. Meanwhile, the distribution of $\overline{\delta}(M)$ given no prior information is still a Gaussian of mean zero and variance

$$\sigma^2(M) = \frac{1}{2\pi^2}\int_{k<k_{\rm c}(M)} {\rm d}k\, k^2 P(k). \tag{94}$$

If we now consider $\overline{\delta}$ as a function of scale k_c, we see that we begin from $\overline{\delta} = 0$ at $k_c = 0$ ($M = \infty$) and then add independently random pieces as k_c increases. This generates a random walk, albeit one whose stepsize varies with k_c. We then assume that, at redshift z, a given function $\overline{\delta}(k_c)$ represents a collapsed mass M corresponding to the k_c where the function first crosses the critical value $\delta_c(z)$. With this assumption, we may use the properties of random walks to calculate the evolution of the mass as a function of redshift.

It is now easy to rederive the Press-Schechter mass function, including the previously unexplained factor of 2 [52, 210, 381]. The fraction of mass elements included in halos of mass less than M is just the probability that a random walk remains below $\delta_c(z)$ for all k_c less than K_c, the filter cutoff appropriate to M. This probability must be the complement of the sum of the probabilities that: (a) $\overline{\delta}(K_c) > \delta_c(z)$; or that (b) $\overline{\delta}(K_c) < \delta_c(z)$ but $\overline{\delta}(k_c') > \delta_c(z)$ for some $k_c' < K_c$. But these two cases in fact have equal probability; any random walk belonging to class (a) may be reflected around its first upcrossing of $\delta_c(z)$ to produce a walk of class (b), and vice versa. Since the distribution of $\overline{\delta}(K_c)$ is simply Gaussian with variance $\sigma^2(M)$, the fraction of random walks falling into class (a) is simply $(1/\sqrt{2\pi\sigma^2}) \int_{\delta_c(z)}^{\infty} \mathrm{d}\delta \ \exp\{-\delta^2/2\sigma^2(M)\}$. Hence, the fraction of mass elements included in halos of mass less than M at redshift z is simply

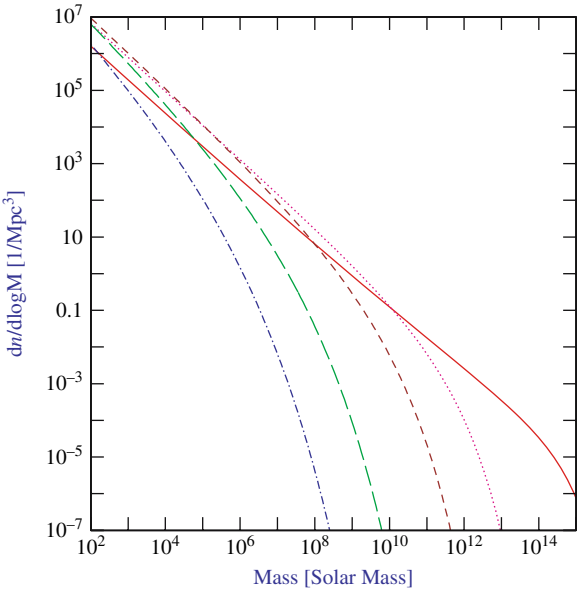

Fig. 18. Halo mass function at several redshifts: $z = 0$ (*solid curve*), $z = 5$ (*dotted curve*), $z = 10$ (*short-dashed curve*), $z = 20$ (*long-dashed curve*), and $z = 30$ (*dot-dashed curve*). (From Barkana & Loeb 2001 [23])

$$F(<M) = 1 - 2 \times \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\delta_{\rm c}(z)}^{\infty} {\rm d}\delta \, \exp\{-\delta^2/2\sigma^2(M)\} \tag{95}$$

which may be differentiated to yield the Press-Schechter mass function. We may now go further and consider how halos at one redshift are related to those at another redshift. If we are given that a halo of mass M_2 exists at redshift z_2, then we know that the random function $\overline{\delta}(k_{\rm c})$ for each mass element within the halo first crosses $\delta(z_2)$ at $k_{\rm c2}$ corresponding to M_2. Given this constraint, we may study the distribution of $k_{\rm c}$ where the function $\overline{\delta}(k_{\rm c})$ crosses other thresholds. It is particularly easy to construct the probability distribution for when trajectories first cross some $\delta_{\rm c}(z_1) > \delta_{\rm c}(z_2)$ (implying $z_1 > z_2$); clearly this occurs at some $k_{\rm c1} > k_{\rm c2}$. This problem reduces to the previous one if we translate the origin of the random walks from $(k_{\rm c}, \overline{\delta}) = (0, 0)$ to $(k_{\rm c2}, \delta_{\rm c}(z_2))$. We therefore find the probability of a halo masses M_1 finding itself in at redshift z_1, given that it is part of a larger halo of mass M_2 at a later redshift z_2 is [52, 55])

$$\frac{{\rm d}P}{{\rm d}M_1}(M_1, z_1 | M_2, z_2) = \sqrt{\frac{2}{\pi}} \frac{\delta_{\rm c}(z_1) - \delta_{\rm c}(z_2)}{[\sigma^2(M_1) - \sigma^2(M_2)]^{3/2}} \left| \frac{{\rm d}\sigma(M_1)}{{\rm d}M_1} \right| \exp\left\{ -\frac{[\delta_{\rm c}(z_1) - \delta_{\rm c}(z_2)]^2}{2[\sigma^2(M_1) - \sigma^2(M_2)]} \right\}. \tag{96}$$

This may be rewritten as saying that the quantity

$$\tilde{v} = \frac{\delta_{\rm c}(z_1) - \delta_{\rm c}(z_2)}{\sqrt{\sigma^2(M_1) - \sigma^2(M_2)}} \tag{97}$$

is distributed as the positive half of a Gaussian with unit variance; (97) may be inverted to find $M_1(\tilde{v})$.

We seek to interpret the statistics of these random walks as those of merging and accreting halos. For a single halo, we may imagine that as we look back in time, the object breaks into ever smaller pieces, similar to the branching of a tree. Equation (96) is the distribution of the sizes of these branches at some given earlier time. However, using this description of the ensemble distribution to generate Monte Carlo realizations of single merger trees has proven to be difficult. In all cases, one recursively steps back in time, at each step breaking the final object into two or more pieces. An elaborate scheme (Kauffmann & White 1993 [193]) picks a large number of progenitors from the ensemble distribution and then randomly groups them into sets with the correct total mass. This generates many (hundreds) possible branching schemes of equal likelihood. A simpler scheme (Lacey & Cole 1993 [210]) assumes that at each time step, the object breaks into two pieces. One value from the distribution (96) then determines the mass ratio of the two branchs.

One may also use the distribution of the ensemble to derive some additional analytic results. A useful example is the distribution of the epoch at which an object that has mass M_2 at redshift z_2 has accumulated half of its mass

[210]. The probability that the formation time is earlier than z_1 is equal to the probability that at redshift z_1 a progenitor whose mass exceeds $M_2/2$ exists:

$$P(z_{\rm f} > z_1) = \int_{M_2/2}^{M_2} \frac{M_2}{M} \frac{{\rm d}P}{{\rm d}M}(M, z_1|M_2, z_2){\rm d}M, \tag{98}$$

where ${\rm d}P/{\rm d}M$ is given in (96). The factor of M_2/M corrects the counting from mass weighted to number weighted; each halo of mass M_2 can have only one progenitor of mass greater than $M_2/2$. Differentiating (98) with respect to time gives the distribution of formation times. This analytic form is an excellent match to scale-free N-body simulations [211]. On the other hand, simple Monte Carlo implementations of (96) produce formation redshifts about 40% higher [210]. As there may be correlations between the various branches, there is no unique Monte Carlo scheme.

Numerical tests of the excursion set formalism are quite encouraging. Its predictions for merger rates are in very good agreement with those measured in scale-free N-body simulations for mass scales down to around 10% of the nonlinear mass scale (that scale at which $\sigma_{\rm M} = 1$), and distributions of formation times closely match the analytic predictions [211]. The model appears to be a promising method for tracking the merging of halos, with many applications to cluster and galaxy formation modeling. In particular, one may use the formalism as the foundation of semi-analytic galaxy formation models [194]. The excursion set formalism may also be used to derive the correlations of halos in the nonlinear regime [257].

4.3 Response of Baryons to Nonlinear Dark Matter Potentials

The dark matter is assumed to be cold and to dominate gravity, and so its collapse and virialization proceeds unimpeded by pressure effects. In order to estimate the minimum mass of baryonic objects, we must go beyond linear perturbation theory and examine the baryonic mass that can accrete into the final gravitational potential well of the dark matter.

For this purpose, we assume that the dark matter had already virialized and produced a gravitational potential $\phi(\mathbf{r})$ at a redshift $z_{\rm vir}$ (with $\phi \to 0$ at large distances, and $\phi < 0$ inside the object) and calculate the resulting overdensity in the gas distribution, ignoring cooling (an assumption justified by spherical collapse simulations which indicate that cooling becomes important only after virialization; see Haiman et al. 1996a [167]).

After the gas settles into the dark matter potential well, it satisfies the hydrostatic equilibrium equation,

$$\nabla p_{\rm b} = -\rho_{\rm b} \nabla \phi \tag{99}$$

where $p_{\rm b}$ and $\rho_{\rm b}$ are the pressure and mass density of the gas. At $z \lesssim 100$ the gas temperature is decoupled from the CMB, and its pressure evolves adiabatically (ignoring atomic or molecular cooling),

$$\frac{p_{\rm b}}{\bar{p}_{\rm b}} = \left(\frac{\rho_{\rm b}}{\bar{\rho}_{\rm b}}\right)^{5/3} \tag{100}$$

where a bar denotes the background conditions. We substitute (100) into (99) and get the solution,

$$\frac{\rho_{\rm b}}{\bar{\rho}_{\rm b}} = \left(1 - \frac{2}{5}\frac{\mu m_{\rm p}\phi}{k\bar{T}}\right)^{3/2} \tag{101}$$

where $\bar{T} = \bar{p}_{\rm b}\mu m_{\rm p}/(k\bar{\rho}_{\rm b})$ is the background gas temperature. If we define $T_{\rm vir} = -\frac{1}{3}m_{\rm p}\phi/k$ as the virial temperature for a potential depth $-\phi$, then the overdensity of the baryons at the virialization redshift is

$$\delta_{\rm b} = \frac{\rho_{\rm b}}{\bar{\rho}_{\rm b}} - 1 = \left(1 + \frac{6}{5}\frac{T_{\rm vir}}{\bar{T}}\right)^{3/2} - 1. \tag{102}$$

This solution is approximate for two reasons: (i) we assumed that the gas is stationary throughout the entire region and ignored the transitions to infall and the Hubble expansion at the interface between the collapsed object and the background intergalactic medium (henceforth IGM), and (ii) we ignored entropy production at the virialization shock surrounding the object. Nevertheless, the result should provide a better estimate for the minimum mass of collapsed baryonic objects than the Jeans mass does, since it incorporates the nonlinear potential of the dark matter.

We may define the threshold for the collapse of baryons by the criterion that their mean overdensity, $\delta_{\rm b}$, exceeds a value of 100, amounting to $\gtrsim 50\%$ of the baryons that would assemble in the absence of gas pressure, according to the spherical top-hat collapse model. Equation (102) then implies that $T_{\rm vir} > 17.2\,\bar{T}$.

As mentioned before, the gas temperature evolves at $z \lesssim 160$ according to the relation $\bar{T} \approx 170[(1+z)/100]^2\,{\rm K}$. This implies that baryons are overdense by $\delta_{\rm b} > 100$ only inside halos with a virial temperature $T_{\rm vir} \gtrsim 2.9 \times 10^3\ [(1+z)/100]^2\,{\rm K}$. Based on the top-hat model, this implies a minimum halo mass for baryonic objects of

$$M_{\rm min} = 5.0 \times 10^3 \left(\frac{\Omega_{\rm m}h^2}{0.15}\right)^{-1/2} \left(\frac{\Omega_{\rm b}h^2}{0.022}\right)^{-3/5} \left(\frac{1+z}{10}\right)^{3/2} {\rm M}_\odot, \tag{103}$$

where we consider sufficiently high redshifts so that $\Omega_{\rm m}^z \approx 1$. This minimum mass is coincidentally almost identical to the naive Jeans mass calculation of linear theory in (62) despite the fact that it incorporates shell crossing by the dark matter, which is not accounted for by linear theory. Unlike the Jeans mass, the minimum mass depends on the choice for an overdensity threshold [taken arbitrarily as $\delta_{\rm b} > 100$ in (103)]. To estimate the minimum halo mass which produces any significant accretion we set, e.g., $\delta_{\rm b} = 5$, and get a mass which is lower than $M_{\rm min}$ by a factor of 27.

Of course, once the first stars and quasars form they heat the surrounding IGM by either outflows or radiation. As a result, the Jeans mass which is relevant for the formation of new objects changes [148, 152]). The most dramatic change occurs when the IGM is photo-ionized and is consequently heated to a temperature of $\sim (1\text{–}2) \times 10^4\,\mathrm{K}$.

5 Fragmentation of the First Gaseous Objects to Stars

5.1 Star Formation

As mentioned in the preface, the fragmentation of the first gaseous objects is a well-posed physics problem with well specified initial conditions, for a given power-spectrum of primordial density fluctuations. This problem is ideally suited for three-dimensional computer simulations, since it cannot be reliably addressed in idealized 1D or 2D geometries.

Recently, two groups have attempted detailed 3D simulations of the formation process of the first stars in a halo of $\sim 10^6$ $\mathrm{M}_\odot$ by following the dynamics of both the dark matter and the gas components, including H_2 chemistry and cooling. Bromm et al. (1999) [57] have used a Smooth Particle Hydrodynamics (SPH) code to simulate the collapse of a top-hat overdensity with a prescribed solid-body rotation (corresponding to a spin parameter $\lambda = 5\%$) and additional small perturbations with $P(k) \propto k^{-3}$ added to the top-hat profile. Abel et al. (2002) [5] isolated a high-density filament out of a larger simulated cosmological volume and followed the evolution of its density maximum with exceedingly high resolution using an Adaptive Mesh Refinement (AMR) algorithm.

The generic results of Bromm et al. (1999 [57]; see also Bromm 2000 [58]) are illustrated in Fig. 19. The collapsing region forms a disk which fragments into many clumps. The clumps have a typical mass $\sim 10^2$–10^3 $\mathrm{M}_\odot$. This mass scale corresponds to the Jeans mass for a temperature of $\sim 500\,\mathrm{K}$ and the density $\sim 10^4\,\mathrm{cm}^{-3}$ where the gas lingers because its cooling time is longer than its collapse time at that point (see Fig. 20). Each clump accretes mass slowly until it exceeds the Jeans mass and collapses at a roughly constant temperature (isothermally) due to H_2 cooling that brings the gas to a fixed temperature floor. The clump formation efficiency is high in this simulation due to the synchronized collapse of the overall top-hat perturbation.

Bromm (2000) [58] has simulated the collapse of one of the above-mentioned clumps with ~ 1000 $\mathrm{M}_\odot$ and demonstrated that it does not tend to fragment into sub-components. Rather, the clump core of ~ 100 $\mathrm{M}_\odot$ free-falls towards the center leaving an extended envelope behind with a roughly isothermal density profile. At very high gas densities, three-body reactions become important in the chemistry of H_2. Omukai & Nishi (1998) [273] have included these reactions as well as radiative transfer and followed the collapse in spherical symmetry up to stellar densities. Radiation pressure from nuclear

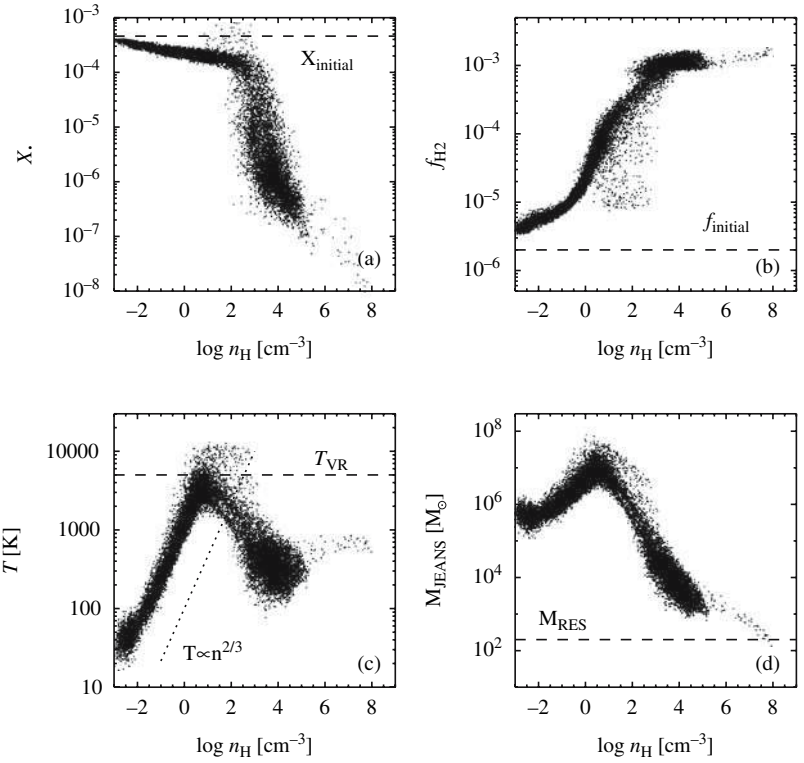

Fig. 19. Numerical results from Bromm et al. (1999) [57], showing gas properties at $z = 31.2$ for a collapsing slightly inhomogeneous top-hat region with a prescribed solid-body rotation. (**a**) Free electron fraction (by number) vs. hydrogen number density (in cm^{-3}). At densities exceeding $n \sim 10^3\,\mathrm{cm}^{-3}$, recombination is very efficient, and the gas becomes almost completely neutral. (**b**) Molecular hydrogen fraction vs. number density. After a quick initial rise, the H_2 fraction approaches the asymptotic value of $f \sim 10^{-3}$, due to the H^- channel. (**c**) Gas temperature vs. number density. At densities below $\sim 1\,\mathrm{cm}^{-3}$, the gas temperature rises because of adiabatic compression until it reaches the virial value of $T_{vir} \simeq 5000\,\mathrm{K}$. At higher densities, cooling due to H_2 drives the temperature down again, until the gas settles into a quasi-hydrostatic state at $T \sim 500\,\mathrm{K}$ and $n \sim 10^4\,\mathrm{cm}^{-3}$. Upon further compression due to accretion and the onset of gravitational collapse, the gas shows a further modest rise in temperature. (**d**) Jeans mass (in $\mathrm{M}_\odot$) vs. number density. The Jeans mass reaches a value of $M_\mathrm{J} \sim 10^3\ \mathrm{M}_\odot$ for the quasi-hydrostatic gas in the center of the potential well, and reaches the resolution limit of the simulation, $M_\mathrm{res} \simeq 200\ \mathrm{M}_\odot$, for densities close to $n = 10^8\,\mathrm{cm}^{-3}$

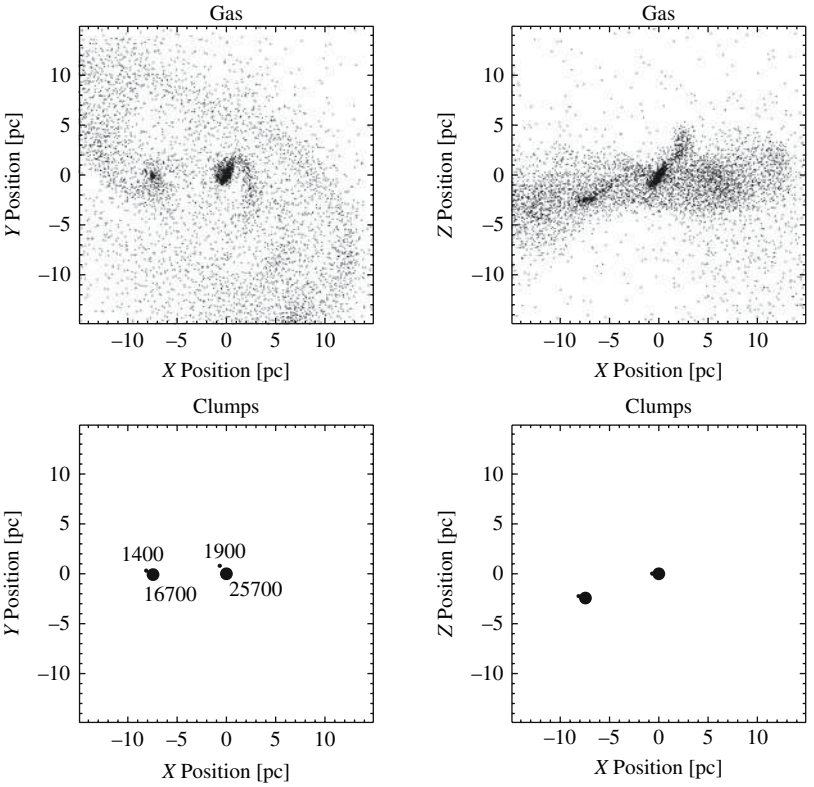

Fig. 20. Gas and clump morphology at $z = 28.9$ in the simulation of Bromm et al. (1999) [57]. *Top row:* The remaining gas in the diffuse phase. *Bottom row:* Distribution of clumps. The numbers next to the dots denote clump mass in units of $M_\odot$. **Left panels:** Face-on view. **Right panels:** Edge-on view. The length of the box is 30 pc. The gas has settled into a flattened configuration with two dominant clumps of mass close to $20,000\ M_\odot$. During the subsequent evolution, the clumps survive without merging, and grow in mass only slightly by accretion of surrounding gas

burning at the center is unlikely to reverse the infall as the stellar mass builds up. These calculations indicate that each clump may end as a single massive star; however, it is conceivable that angular momentum may eventually halt the collapsing cloud and lead to the formation of a binary stellar system instead.

The Jeans mass, which is defined based on small fluctuations in a background of *uniform* density, does not strictly apply in the context of collapsing gas cores. We can instead use a slightly modified critical mass known as the Bonnor-Ebert mass [53, 114]. For baryons in a background of uniform density ρ_b, perturbations are unstable to gravitational collapse in a region more massive than the Jeans mass. Instead of a uniform background, we con-

sider a spherical, non-singular, isothermal, self-gravitating gas in hydrostatic equilibrium, i.e., a centrally-concentrated object which more closely resembles the gas cores found in the above-mentioned simulations. In this case, small fluctuations are unstable and lead to collapse if the sphere is more massive than the Bonnor-Ebert mass $M_{\rm BE}$, given by the same expression the Jeans Mass but with a different coefficient (1.2 instead of 2.9) and with $\rho_{\rm b}$ denoting in this case the gas (volume) density at the surface of the sphere,

$$M_{\rm BE} = 1.2 \frac{1}{\sqrt{\rho_{\rm b}}} \left(\frac{kT}{G\mu m_{\rm p}} \right)^{3/2} . \tag{104}$$

In their simulation, Abel et al. (2000) [4] adopted the actual cosmological density perturbations as initial conditions. The simulation focused on the density peak of a filament within the IGM, and evolved it to very high densities (Fig. 21). Following the initial collapse of the filament, a clump core formed with $\sim$200 $M_\odot$, amounting to only $\sim$1% of the virialized mass. Subsequently due to slow cooling, the clump collapsed subsonically in a state close to hydrostatic equilibrium (see Fig. 22). Unlike the idealized top-hat simulation of Bromm et al. (2001) [59], the collapse of the different clumps within the

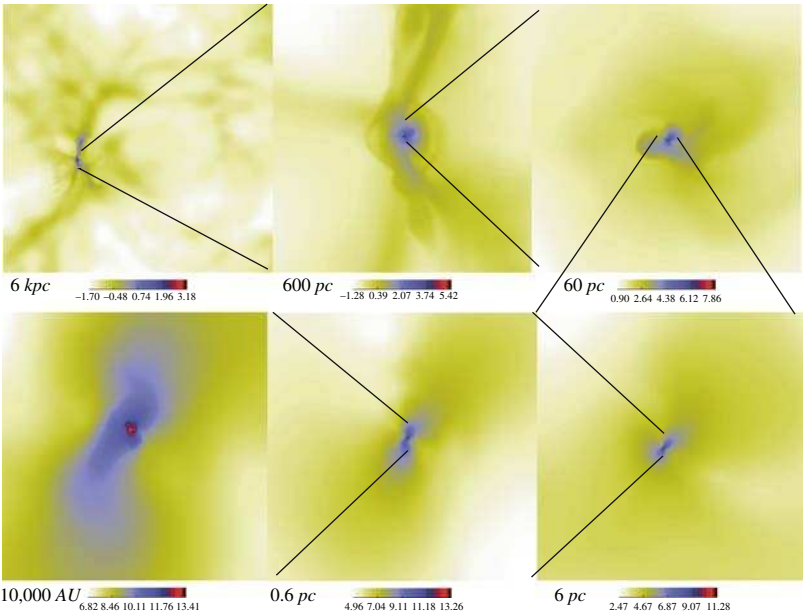

Fig. 21. Zooming in on the core of a star forming region with the *Adaptive Mesh Refinement* simulation of Abel et al. (2000) [4]. The panels show different length scales, decreasing clockwise by an order of magnitude between adjacent panels. Note the large dynamic range of scales which are being resolved, from 6 kpc (**top left panel**) down to 10,000 AU (**bottom left panel**, from Barkana & Loeb 2001 [23])

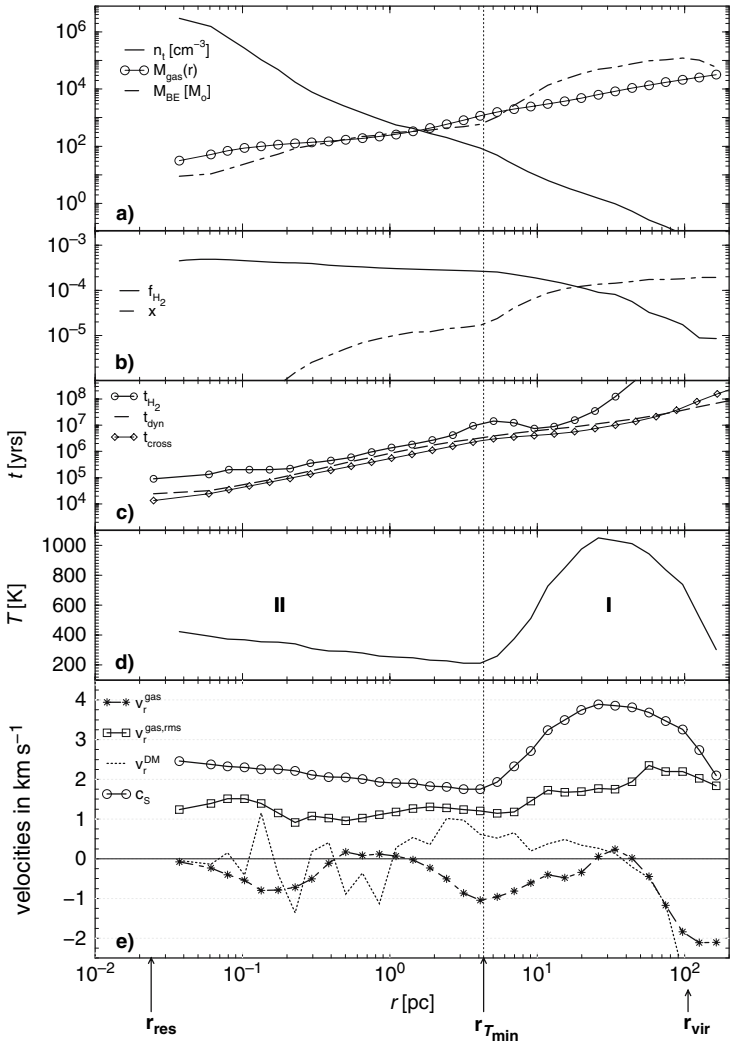

Fig. 22. Gas profiles from the simulation of Abel et al. (2000) [4]. The cell size on the finest grid corresponds to 0.024 pc, while the simulation box size corresponds to 6.4 kpc. Shown are spherically-averaged mass-weighted profiles around the baryon density peak shortly before a well defined fragment forms ($z = 19.1$). Panel (**a**) shows the baryonic number density, enclosed gas mass in solar mass, and the local Bonnor-Ebert mass $M_{\rm BE}$ (see text). Panel (**b**) plots the molecular hydrogen fraction (by number) $f_{\rm H_2}$ and the free electron fraction x. The H_2 cooling time, $t_{\rm H_2}$, the time it takes a sound wave to travel to the center, $t_{\rm cross}$, and the free–fall time $t_{\rm ff} = [3\pi/(32G\rho)]^{1/2}$ are given in panel (**c**). Panel (**d**) gives the temperature in K as a function of radius. The bottom panel gives the local sound speed, $c_{\rm s}$ (*solid line with circles*), the rms radial velocities of the dark matter (*dashed line*) and the gas (*dashed line with asterisks*) as well as the rms gas velocity (*solid line with square symbols*). The vertical dotted line indicates the radius ($\sim$5 pc) at which the gas has reached its minimum temperature allowed by H_2 cooling. The virial radius of the 5.6×10^6 $M_\odot$ halo is 106 pc

filament is not synchronized. Once the first star forms at the center of the first collapsing clump, it is likely to affect the formation of other stars in its vicinity.

As soon as nuclear burning sets in the core of the proto-star, the radiation emitted by the star starts to affect the infall of the surrounding gas towards it. The radiative feedback involves photo-dissociation of H_2, Lyα radiation pressure, and photo-evaporation of the accretion disk. Tan & McKee [356] studied these effects by extrapolating analytically the infall of gas from the final snapshot of the above resolution-limited simulations to the scale of a proto-star; they concluded that nuclear burning (and hence the feedback) starts when the proton-star accretes ~ 30 $M_\odot$ and accretion is likely to be terminated when the star reaches ~ 200 $M_\odot$.

If the clumps in the above simulations end up forming individual very massive stars, then these stars will likely radiate copious amounts of ionizing radiation [50, 369, 59] and expel strong winds. Hence, the stars will have a large effect on their interstellar environment, and feedback is likely to control the overall star formation efficiency. This efficiency is likely to be small in galactic potential wells which have a virial temperature lower than the temperature of photoionized gas, $\sim 10^4$ K. In such potential wells, the gas may go through only a single generation of star formation, leading to a "suicidal" population of massive stars.

The final state in the evolution of these stars is uncertain; but if their mass loss is not too extensive, then they are likely to end up as black holes [50, 137]. The remnants may provide the seeds of quasar black holes [213]. Some of the massive stars may end their lives by producing gamma-ray bursts. If so then the broad-band afterglows of these bursts could provide a powerful tool for probing the epoch of reionization [212, 94]. There is no better way to end the dark ages than with γ-ray burst fireworks.

Where are the first stars or their remnants located today? The very first stars formed in rare high-σ peaks and hence are likely to populate the cores of present-day galaxies [379]. However, the bulk of the stars which formed in low-mass systems at later times are expected to behave similarly to the collisionless dark matter particles and populate galaxy halos [220].

5.2 The Mass Function of Stars

Currently, we do not have direct observational constraints on how the first stars, the so-called Population III stars, formed at the end of the cosmic dark ages. It is, therefore, instructive to briefly summarize what we have learned about star formation in the present-day Universe, where theoretical reasoning is guided by a wealth of observational data (see [292] for a recent review).

Population I stars form out of cold, dense molecular gas that is structured in a complex, highly inhomogeneous way. The molecular clouds are supported against gravity by turbulent velocity fields and pervaded on large scales by magnetic fields. Stars tend to form in clusters, ranging from a few hundred

up to $\sim 10^6$ stars. It appears likely that the clustered nature of star formation leads to complicated dynamics and tidal interactions that transport angular momentum, thus allowing the collapsing gas to overcome the classical centrifugal barrier [214]. The initial mass function (IMF) of Pop I stars is observed to have the approximate Salpeter form (e.g., [206])

$$\frac{\mathrm{d}N}{\mathrm{d}\log M} \propto M^x \ , \tag{105}$$

where

$$x \simeq \begin{cases} -1.35 \text{ for } M \geqslant 0.5 \text{ M}_\odot \\ 0.0 \text{ for } 0.007 \leqslant M \leqslant 0.5 \text{ M}_\odot \end{cases} . \tag{106}$$

The lower cutoff in mass corresponds roughly to the opacity limit for fragmentation. This limit reflects the minimum fragment mass, set when the rate at which gravitational energy is released during the collapse exceeds the rate at which the gas can cool (e.g., [297]). The most important feature of the observed IMF is that ~ 1 M$_\odot$ is the characteristic mass scale of Pop I star formation, in the sense that most of the mass goes into stars with masses close to this value. In Fig. 23, we show the result from a recent hydrodynamical simulation of the collapse and fragmentation of a molecular cloud core [31, 32].

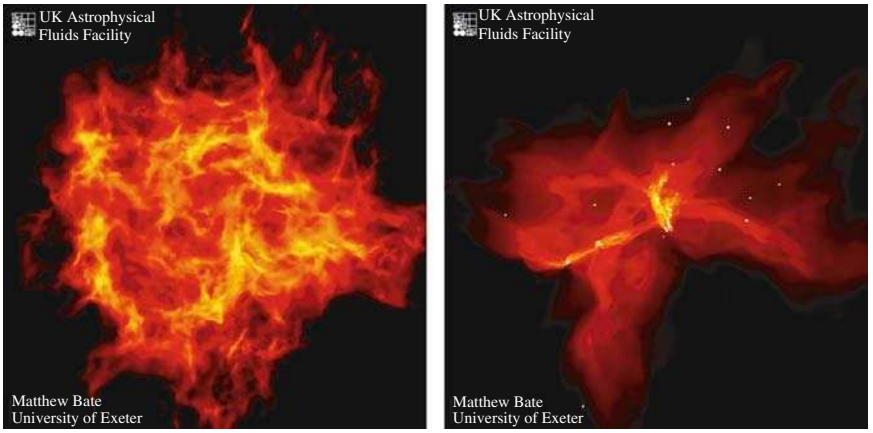

Fig. 23. A hydrodynamic simulation of the collapse and fragmentation of a turbulent molecular cloud in the present-day Universe (from Bate et al. 2003 [32]). The cloud has a mass of 50 M$_\odot$. The panels show the column density through the cloud, and span a scale of 0.4 pc across. **Left:** The initial phase of the collapse. The turbulence organizes the gas into a network of filaments, and decays thereafter through shocks. **Right:** A snapshot taken near the end of the simulation, after 1.4 initial free-fall times of 2×10^5 yr. Fragmentation has resulted in ~ 50 stars and brown dwarfs. The star formation efficiency is $\sim 10\%$ on the scale of the overall cloud, but can be much larger in the dense sub-condensations. This result is in good agreement with what is observed in local star-forming regions

This simulation illustrates the highly dynamic and chaotic nature of the star formation process[6].

The metal-rich chemistry, magnetohydrodynamics, and radiative transfer involved in present-day star formation is complex, and we still lack a comprehensive theoretical framework that predicts the IMF from first principles. Star formation in the high redshift Universe, on the other hand, poses a theoretically more tractable problem due to a number of simplifying features, such as: (i) the initial absence of heavy metals and therefore of dust; and (ii) the absence of dynamically-significant magnetic fields, in the pristine gas left over from the big bang. The cooling of the primordial gas does then only depend on hydrogen in its atomic and molecular form. Whereas in the present-day interstellar medium, the initial state of the star forming cloud is poorly constrained, the corresponding initial conditions for primordial star formation are simple, given by the popular ΛCDM model of cosmological structure formation. We now turn to a discussion of this theoretically attractive and important problem.

How did the first stars form? A complete answer to this question would entail a theoretical prediction for the Population III IMF, which is rather challenging. Let us start by addressing the simpler problem of estimating the characteristic mass scale of the first stars. As mentioned before, this mass scale is observed to be ~ 1 $M_\odot$ in the present-day Universe.

Bromm & Loeb (2004) [67] carried out idealized simulations of the protostellar accretion problem and estimated the final mass of a Population III star. Using the smoothed particle hydrodynamics (SPH) method, they included the chemistry and cooling physics relevant for the evolution of metal-free gas (see [62] for details). Improving on earlier work [57, 62] by initializing the simulations according to the ΛCDM model, they focused on an isolated overdense region that corresponds to a $3\sigma-$peak [67]: a halo containing a total mass of 10^6 $M_\odot$, and collapsing at a redshift $z_{\rm vir} \simeq 20$. In these runs, one high-density clump has formed at the center of the minihalo, possessing a gas mass of a few hundred solar masses. Soon after its formation, the clump becomes gravitationally unstable and undergoes runaway collapse. Once the gas clump has exceeded a threshold density of $10^7\,{\rm cm}^{-3}$, it is replaced by a sink particle which is a collisionless point-like particle that is inserted into the simulation. This choice for the density threshold ensures that the local Jeans mass is resolved throughout the simulation. The clump (i.e., sink particle) has an initial mass of $M_{\rm Cl} \simeq 200$ $M_\odot$, and grows subsequently by ongoing accretion of surrounding gas. High-density clumps with such masses result from the chemistry and cooling rate of molecular hydrogen, H_2, which imprint characteristic values of temperature, $T \sim 200\,{\rm K}$, and density, $n \sim 10^4\,{\rm cm}^{-3}$, into the metal-free gas [62]. Evaluating the Jeans mass for these characteristic values results in $M_{\rm J} \sim$ a few $\times\, 10^2$ $M_\odot$, which is close to the initial clump masses found in the simulations.

[6] See http:// www.ukaff.ac.uk/starcluster for an animation.

The high-density clumps are clearly not stars yet. To probe the subsequent fate of a clump, Bromm & Loeb (2004) [67] have re-simulated the evolution of the central clump with sufficient resolution to follow the collapse to higher densities. Figure 24 (*right panel*) shows the gas density on a scale of 0.5 pc, which is two orders of magnitude smaller than before. Several features are evident in this plot. First, the central clump does not undergo further subfragmentation, and is likely to form a single Population III star. Second, a companion clump is visible at a distance of ~ 0.25 pc. If negative feedback from the first-forming star is ignored, this companion clump would undergo runaway collapse on its own approximately ~ 3 Myr later. This timescale is comparable to the lifetime of a very massive star (VMS)[59]. If the second clump was able to survive the intense radiative heating from its neighbor, it could become a star before the first one explodes as a supernova (SN). Whether more than one star can form in a low-mass halo thus crucially depends on the degree of synchronization of clump formation. Finally, the non-axisymmetric disturbance induced by the companion clump, as well as the angular momentum stored in the orbital motion of the binary system, allow the system to overcome the angular momentum barrier for the collapse of the central clump (see [214]).

The recent discovery of stars like HE0107-5240 with a mass of 0.8 $M_\odot$ and an iron abundance of [Fe/H] $= -5.3$ [90] shows that at least some low mass stars could have formed out of extremely low-metallicity gas. The above

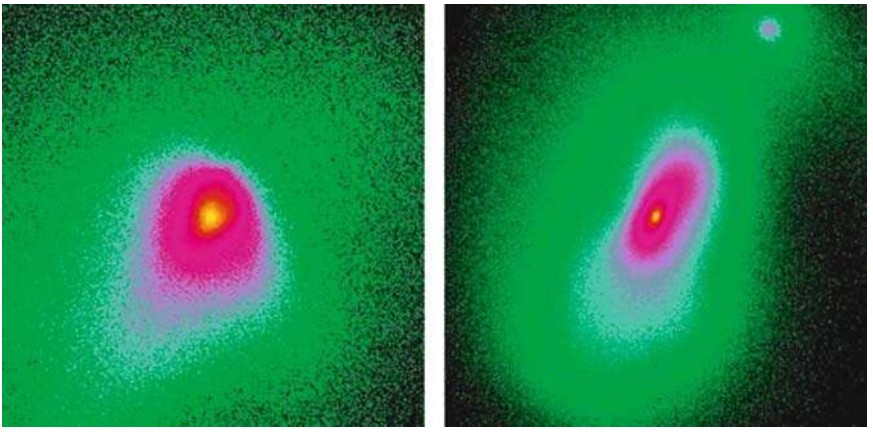

Fig. 24. Collapse and fragmentation of a primordial cloud (from Bromm & Loeb 2004 [67]). Shown is the projected gas density at a redshift $z \simeq 21.5$, briefly after gravitational runaway collapse has commenced in the center of the cloud. **Left:** The coarse-grained morphology in a box with linear physical size of 23.5 pc. At this time in the unrefined simulation, a high-density clump (sink particle) has formed with an initial mass of $\sim 10^3$ $M_\odot$. **Right:** The refined morphology in a box with linear physical size of 0.5 pc. The central density peak, vigorously gaining mass by accretion, is accompanied by a secondary clump

simulations show that although the majority of clumps are very massive, a few of them, like the secondary clump in Fig. 24, are significantly less massive. Alternatively, low-mass fragments could form in the dense, shock-compressed shells that surround the first hypernovae [233].

How massive were the first stars? Star formation typically proceeds from the "inside-out", through the accretion of gas onto a central hydrostatic core. Whereas the initial mass of the hydrostatic core is very similar for primordial and present-day star formation [273], the accretion process – ultimately responsible for setting the final stellar mass, is expected to be rather different. On dimensional grounds, the accretion rate is simply related to the sound speed cubed over Newton's constant (or equivalently given by the ratio of the Jeans mass and the free-fall time): $\dot{M}_{\rm acc} \sim c_{\rm s}^3/G \propto T^{3/2}$. A simple comparison of the temperatures in present-day star forming regions ($T \sim 10\,{\rm K}$) with those in primordial ones ($T \sim 200 - 300\,{\rm K}$) already indicates a difference in the accretion rate of more than two orders of magnitude.

The above refined simulation enables one to study the three-dimensional accretion flow around the protostar (see also [275, 303, 356]). The gas may now reach densities of $10^{12}\,{\rm cm}^{-3}$ before being incorporated into a central sink particle. At these high densities, three-body reactions [279] convert the gas into a fully molecular form. Figure 25 shows how the molecular core grows in mass over the first $\sim 10^4\,{\rm yr}$ after its formation. The accretion rate (*left panel*) is initially very high, $\dot{M}_{\rm acc} \sim 0.1\ {\rm M}_\odot\,{\rm yr}^{-1}$, and subsequently declines according to a power law, with a possible break at $\sim 5000\,{\rm yr}$. The mass of the molecular core (*right panel*), taken as an estimator of the proto-stellar mass,

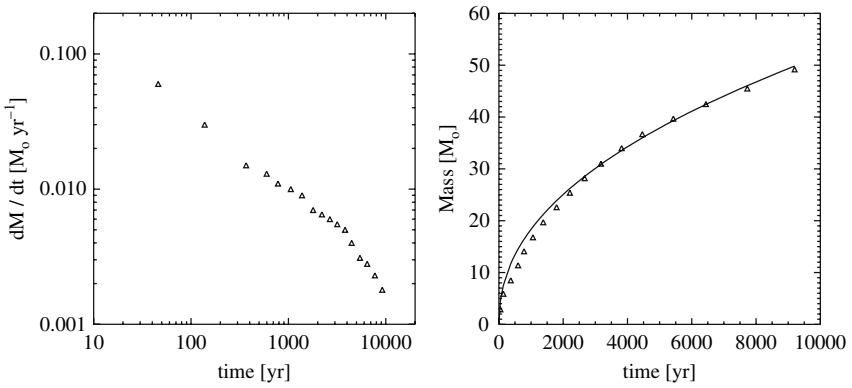

Fig. 25. Accretion onto a primordial protostar (from Bromm & Loeb 2004 [67]). The morphology of this accretion flow is shown in Fig. 24. **Left:** Accretion rate (in ${\rm M}_\odot\,{\rm yr}^{-1}$) vs. time (in yr) since molecular core formation. **Right:** Mass of the central core (in ${\rm M}_\odot$) vs. time. *Solid line:* Accretion history approximated as: $M_* \propto t^{0.45}$. Using this analytical approximation, we extrapolate that the protostellar mass has grown to $\sim 150\ {\rm M}_\odot$ after $\sim 10^5\,{\rm yr}$, and to $\sim 700\ {\rm M}_\odot$ after $\sim 3 \times 10^6\,{\rm yr}$, the total lifetime of a very massive star

grows approximately as: $M_* \sim \int \dot{M}_{\rm acc} {\rm d}t \propto t^{0.45}$. A rough upper limit for the final mass of the star is then: $M_*(t = 3 \times 10^6\,{\rm yr}) \sim 700\ {\rm M}_\odot$. In deriving this upper bound, we have conservatively assumed that accretion cannot go on for longer than the total lifetime of a massive star.

Can a Population III star ever reach this asymptotic mass limit? The answer to this question is not yet known with any certainty, and it depends on whether the accretion from a dust-free envelope is eventually terminated by feedback from the star (e.g., [275, 303, 356, 276]). The standard mechanism by which accretion may be terminated in metal-rich gas, namely radiation pressure on dust grains [385], is evidently not effective for gas with a primordial composition. Recently, it has been speculated that accretion could instead be turned off through the formation of an H II region [276], or through the photo-evaporation of the accretion disk [356]. The termination of the accretion process defines the current unsolved frontier in studies of Population III star formation. Current simulations indicate that the first stars were predominantly very massive ($\gtrsim 30\ {\rm M}_\odot$), and consequently rather different from present-day stellar populations. The crucial question then arises: *How and when did the transition take place from the early formation of massive stars to that of low-mass stars at later times?* We address this problem next.

The very first stars, marking the cosmic Renaissance of structure formation, formed under conditions that were much simpler than the highly complex environment in present-day molecular clouds. Subsequently, however, the situation rapidly became more complicated again due to the feedback from the first stars on the IGM. Supernova explosions dispersed the nucleosynthetic products from the first generation of stars into the surrounding gas (e.g., [240, 260, 360]), including also dust grains produced in the explosion itself [221, 363]. Atomic and molecular cooling became much more efficient after the addition of these metals. Moreover, the presence of ionizing cosmic rays, as well as of UV and X-ray background photons, modified the thermal and chemical behavior of the gas in important ways (e.g., [231, 232]).

Early metal enrichment was likely the dominant effect that brought about the transition from Population III to Population II star formation. Recent numerical simulations of collapsing primordial objects with overall masses of $\sim 10^6\ {\rm M}_\odot$, have shown that the gas has to be enriched with heavy elements to a minimum level of $Z_{\rm crit} \simeq 10^{-3.5}\ {\rm Z}_\odot$, in order to have any effect on the dynamics and fragmentation properties of the system [274, 60, 64]. Normal, low-mass (Population II) stars are hypothesized to only form out of gas with metallicity $Z \geqslant Z_{\rm crit}$. Thus, the characteristic mass scale for star formation is expected to be a function of metallicity, with a discontinuity at $Z_{\rm crit}$ where the mass scale changes by $\sim$ two orders of magnitude. The redshift where this transition occurs has important implications for the early growth of cosmic structure, and the resulting observational signature (e.g., [391, 141, 233, 321]) include the extended nature of reionization [144].

For additional detailes about the properties of the first stars, see the comprehensive review by Bromm & Larson (2004) [66].

5.3 Gamma-ray Bursts: Probing the First Stars One Star at a Time

Gamma-Ray Bursts (GRBs) are believed to originate in compact remnants (neutron stars or black holes) of massive stars. Their high luminosities make them detectable out to the edge of the visible Universe [94, 212]. GRBs offer the opportunity to detect the most distant (and hence earliest) population of massive stars, the so-called Population III (or Pop III), one star at a time. In the hierarchical assembly process of halos which are dominated by cold dark matter (CDM), the first galaxies should have had lower masses (and lower stellar luminosities) than their low-redshift counterparts. Consequently, the characteristic luminosity of galaxies or quasars is expected to decline with increasing redshift. GRB afterglows, which already produce a peak flux comparable to that of quasars or starburst galaxies at $z \sim 1$–2, are therefore expected to outshine any competing source at the highest redshifts, when the first dwarf galaxies have formed in the Universe.

The first-year polarization data from the *Wilkinson Microwave Anisotropy Probe* (*WMAP*) indicates an optical depth to electron scattering of $\sim 17 \pm 4\%$ after cosmological recombination [201, 347]. This implies that the first stars must have formed at a redshift $z \sim$10–20, and reionized a substantial fraction of the intergalactic hydrogen around that time [83, 93, 344, 393, 402]. Early reionization can be achieved with plausible star formation parameters in the standard ΛCDM cosmology; in fact, the required optical depth can be achieved in a variety of very different ionization histories since *WMAP* places only an integral constraint on these histories [174]. One would like to probe the full history of reionization in order to disentangle the properties and formation history of the stars that are responsible for it. GRB afterglows offer the opportunity to detect stars as well as to probe the metal enrichment level [141] of the intervening IGM.

GRBs, the electromagnetically-brightest explosions in the Universe, should be detectable out to redshifts $z > 10$ [94, 212] (Fig. 26). High-redshift GRBs can be identified through infrared photometry, based on the Lyα break induced by absorption of their spectrum at wavelengths below $1.216\,\mu\mathrm{m}\,[(1+z)/10]$. Follow-up spectroscopy of high-redshift candidates can then be performed on a 10-m-class telescope. Recently, the ongoing *Swift* mission [147] has detected a GRB originating at $z \simeq 6.3$ (e.g., [177]), thus demonstrating the viability of GRBs as probes of the early Universe.

There are four main advantages of GRBs relative to traditional cosmic sources such as quasars:

(i) The GRB afterglow flux at a given observed time lag after the γ-ray trigger is not expected to fade significantly with increasing redshift, since higher redshifts translate to earlier times in the source frame, during which the afterglow is intrinsically brighter [94]. For standard afterglow lightcurves and spectra, the increase in the luminosity distance with redshift is compensated by this *cosmological time-stretching* effect.

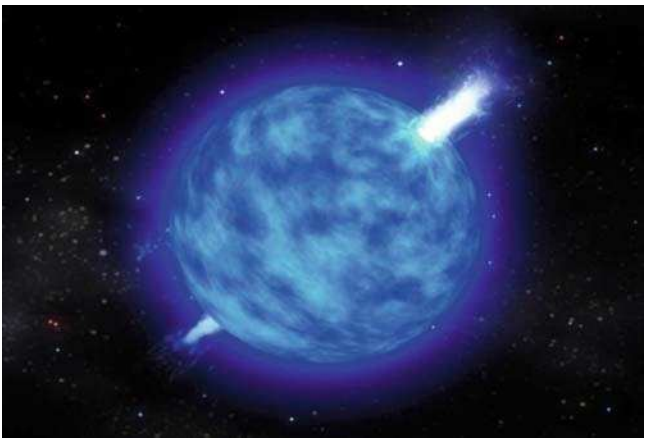

Fig. 26. Illustration of a long-duration gamma-ray burst in the popular "collapsar" model. (Image from http://imagine.gsfc.nasa.gov/docs/science/, credit: NASA/ Sky Works Digital) The collapse of the core of a massive star (which lost its hydrogen envelope) to a black hole generates two opposite jets moving out at a speed close to the speed of light. The jets drill a hole in the star and shine brightly towards an observer who happened to be located within with the collimation cones of the jets. The jets emenating from a single massive star are so bright that they can be seen across the Universe out to the epoch when the first stars have formed. Upcoming observations by the *Swift* satellite will have the sensitivity to reveal whether the first stars served as progenitors of gamma-ray bursts (for updates see http://swift.gsfc.nasa.gov/)

(ii) As already mentioned, in the standard ΛCDM cosmology, galaxies form hierarchically, starting from small masses and increasing their average mass with cosmic time. Hence, the characteristic mass of quasar black holes and the total stellar mass of a galaxy were smaller at higher redshifts, making these sources intrinsically fainter [390]. However, GRBs are believed to originate from a stellar mass progenitor and so the intrinsic luminosity of their engine should not depend on the mass of their host galaxy. GRB afterglows are therefore expected to outshine their host galaxies by a factor that gets larger with increasing redshift.

(iii) Since the progenitors of GRBs are believed to be stellar, they likely originate in the most common star-forming galaxies at a given redshift rather than in the most massive host galaxies, as is the case for bright quasars [26]. Low-mass host galaxies induce only a weak ionization effect on the surrounding IGM and do not greatly perturb the Hubble flow around them. Hence, the Lyα damping wing should be closer to the idealized unperturbed IGM case and its detailed spectral shape should be easier to interpret. Note also that unlike the case of a quasar, a GRB afterglow can itself ionize at most $\sim 4 \times 10^4 E_{51}$ $M_\odot$ of hydrogen if its UV energy is E_{51} in units of 10^{51} ergs (based on the available number of ionizing photons), and so it should have a negligible cosmic effect on the surrounding IGM.

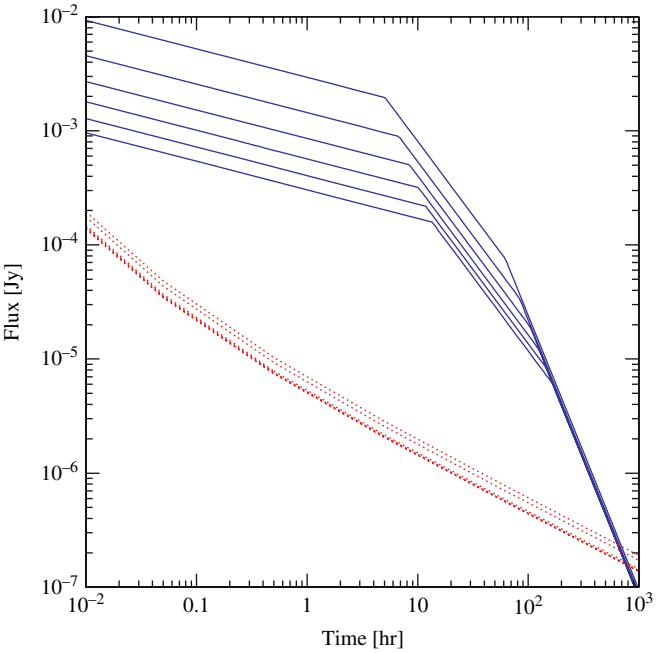

Fig. 27. GRB afterglow flux as a function of time since the γ-ray trigger in the observer frame (taken from Bromm & Loeb 2004 [67]). The flux (*solid curves*) is calculated at the redshifted Lyα wavelength. The dotted curves show the planned detection threshold for the *James Webb Space Telescope* (*JWST*), assuming a spectral resolution $R = 5000$ with the near infrared spectrometer, a signal to noise ratio of 5 per spectral resolution element, and an exposure time equal to 20% of the time since the GRB. Each set of curves shows a sequence of redshifts, namely $z = 5$, 7, 9, 11, 13, and 15, respectively, **from top to bottom**

(iv) GRB afterglows have smooth (broken power-law) continuum spectra unlike quasars which show strong spectral features (such as broad emission lines or the so-called "blue bump") that complicate the extraction of IGM absorption features (Fig. 27). In particular, the continuum extrapolation into the Lyα damping wing (the so-called Gunn-Peterson absorption trough) during the epoch of reionization is much more straightforward for the smooth UV spectra of GRB afterglows than for quasars with an underlying broad Lyα emission line [26].

The optical depth of the uniform IGM to Lyα absorption is given by (Gunn & Peterson 1965 [163]),

$$\tau_{\rm s} = \frac{\pi e^2 f_\alpha \lambda_\alpha n_{\rm HI}\,(z_{\rm s})}{m_{\rm e} c H(z_{\rm s})} \approx 6.45 \times 10^5 x_{\rm HI} \left(\frac{\Omega_{\rm b} h}{0.03}\right) \left(\frac{\Omega_{\rm m}}{0.3}\right)^{-1/2} \left(\frac{1+z_{\rm s}}{10}\right)^{3/2} \tag{107}$$

where $H \approx 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}\Omega_{\rm m}^{1/2}(1+z_{\rm s})^{3/2}$ is the Hubble parameter at the source redshift $z_{\rm s} >> 1$, $f_\alpha = 0.4162$ and $\lambda_\alpha = 1216\text{Å}$ are the oscillator strength and the wavelength of the Lyα transition; $n_{\rm HI}\,(z_{\rm s})$ is the neutral hydrogen density at the source redshift (assuming primordial abundances); $\Omega_{\rm m}$ and $\Omega_{\rm b}$ are the present-day density parameters of all matter and of baryons, respectively; and $x_{\rm HI}$ is the average fraction of neutral hydrogen. In the second equality we have implicitly considered high-redshifts, $(1+z) \gg \max\left[(1-\Omega_{\rm m}-\Omega_\Lambda)/\Omega_{\rm m}, (\Omega_\Lambda/\Omega_{\rm m})^{1/3}\right]$, at which the vacuum energy density is negligible relative to matter ($\Omega_\Lambda \ll \Omega_{\rm m}$) and the Universe is nearly flat; for $\Omega_{\rm m} = 0.3, \Omega_\Lambda = 0.7$ this corresponds to the condition $z \gg 1.3$ which is well satisfied by the reionization redshift.

At wavelengths longer than Lyα at the source, the optical depth obtains a small value; these photons redshift away from the line center along its red wing and never resonate with the line core on their way to the observer. The red damping wing of the Gunn-Peterson trough (Miralda-Escudé 1998 [253])

$$\tau(\lambda_{\rm obs}) = \tau_{\rm s}\left(\frac{\Lambda}{4\pi^2\nu_\alpha}\right)\tilde{\lambda}_{\rm obs}^{3/2}\left[I(\tilde{\lambda}_{\rm obs}^{-1}) - I([(1+z_{\rm i})/(1+z_{\rm s})]\tilde{\lambda}_{\rm obs}^{-1})\right] \text{ for } \tilde{\lambda}_{\rm obs} \geqslant 1\,, \tag{108}$$

where $\tau_{\rm s}$ is given in (107), also we define

$$\tilde{\lambda}_{\rm obs} \equiv \frac{\lambda_{\rm obs}}{(1+z_{\rm s})\lambda_\alpha} \tag{109}$$

and

$$I(x) \equiv \frac{x^{9/2}}{1-x} + \frac{9}{7}x^{7/2} + \frac{9}{5}x^{5/2} + 3x^{3/2} + 9x^{1/2} - \frac{9}{2}\ln\left[\frac{1+x^{1/2}}{1-x^{1/2}}\right]\,. \tag{110}$$

Although the nature of the central engine that powers the relativistic jets of GRBs is still unknown, recent evidence indicates that long-duration GRBs trace the formation of massive stars (e.g., [364, 382, 45, 209, 47, 263]) and in particular that long-duration GRBs are associated with Type Ib/c supernovae [350]. Since the first stars in the Universe are predicted to be predominantly massive [5, 62, 66], their death might give rise to large numbers of GRBs at high redshifts. In contrast to quasars of comparable brightness, GRB afterglows are short-lived and release $\sim$10 orders of magnitude less energy into the surrounding IGM. Beyond the scale of their host galaxy, they have a negligible effect on their cosmological environment[7]. Consequently, they are ideal probes of the IGM during the reionization epoch. Their rest-frame UV spectra can be used to probe the ionization state of the IGM through the spectral shape of the Gunn-Peterson (Lyα) absorption trough, or its metal enrichment history through the intersection of enriched bubbles of supernova (SN) ejecta from

[7] Note, however, that feedback from a single GRB or supernova on the gas confined within early dwarf galaxies could be dramatic, since the binding energy of most galaxies at $z > 10$ is lower than 10^{51} ergs [23].

early galaxies [141]. Afterglows that are unusually bright ($> 10\,\mathrm{mJy}$) at radio frequencies should also show a detectable forest of 21 cm absorption lines due to enhanced HI column densities in sheets, filaments, and collapsed minihalos within the IGM [76, 140].

Another advantage of GRB afterglows is that once they fade away, one may search for their host galaxies. Hence, GRBs may serve as signposts of the earliest dwarf galaxies that are otherwise too faint or rare on their own for a dedicated search to find them. Detection of metal absorption lines from the host galaxy in the afterglow spectrum, offers an unusual opportunity to study the physical conditions (temperature, metallicity, ionization state, and kinematics) in the interstellar medium of these high-redshift galaxies. As Fig. 28 indicates, damped Lyα absorption within the host galaxy may mask the clear signature of the Gunn-Peterson trough in some galaxies [67]. A small fraction ($\sim$10) of the GRB afterglows are expected to originate at redshifts $z > 5$ [61, 68]. This subset of afterglows can be selected photometrically using a small telescope, based on the Lyα break at a wavelength of $1.216\,\mu\mathrm{m}\,[(1+z)/10]$, caused by intergalactic HI absorption. The challenge in the upcoming years will be to follow-up on these candidates spectroscopically, using a large (10-m class) telescope. GRB afterglows are likely to revolutionize observational cosmology and replace traditional sources like quasars, as probes of the IGM at $z > 5$. The near future promises to be exciting for GRB astronomy as well as for studies of the high-redshift Universe.

It is of great importance to constrain the Pop III star formation mode, and in particular to determine down to which redshift it continues to be prominent. The extent of the Pop III star formation will affect models of the initial stages of reionization (e.g., [393, 93, 342, 402, 12]) and metal enrichment (e.g., [233, 141, 144, 319, 339]), and will determine whether planned surveys will be able to effectively probe Pop III stars (e.g., [318]). The constraints on Pop III star formation will also determine whether the first stars could have contributed a significant fraction to the cosmic near-IR background (e.g., [310, 309, 191, 241, 113]). To constrain high-redshift star formation from GRB observations, one has to address two major questions:

(1) *What is the signature of GRBs that originate in metal-free, Pop III progenitors?* Simply knowing that a given GRB came from a high redshift is not sufficient to reach a definite conclusion as to the nature of the progenitor. Pregalactic metal enrichment was likely inhomogeneous, and we expect normal Pop I and II stars to exist in galaxies that were already metal-enriched at these high redshifts [68]. Pop III and Pop I/II star formation is thus predicted to have occurred concurrently at $z > 5$. How is the predicted high mass-scale for Pop III stars reflected in the observational signature of the resulting GRBs? Preliminary results from numerical simulations of Pop III star formation indicate that circumburst densities are systematically higher in Pop III environments. GRB afterglows will then be much brighter than for conventional GRBs. In addition, due to

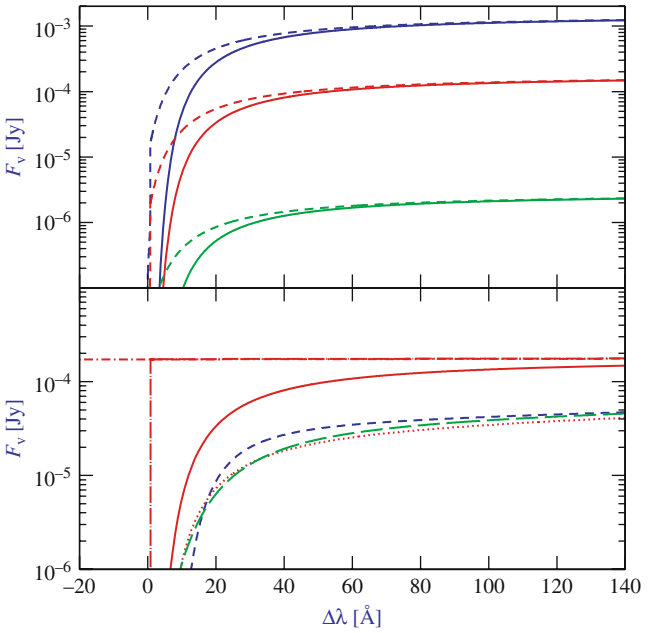

Fig. 28. Expected spectral shape of the Lyα absorption trough due to intergalactic absorption in GRB afterglows (taken from Bromm & Loeb 2004 [67]). The spectrum is presented in terms of the flux density F_ν versus relative observed wavelength $\Delta\lambda$, for a source redshift $z = 7$ (assumed to be prior to the final reionization phase) and the typical halo mass $M = 4 \times 10^8$ $M_\odot$ expected for GRB host galaxies that cool via atomic transitions. **Top panel:** Two examples for the predicted spectrum including IGM HI absorption (both resonant and damping wing), for host galaxies with (i) an age $t_S = 10^7$ yr, a UV escape fraction $f_{esc} = 10\%$ and a Scalo initial mass function (IMF) in solid curves, or (ii) $t_S = 10^8$ yr, $f_{esc} = 90\%$ and massive (> 100 $M_\odot$) Pop III stars in dashed curves. The observed time after the γ-ray trigger is one hour, one day, and ten days, from top to bottom, respectively. **Bottom panel:** Predicted spectra one day after a GRB for a host galaxy with $t_S = 10^7$ yr, $f_{esc} = 10\%$ and a Scalo IMF. Shown is the unabsorbed GRB afterglow (*dot-short dashed curve*), the afterglow with resonant IGM absorption only (*dot-long dashed curve*), and the afterglow with full (resonant and damping wing) IGM absorption (*solid curve*). Also shown, with 1.7 magnitudes of extinction, are the afterglow with full IGM absorption (*dotted curve*), and attempts to reproduce this profile with a damped Lyα absorption system in the host galaxy (*dashed curves*). (Note, however, that damped absorption of this type could be suppressed by the ionizing effect of the afterglow UV radiation on the surrounding interstellar medium of its host galaxy Perna & Loeb 1998 [288].) Most importantly, the overall spectral shape of the Lyα trough carries precious information about the neutral fraction of the IGM at the source redshift; averaging over an ensemble of sources with similar redshifts can reduce ambiguities in the interpretation of each case due to particular local effects

the systematically increased progenitor masses, the Pop III distribution may be biased toward long-duration events.

(2) The modelling of Pop III cosmic star formation histories has a number of free parameters, such as the star formation efficiency and the strength of the chemical feedback. The latter refers to the timescale for, and spatial extent of, the distribution of the first heavy elements that were produced inside of Pop III stars, and subsequently dispersed into the IGM by supernova blast waves. Comparing with theoretical GRB redshift distributions one can use the GRB redshift distribution observed by *Swift* to calibrate the free model parameters. In particular, one can use this strategy to measure the redshift where Pop III star formation terminates.

Figures 29 and 30 illustrate these issues (based on [68]). Figure 30 leads to the robust expectation that $\sim 10\%$ of all *Swift* bursts should originate at $z > 5$. This prediction is based on the contribution from Population I/II stars which are known to exist even at these high redshifts. Additional GRBs could be triggered by Pop III stars, with a highly uncertain efficiency. Assuming that long-duration GRBs are produced by the collapsar mechanism, a Pop III

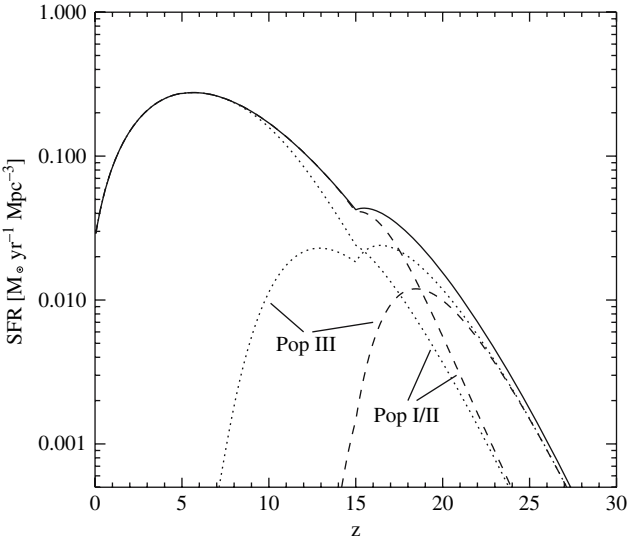

Fig. 29. Theoretical prediction for the comoving star formation rate (SFR) in units of $M_\odot\,yr^{-1}\,Mpc^{-3}$, as a function of redshift (from Bromm & Loeb 2006a [68]). We assume that cooling in primordial gas is due to atomic hydrogen only, a star formation efficiency of $\eta_* = 10\%$, and reionization beginning at $z_{\rm reion} \approx 17$. *Solid line*: Total comoving SFR. *Dotted lines*: Contribution to the total SFR from Pop I/II and Pop III for the case of weak chemical feedback. *Dashed lines*: Contribution to the total SFR from Pop I/II and Pop III for the case of strong chemical feedback. Pop III star formation is restricted to high redshifts, but extends over a significant range, $\Delta z \sim 10$–15

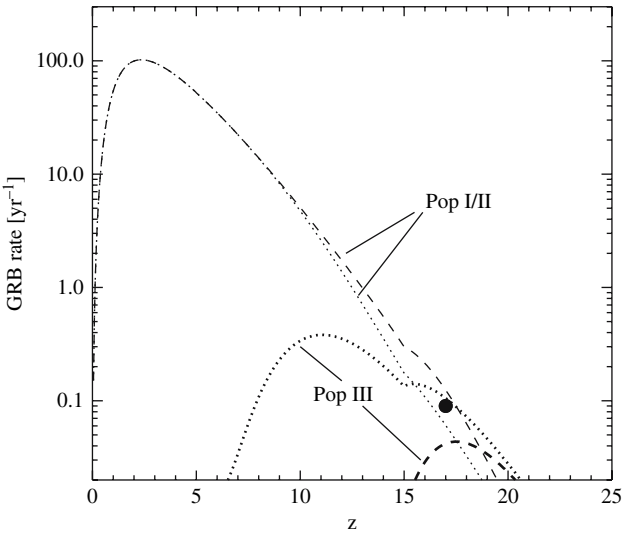

Fig. 30. Predicted GRB rate to be observed by *Swift* (from Bromm & Loeb 2006a [68]). Shown is the observed number of bursts per year, $\mathrm{d}N_{\mathrm{GRB}}^{\mathrm{obs}}/\mathrm{d}\ln(1+z)$, as a function of redshift. All rates are calculated with a constant GRB efficiency, $\eta_{\mathrm{GRB}} \simeq 2\times10^{-9}$ bursts $\mathrm{M}_{\odot}^{-1}$, using the cosmic SFRs from Fig. 29. *Dotted lines*: Contribution to the observed GRB rate from Pop I/II and Pop III for the case of weak chemical feedback. *Dashed lines*: Contribution to the GRB rate from Pop I/II and Pop III for the case of strong chemical feedback. *Filled circle*: GRB rate from Pop III stars if these were responsible for reionizing the Universe at $z \sim 17$

star with a close binary companion provides a plausible GRB progenitor. The Pop III GRB efficiency, reflecting the probability of forming sufficiently close and massive binary systems, lies between zero (if tight Pop III binaries do not exist) and $\sim$10 times the empirically inferred value for Population I/II (due to the increased fraction of black hole forming progenitors among the massive Pop III stars).

A key ingredient in determining the underlying star formation history from the observed GRB redshift distribution is the GRB luminosity function, which is only poorly constrained at present. The improved statistics provided by *Swift* will enable the construction of an empirical luminosity function. With an improved luminosity function it would be possible to re-calibrate the theoretical prediction in Fig. 29 more reliably.

In order to predict the observational signature of high-redshift GRBs, it is important to know the properties of the GRB host systems. Within variants of the popular CDM model for structure formation, where small objects form first and subsequently merge to build up more massive ones, the first stars are predicted to form at $z \sim$ 20–30 in minihalos of total mass (dark matter plus gas) $\sim 10^6$ $\mathrm{M}_{\odot}$ [358, 23, 402]. These objects are the sites for the formation of the first stars, and thus are the potential hosts of the highest-redshift GRBs.

What is the environment in which the earliest GRBs and their afterglows did occur? This problem breaks down into two related questions: (i) what type of stars (in terms of mass, metallicity, and clustering properties) will form in each minihalo?, and (ii) how will the ionizing radiation from each star modify the density structure of the surrounding gas? These two questions are fundamentally intertwined. The ionizing photon production strongly depends on the stellar mass, which in turn is determined by how the accretion flow onto the growing protostar proceeds under the influence of this radiation field. In other words, the assembly of the Population III stars and the development of an H II region around them proceed simultaneously, and affect each other. As a preliminary illustration, Fig. 31 describes the photo-evaporation as a self-similar champagne flow [336] with parameters appropriate for the Pop III case.

Notice that the central density is significantly reduced by the end of the life of a massive star, and that a central core has developed where the density is nearly constant. Such a flat density profile is markedly different from that created by stellar winds ($\rho \propto r^{-2}$). Winds, and consequently mass-loss, may not be important for massive Population III stars [18, 208], and such a flat

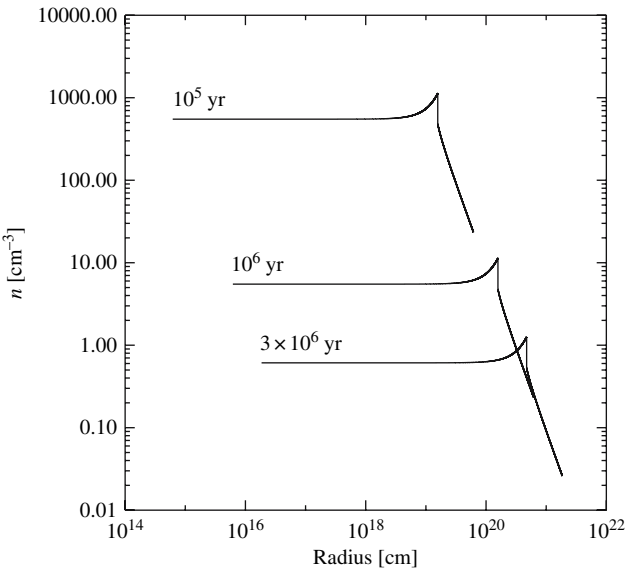

Fig. 31. Effect of photoheating from a Population III star on the density profile in a high-redshift minihalo (from Bromm & Loeb 2006b [69]). The curves, labeled by the time after the onset of the central point source, are calculated according to a self-similar model for the expansion of an H II region. Numerical simulations closely conform to this analytical behavior. Notice that the central density is significantly reduced by the end of the life of a massive star, and that a central core has developed where the density is constant

density profile may be characteristic of GRBs that originate from metal-free Population III progenitors.

The first galaxies may be surrounded by a shell of highly enriched material that was carried out in a SN-driven wind (see Fig. 32). A GRB in that galaxy may show strong absorption lines at a velocity separation associated with the wind velocity. Simulating these winds and calculating the absorption profile in the featureless spectrum of a GRB afterglow, will allow us to use the observed spectra of high-z GRBs and directly probe the degree of metal enrichment in the vicinity of the first star forming regions (see [141] for a semi-analytic treatment).

As the early afterglow radiation propagates through the interstellar environment of the GRB, it will likely modify the gas properties close to the source; these changes could in turn be noticed as time-dependent spectral features in the spectrum of the afterglow and used to derive the properties of the gas cloud (density, metal abundance, and size). The UV afterglow radiation can induce detectable changes to the interstellar absorption features of the host galaxy [288]; dust destruction could have occurred due to the GRB X-rays [374, 136], and molecules could have been destroyed near the GRB source [112]. Quantitatively, all of the effects mentioned above strongly depend on the exact properties of the gas in the host system.

Most studies to date have assumed a constant efficiency of forming GRBs per unit mass of stars. This simplifying assumption could lead, under different

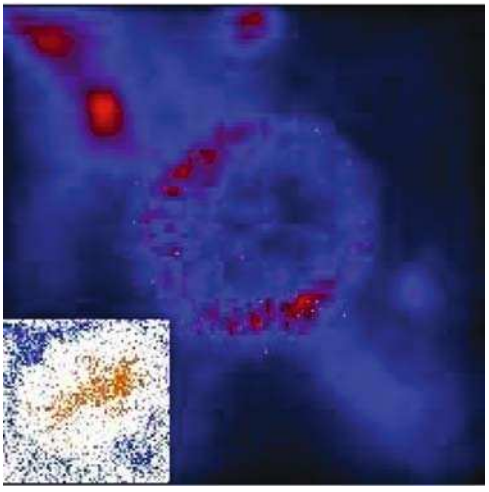

Fig. 32. Supernova explosion in the high-redshift Universe (from Bromm et al. 2003 [65]). The snapshot is taken $\sim 10^6$ yr after the explosion with total energy $E_{\rm SN} \simeq 10^{53}$ ergs. We show the projected gas density within a box of linear size 1 kpc. The SN bubble has expanded to a radius of $\sim$200 pc, having evacuated most of the gas in the minihalo. *Inset:* Distribution of metals. The stellar ejecta (*gray dots*) trace the metals and are embedded in pristine metal-poor gas (*black dots*)

circumstances, to an overestimation or an underestimation of the frequency of GRBs. Metal-free stars are thought to be massive [5, 62] and their extended envelopes may suppress the emergence of relativistic jets out of their surface (even if such jets are produced through the collapse of their core to a spinning black hole). On the other hand, low-metallicity stars are expected to have weak winds with little angular momentum loss during their evolution, and so they may preferentially yield rotating central configurations that make GRB jets after core collapse.

What kind of metal-free, Pop III progenitor stars may lead to GRBs? Long-duration GRBs appear to be associated with Type Ib/c supernovae [350], namely progenitor massive stars that have lost their hydrogen envelope. This requirement is explained theoretically in the collapse model, in which the relativistic jets produced by core collapse to a black hole are unable to emerge relativistically out of the stellar surface if the hydrogen envelope is retained [230]. The question then arises as to whether the lack of metal line-opacity that is essential for radiation-driven winds in metal-rich stars, would make a Pop III star retain its hydrogen envelope, thus quenching any relativistic jets and GRBs.

Aside from mass transfer in a binary system, individual Pop III stars could lose their hydrogen envelope due to either: (i) violent pulsations, particularly in the mass range 100–140 $M_\odot$, or (ii) a wind driven by helium lines. The outer stellar layers are in a state where gravity only marginally exceeds radiation pressure due to electron-scattering (Thomson) opacity. Adding the small, but still non-negligible contribution from the bound-free opacity provided by singly-ionized helium, may be able to unbind the atmospheric gas. Therefore, mass-loss might occur even in the absence of dust or any heavy elements.

5.4 Emission Spectrum of Metal-Free Stars

The evolution of metal-free (Population III) stars is qualitatively different from that of enriched (Population I and II) stars. In the absence of the catalysts necessary for the operation of the CNO cycle, nuclear burning does not proceed in the standard way. At first, hydrogen burning can only occur via the inefficient PP chain. To provide the necessary luminosity, the star has to reach very high central temperatures ($T_c \simeq 10^{8.1}\,K$). These temperatures are high enough for the spontaneous turn-on of helium burning via the triple-α process. After a brief initial period of triple-α burning, a trace amount of heavy elements forms. Subsequently, the star follows the CNO cycle. In constructing main-sequence models, it is customary to assume that a trace mass fraction of metals ($Z \sim 10^{-9}$) is already present in the star (El Eid 1983 [115]; Castellani et al. 1983 [78]).

Figures 33 and 34 show the luminosity L vs. effective temperature T for zero-age main sequence stars in the mass ranges of 2–90 $M_\odot$ (Fig. 33) and 100–1000 $M_\odot$ (Fig. 34). Note that above $\sim$100 $M_\odot$ the effective temperature is roughly constant, $T_{\rm eff} \sim 10^5\,K$, implying that the spectrum is independent

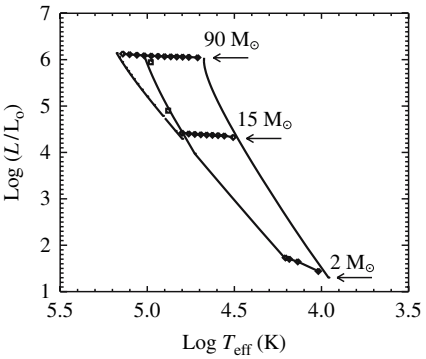

Fig. 33. Luminosity vs. effective temperature for zero-age main sequences stars in the mass range of 2–90 $M_\odot$ (from Tumlinson & Shull 2000 [369]). The curves show Pop I ($Z_\odot = 0.02$) and Pop III stars of mass 2, 5, 8, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, and 90 $M_\odot$. The diamonds mark decades in metallicity in the approach to $Z = 0$ from 10^{-2} down to 10^{-5} at 2 $M_\odot$, down to 10^{-10} at 15 $M_\odot$, and down to 10^{-13} at 90 $M_\odot$. The dashed line along the Pop III ZAMS assumes pure H-He composition, while the solid line (**on the left**) marks the upper MS with $Z_C = 10^{-10}$ for the $M \geqslant 15$ $M_\odot$ models. Squares mark the points corresponding to pre-enriched evolutionary models from El Eid et al. (1983) [115] at 80 $M_\odot$ and from Castellani et al. (1983) [78] for 25 $M_\odot$

of the mass distribution of the stars in this regime (Bromm et al. 2001 [59]). As is evident from these figures (see also Tumlinson & Shull 2000 [369]), both the effective temperature and the ionizing power of metal-free (Pop III) stars are substantially larger than those of metal-rich (Pop I) stars. Metal-free stars with masses $\gtrsim$ 20 $M_\odot$ emit between 10^{47} and 10^{48} H I and He I ionizing photons per second per solar mass of stars, where the lower value applies to stars of $\sim$ 20 $M_\odot$ and the upper value applies to stars of $\gtrsim$ 100 $M_\odot$ (see Tumlinson & Shull 2000 [369] and Bromm et al. 2001 [59] for more details). Over a lifetime of $\sim 3 \times 10^6$ yr these massive stars produce 10^4–10^5 ionizing photons per stellar baryon. However, this powerful UV emission is suppressed as soon as the interstellar medium out of which new stars form is enriched by trace amounts of metals. Even though the collapsed fraction of baryons is small at the epoch of reionization, it is likely that most of the stars responsible for the reionization of the Universe formed out of enriched gas.

Will it be possible to infer the initial mass function (IMF) of the first stars from spectroscopic observations of the first galaxies? Figure 35 compares the observed spectrum from a Salpeter IMF ($\mathrm{d}N_\star/\mathrm{d}M \propto M^{-2.35}$) and a top-heavy IMF (with all stars more massive than 100 $M_\odot$) for a galaxy at $z_s = 10$. The latter case follows from the assumption that each of the dense clumps in the simulations described in the previous section ends up as a single star with no significant fragmentation or mass loss. The difference between the plotted spectra cannot be confused with simple reddening due to normal

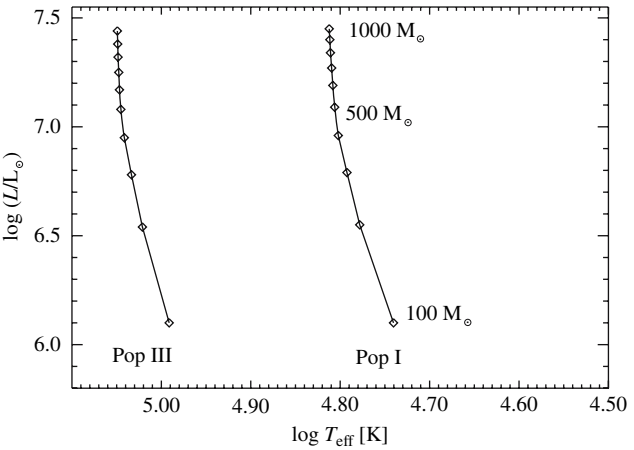

Fig. 34. Same as Fig. 33 but for very massive stars above 100 $M_\odot$ (from Bromm et al. 2001 [59]). *Left solid line:* Pop III zero-age main sequence (ZAMS). *Right solid line:* Pop I ZAMS. In each case, stellar luminosity (in $L_\odot$) is plotted vs. effective temperature (in K). *Diamond-shaped symbols:* Stellar masses along the sequence, from 100 $M_\odot$ (**bottom**) to 1000 $M_\odot$ (**top**) in increments of 100 $M_\odot$. The Pop III ZAMS is systematically shifted to higher effective temperature, with a value of $\sim 10^5$ K which is approximately independent of mass. The luminosities, on the other hand, are almost identical in the two cases

dust. Another distinguishing feature of the IMF is the expected flux in the hydrogen and helium recombination lines, such as Lyα and He II 1640 Å, from the interstellar medium surrounding these stars. We discuss this next.

5.5 Emission of Recombination Lines from the First Galaxies

The hard UV emission from a star cluster or a quasar at high redshift is likely reprocessed by the surrounding interstellar medium, producing very strong recombination lines of hydrogen and helium (Oh 1999 [269]; Tumlinson & Shull 2000 [369]; see also Baltz et al. 1998 [17]). We define $\dot{N}_{\rm ion}$ to be the production rate of ionizing photons by the source. The emitted luminosity $L^{\rm em}_{\rm line}$ per unit stellar mass in a particular recombination line is then estimated to be

$$L^{\rm em}_{\rm line} = p^{\rm em}_{\rm line} h\nu \dot{N}_{\rm ion}(1 - p^{\rm esc}_{\rm cont}) p^{\rm esc}_{\rm line} \ , \qquad (111)$$

where $p^{\rm em}_{\rm line}$ is the probability that a recombination leads to the emission of a photon in the corresponding line, ν is the frequency of the line and $p^{\rm esc}_{\rm cont}$ and $p^{\rm esc}_{\rm line}$ are the escape probabilities for the ionizing photons and the line photons, respectively. It is natural to assume that the stellar cluster is surrounded by a finite H II region, and hence that $p^{\rm esc}_{\rm cont}$ is close to zero [386, 301]. In addition, $p^{\rm esc}_{\rm line}$ is likely close to unity in the H II region, due to the lack of dust in the ambient metal-free gas. Although the emitted line photons may be scattered

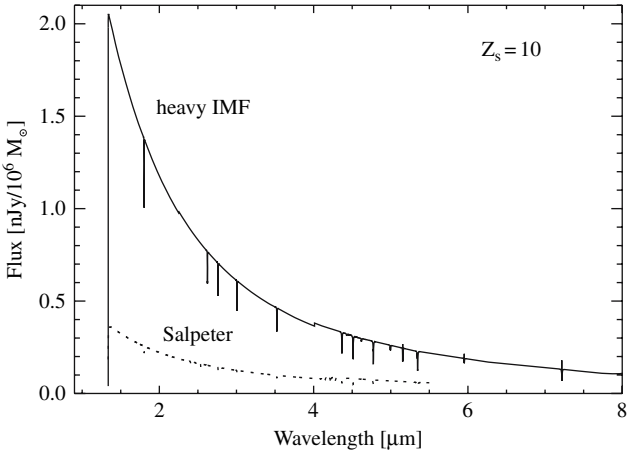

Fig. 35. Comparison of the predicted flux from a Pop III star cluster at $z_s = 10$ for a Salpeter IMF (Tumlinson & Shull 2000 [369]) and a massive IMF (Bromm et al. 2001 [59]). Plotted is the observed flux (in nJy per 10^6 $M_\odot$ of stars) vs. observed wavelength (in μm) for a flat Universe with $\Omega_\Lambda = 0.7$ and $h = 0.65$. *Solid line:* The case of a heavy IMF. *Dotted line:* The fiducial case of a standard Salpeter IMF. The cutoff below $\lambda_{\rm obs} = 1216\,\text{Å}(1 + z_s) = 1.34\,\mu\text{m}$ is due to complete Gunn-Peterson absorption (which is artificially assumed to be sharp). Clearly, for the same total stellar mass, the observable flux is larger by an order of magnitude for stars which are biased towards having masses $\gtrsim 100$ $M_\odot$

by neutral gas, they diffuse out to the observer and in the end survive if the gas is dust free. Thus, for simplicity, we adopt a value of unity for $p_{\rm line}^{\rm esc}$.

As a particular example we consider case B recombination which yields $p_{\rm line}^{\rm em}$ of about 0.65 and 0.47 for the Lyα and He II 1640 Å lines, respectively. These numbers correspond to an electron temperature of $\sim 3 \times 10^4$ K and an electron density of $\sim 10^2 - 10^3\,\text{cm}^{-3}$ inside the H II region [353]. For example, we consider the extreme and most favorable case of metal-free stars all of which are more massive than ~ 100 $M_\odot$. In this case $L_{\rm line}^{\rm em} = 1.7 \times 10^{37}$ and $2.2 \times 10^{36}\,\text{erg}\,\text{s}^{-1}$ $M_\odot^{-1}$ for the recombination luminosities of Lyα and He II 1640 Å per stellar mass [59]. A cluster of 10^6 $M_\odot$ in such stars would then produce 4.4 and 0.6 $\times 10^9$ $L_\odot$ in the Lyα and He II 1640 Å lines. Comparably-high luminosities would be produced in other recombination lines at longer wavelengths, such as He II 4686 Å and Hα [269, 270].

The rest–frame equivalent width of the above emission lines measured against the stellar continuum of the embedded star cluster at the line wavelengths is given by

$$W_\lambda = \left(\frac{L_{\rm line}^{\rm em}}{L_\lambda} \right) \ , \tag{112}$$

where L_λ is the spectral luminosity per unit wavelength of the stars at the line resonance. The extreme case of metal-free stars which are more massive than

100 $M_\odot$ yields a spectral luminosity per unit frequency $L_\nu = 2.7 \times 10^{21}$ and $1.8 \times 10^{21}\,\mathrm{erg\,s^{-1}\,Hz^{-1}\,M_\odot^{-1}}$ at the corresponding wavelengths [59]. Converting to L_λ, this yields rest-frame equivalent widths of $W_\lambda = 3100$ Åand 1100 Åfor Lyα and He II 1640 Å, respectively. These extreme emission equivalent widths are more than an order of magnitude larger than the expectation for a normal cluster of hot metal-free stars with the same total mass and a Salpeter IMF under the same assumptions concerning the escape probabilities and recombination [207]. The equivalent widths are, of course, larger by a factor of $(1+z_s)$ in the observer frame. Extremely strong recombination lines, such as Lyα and He II 1640 Å, are therefore expected to be an additional spectral signature that is unique to very massive stars in the early Universe. The strong recombination lines from the first luminous objects are potentially detectable with *JWST* [270].

6 Supermassive Black Holes

6.1 The Principle of Self-Regulation

The fossil record in the present-day Universe indicates that every bulged galaxy hosts a supermassive black hole (BH) at its center [204]. This conclusion is derived from a variety of techniques which probe the dynamics of stars and gas in galactic nuclei. The inferred BHs are dormant or faint most of the time, but ocassionally flash in a short burst of radiation that lasts for a small fraction of the Hubble time. The short duty cycle acounts for the fact that bright quasars are much less abundant than their host galaxies, but it begs the more fundamental question: *why is the quasar activity so brief?* A natural explanation is that quasars are suicidal, namely the energy output from the BHs regulates their own growth.

Supermassive BHs make up a small fraction, $< 10^{-3}$, of the total mass in their host galaxies, and so their direct dynamical impact is limited to the central star distribution where their gravitational influence dominates. Dynamical friction on the background stars keeps the BH close to the center. Random fluctuations in the distribution of stars induces a Brownian motion of the BH. This motion can be decribed by the same Langevin equation that captures the motion of a massive dust particle as it responds to random kicks from the much lighter molecules of air around it [86]. The characteristic speed by which the BH wanders around the center is small, $\sim (m_\star/M_{\rm BH})^{1/2}\sigma_\star$, where $m_\star$ and $M_{\rm BH}$ are the masses of a single star and the BH, respectively, and $\sigma_\star$ is the stellar velocity dispersion. Since the random force fluctuates on a dynamical time, the BH wanders across a region that is smaller by a factor of $\sim (m_\star/M_{\rm BH})^{1/2}$ than the region traversed by the stars inducing the fluctuating force on it.

The dynamical insignificance of the BH on the global galactic scale is misleading. The gravitational binding energy per rest-mass energy of

galaxies is of order $\sim (\sigma_\star/c)^2 < 10^{-6}$. Since BH are relativistic objects, the gravitational binding energy of material that feeds them amounts to a substantial fraction its rest mass energy. Even if the BH mass occupies a fraction as small as $\sim 10^{-4}$ of the baryonic mass in a galaxy, and only a percent of the accreted rest-mass energy leaks into the gaseous environment of the BH, this slight leakage can unbind the entire gas reservoir of the host galaxy! This order-of-magnitude estimate explains why quasars are short lived. As soon as the central BH accretes large quantities of gas so as to significantly increase its mass, it releases large amounts of energy that would suppress further accretion onto it. In short, the BH growth is *self-regulated*.

The principle of *self-regulation* naturally leads to a correlation between the final BH mass, $M_{\rm bh}$, and the depth of the gravitational potential well to which the surrounding gas is confined which can be characterized by the velocity dispersion of the associated stars, $\sim \sigma_\star^2$. Indeed such a correlation is observed in the present-day Universe [367]. The observed power-law relation between $M_{\rm bh}$ and $\sigma_\star$ can be generalized to a correlation between the BH mass and the circular velocity of the host halo, $v_{\rm c}$ [130], which in turn can be related to the halo mass, $M_{\rm halo}$, and redshift, z [393]

$$\begin{aligned} M_{\rm bh}(M_{\rm halo}, z) &= {\rm const} \times v_{\rm c}^5 \\ &= \epsilon_{\rm o} M_{\rm halo} \left(\frac{M_{\rm halo}}{10^{12}\ {\rm M}_\odot} \right)^{\frac{2}{3}} [\zeta(z)]^{\frac{5}{6}} (1+z)^{\frac{5}{2}}, \end{aligned} \tag{113}$$

where $\epsilon_{\rm o} \approx 10^{-5.7}$ is a constant, and as before $\zeta \equiv [(\Omega_{\rm m}/\Omega_{\rm m}^z)(\Delta_{\rm c}/18\pi^2)]$, $\Omega_{\rm m}^z \equiv [1 + (\Omega_\Lambda/\Omega_{\rm m})(1+z)^{-3}]^{-1}$, $\Delta_{\rm c} = 18\pi^2 + 82d - 39d^2$, and $d = \Omega_{\rm m}^z - 1$. If quasars shine near their Eddington limit as suggested by observations of low and high-redshift quasars [134, 383], then the above value of $\epsilon_{\rm o}$ implies that a fraction of ∼5–10% of the energy released by the quasar over a galactic dynamical time needs to be captured in the surrounding galactic gas in order for the BH growth to be self-regulated [393].

With this interpretation, the $M_{\rm bh}$–$\sigma_\star$ relation reflects the limit introduced to the BH mass by self-regulation; deviations from this relation are inevitable during episodes of BH growth or as a result of mergers of galaxies (see Fig. 36) that have no cold gas in them. A physical scatter around this upper envelope could also result from variations in the efficiency by which the released BH energy couples to the surrounding gas.

Various prescriptions for self-regulation were sketched by Silk & Rees [338]. These involve either energy or momentum-driven winds, where the latter type is a factor of $\sim v_{\rm c}/c$ less efficient [35, 197, 261]. Wyithe & Loeb [393] demonstrated that a particularly simple prescription for an energy-driven wind can reproduce the luminosity function of quasars out to highest measured redshift, $z \sim 6$ (see Figs. 37 and 38), as well as the observed clustering properties of quasars at $z \sim 3$ [397] (see Fig. 39). The prescription postulates that: (i) self-regulation leads to the growth of $M_{\rm bh}$ up the redshift-independent limit as a function of $v_{\rm c}$ in (113), for all galaxies throughout their evolution; and

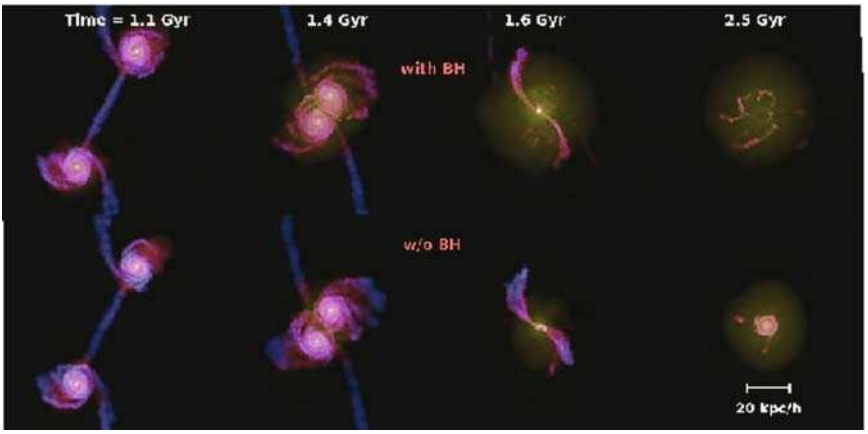

Fig. 36. Simulation images of a merger of galaxies resulting in quasar activity that eventually shuts-off the accretion of gas onto the black hole (from Di Matteo et al. 2005 [108]). The upper (**lower**) panels show a sequence of snapshots of the gas distribution during a merger with (without) feedback from a central black hole. The temperature of the gas is color coded

(ii) the growth of $M_{\rm bh}$ to the limiting mass in (113) occurs through halo merger episodes during which the BH shines at its Eddington luminosity (with the median quasar spectrum) over the dynamical time of its host galaxy, $t_{\rm dyn}$. This model has only one adjustable parameter, namely the fraction of the released quasar energy that couples to the surrounding gas in the host galaxy. This parameter can be fixed based on the $M_{\rm bh}$–$\sigma_\star$ relation in the local Universe [130]. It is remarkable that the combination of the above simple prescription and the standard ΛCDM cosmology for the evolution and merger rate of galaxy halos, lead to a satisfactory agreement with the rich data set on quasar evolution over cosmic history.

The cooling time of the heated gas is typically longer than its dynamical time and so the gas should expand into the galactic halo and escape the galaxy if its initial temperature exceeds the virial temperature of the galaxy [393]. The quasar remains active during the dynamical time of the initial gas reservoir, $\sim 10^7$ yr, and fades afterwards due to the dilution of this reservoir. Accretion is halted as soon as the quasar supplies the galactic gas with more than its binding energy. The BH growth may resume if the cold gas reservoir is replenished through a new merger.

Following the early analytic work, extensive numerical simulations by Springel et al. (2005) [349] (see also Di Matteo et al. 2005 [108]) demonstrated that galaxy mergers do produce the observed correlations between black hole mass and spheroid properties when a similar energy feedback is incorporated. Because of the limited resolution near the galaxy nucleus, these simulations adopt a simple prescription for the accretion flow that feeds the

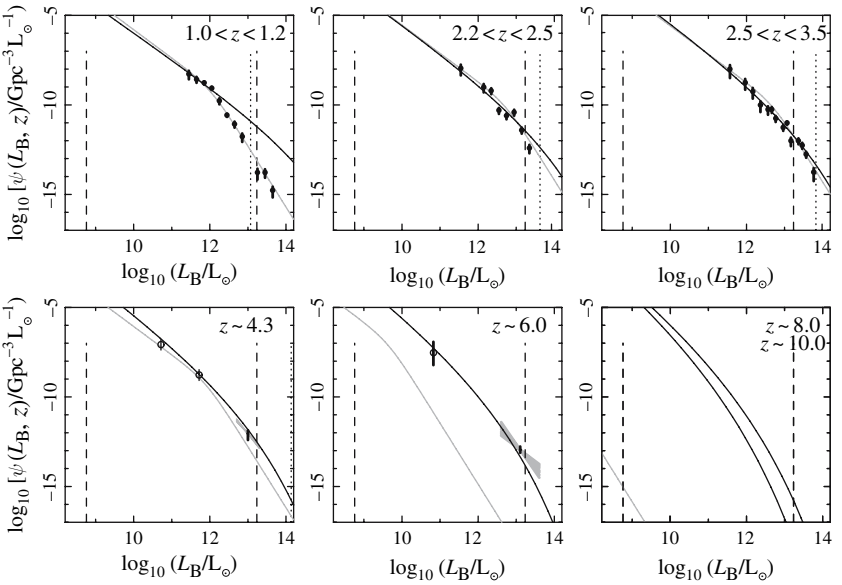

Fig. 37. Comparison of the observed and model luminosity functions (from Wyithe & Loeb 2003 [393]). The data points at $z < 4$ are summarized in Pei (1995) [285], while the light lines show the double power-law fit to the 2dF quasar luminosity function Boyle et al. (2000) [56]. At $z \sim 4.3$ and $z \sim 6.0$ the data is from Fan et al. (2001) [125]. The grey regions show the 1-σ range of logarithmic slope ($[-2.25, -3.75]$ at $z \sim 4.3$ and $[-1.6, -3.1]$ at $z \sim 6$), and the vertical bars show the uncertainty in the normalization. The open circles show data points converted from the X-ray luminosity function Barger et al. (2003) [20] of low luminosity quasars using the median quasar spectral energy distribution. In each panel the vertical dashed lines correspond to the Eddington luminosities of BHs bracketing the observed range of the $M_{\rm bh}$–$v_{\rm c}$ relation, and the vertical dotted line corresponds to a BH in a $10^{13.5}$ $M_\odot$ galaxy

black hole. The actual feedback in reality may depend crucially on the geometry of this flow and the physical mechanism that couples the energy or momentum output of the quasar to the surrounding gas.

Agreement between the predicted and observed correlation function of quasars (Fig. 39) is obtained only if the BH mass scales with redshift as in (113) and the quasar lifetime is of the order of the dynamical time of the host galactic disk [397],

$$t_{\rm dyn} = 10^7 \left[\xi(z)\right]^{-1/2} \left(\frac{1+z}{3}\right)^{-3/2} \text{ yr}. \tag{114}$$

This characterizes the timescale it takes low angular momentum gas to settle inwards and feed the black hole from across the galaxy before feedback

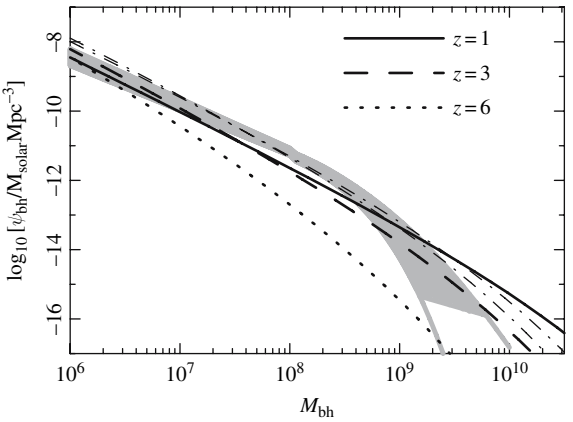

Fig. 38. The comoving density of supermassive BHs per unit BH mass (from Wyithe & Loeb 2003 [393]). The grey region shows the estimate based on the observed velocity distribution function of galaxies in Sheth et al. (2003) [335] and the $M_{\rm bh}$–$v_{\rm c}$ relation in (113). The lower bound corresponds to the lower limit in density for the observed velocity function while the grey lines show the extrapolation to lower densities. We also show the mass function computed at $z = 1$, 3 and 6 from the Press-Schechter (1974) [291] halo mass function and (113), as well as the mass function at $z \sim 2.35$ and $z \sim 3$ implied by the observed density of quasars and a quasar lifetime of order the dynamical time of the host galactic disk, $t_{\rm dyn}$ (*dot-dashed lines*)

sets in and suppresses additional infall. It also characterizes the timescale for establishing an outflow at the escape speed from the host spheroid.

The inflow of cold gas towards galaxy centers during the growth phase of the BH would naturally be accompanied by a burst of star formation. The fraction of gas that is not consumed by stars or ejected by supernovae, will continue to feed the BH. It is therefore not surprising that quasar and starburst activities co-exist in Ultra Luminous Infrared Galaxies [355], and that all quasars show broad metal lines indicating a super-solar metallicity of the surrounding gas [106]. Applying a similar self-regulation principle to the stars, leads to the expectation [393, 195] that the ratio between the mass of the BH and the mass in stars is independent of halo mass (as observed locally [242]) but increases with redshift as $\propto \xi(z)^{1/2}(1+z)^{3/2}$. A consistent trend has indeed been inferred in an observed sample of gravitationally-lensed quasars [304].

The upper mass of galaxies may also be regulated by the energy output from quasar activity. This would account for the fact that cooling flows are suppressed in present-day X-ray clusters [123, 91, 272], and that massive BHs and stars in galactic bulges were already formed at $z \sim 2$. The quasars discovered by the *Sloan Digital Sky Survey* (*SDSS*) at $z \sim 6$ mark the early growth of the most massive BHs and galactic spheroids. The present-day abundance of galaxies capable of hosting BHs of mass $\sim 10^9$ $M_\odot$ (based on 113) already

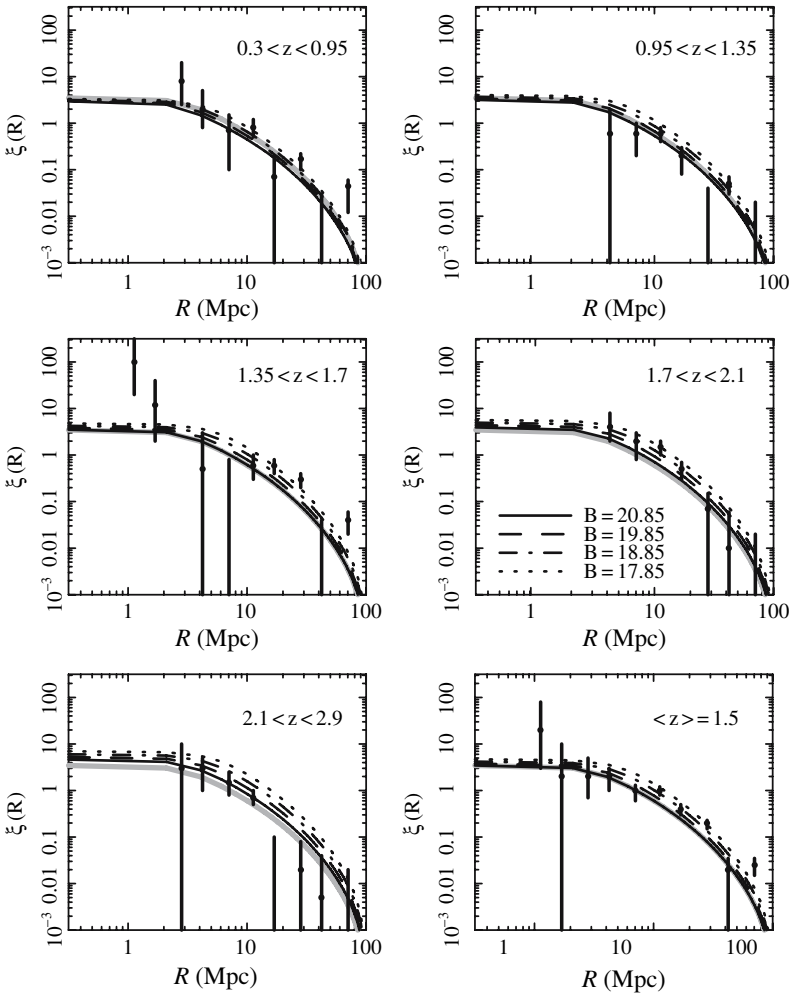

Fig. 39. Predicted correlation function of quasars at various redshifts in comparison to the 2dF data Croom et al. (2001) [101] (from Wyithe & Loeb 2004 [397]). The dark lines show the correlation function predictions for quasars of various apparent B-band magnitudes. The 2dF limit is $B \sim 20.85$. The lower right panel shows data from entire 2dF sample in comparison to the theoretical prediction at the mean quasar redshift of $\langle z \rangle = 1.5$. The $B = 20.85$ prediction at this redshift is also shown by thick gray lines in the other panels to guide the eye. The predictions are based on the scaling $M_{\rm bh} \propto v_{\rm c}^5$ in (113)

existed at $z \sim 6$ [224]. At some epoch, the quasar energy output may have led to the extinction of cold gas in these galaxies and the suppression of further star formation in them, leading to an apparent "anti-hierarchical" mode of galaxy formation where massive spheroids formed early and did not make new stars at late times. In the course of subsequent merger events, the cores of the most massive spheroids acquired an envelope of collisionless matter in the form of already-formed stars or dark matter [224], without the proportional accretion of cold gas into the central BH. The upper limit on the mass of the central BH and the mass of the spheroid is caused by the lack of cold gas and cooling flows in their X-ray halos. In the cores of cooling X-ray clusters, there is often an active central BH that supplies sufficient energy to compensate for the cooling of the gas [91, 123, 35]. The primary physical process by which this energy couples to the gas is still unknown.

6.2 Feedback on Large Intergalactic Scales

Aside from affecting their host galaxy, quasars disturb their large-scale cosmological environment. Powerful quasar outflows are observed in the form of radio jets [34] or broad-absorption-line winds [160]. The amount of energy carried by these outflows is largely unknown, but could be comparable to the radiative output from the same quasars. Furlanetto & Loeb [139] have calculated the intergalactic volume filled by such outflows as a function of cosmic time (see Fig. 40). This volume is likely to contain magnetic fields and metals, providing a natural source for the observed magnetization of the metal-rich gas in X-ray clusters [205] and in galaxies [103]. The injection of energy by quasar outflows may also explain the deficit of Lyα absorption in the vicinity of Lyman-break galaxies [7, 100] and the required pre-heating in X-ray clusters [54, 91].

Beyond the reach of their outflows, the brightest *SDSS* quasars at $z > 6$ are inferred to have ionized exceedingly large regions of gas (tens of comoving Mpc) around them prior to global reionization (see Fig. 41 and [380, 399]). Thus, quasars must have suppressed the faint-end of the galaxy luminosity function in these regions before the same occurred throughout the Universe. The recombination time is comparable to the Hubble time for the mean gas density at $z \sim 7$ and so ionized regions persist [271] on these large scales where inhomogeneities are small. The minimum galaxy mass is increased by at least an order of magnitude to a virial temperature of $\sim 10^5\,\mathrm{K}$ in these ionized regions [23]. It would be particularly interesting to examine whether the faint end ($\sigma_\star < 30\,\mathrm{km\,s^{-1}}$) of the luminosity function of dwarf galaxies shows any moduluation on large-scales around rare massive BHs, such as M87.

To find the volume filling fraction of relic regions from $z \sim 6$, we consider a BH of mass $M_{\rm bh} \sim 3 \times 10^9\ \mathrm{M}_\odot$. We can estimate the comoving density of BHs directly from the observed quasar luminosity function and our estimate of quasar lifetime. At $z \sim 6$, quasars powered by $M_{\rm bh} \sim 3 \times 10^9\ \mathrm{M}_\odot$ BHs had a comoving density of $\sim 0.5\,\mathrm{Gpc}^{-3}$[393]. However, the Hubble time exceeds

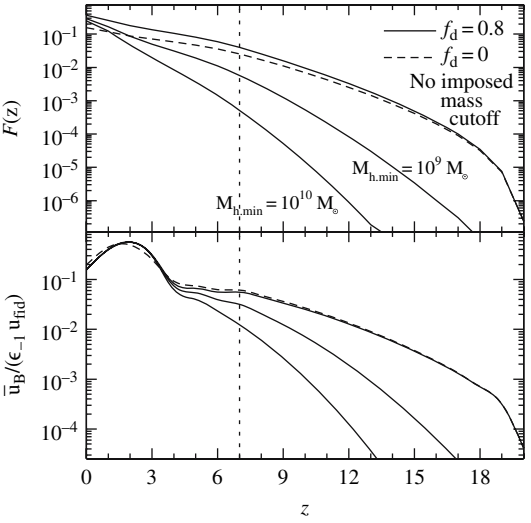

Fig. 40. The global influence of magnetized quasar outflows on the intergalactic medium (from Furlanetto & Loeb 2000 [139]). **Upper Panel:** Predicted volume filling fraction of magnetized quasar bubbles $F(z)$, as a function of redshift. **Lower Panel:** Ratio of normalized magnetic energy density, $\bar{u}_{\rm B}/\epsilon_{-1}$, to the fiducial thermal energy density of the intergalactic medium $u_{\rm fid} = 3n(z)kT_{\rm IGM}$, where $T_{\rm IGM} = 10^4$ K, as a function of redshift (see Furlanetto & Loeb 2000 [139] for more details). In each panel, the solid curves assume that the blast wave created by quasar ouflows is nearly (80%) adiabatic, and that the minimum halo mass of galaxies, $M_{\rm h,min}$, is determined by atomic cooling before reionization and by suppression due to galactic infall afterwards (*top curve*), $M_{\rm h,min} = 10^9$ M$_\odot$ (*middle curve*), and $M_{\rm h,min} = 10^{10}$ M$_\odot$ (*bottom curve*). The dashed curve assumes a fully-radiative blast wave and fixes $M_{h,min}$ by the thresholds for atomic cooling and infall suppression. The vertical dotted line indicates the assumed redshift of complete reionization, $z_{\rm r} = 7$

$t_{\rm dyn}$ by a factor of $\sim 2 \times 10^2$ (reflecting the square root of the density contrast of cores of galaxies relative to the mean density of the Universe), so that the comoving density of the bubbles created by the $z \sim 6$ BHs is $\sim 10^2\,{\rm Gpc}^{-3}$ (see Fig. 38). The density implies that the volume filling fraction of relic $z \sim 6$ regions is small, $< 10\%$, and that the nearest BH that had $M_{\rm bh} \sim 3 \times 10^9$ M$_\odot$ at $z \sim 6$ (and could have been detected as an *SDSS* quasar then) should be at a distance $d_{\rm bh} \sim \left(4\pi/3 \times 10^2\right)^{1/3}$ Gpc ~ 140 Mpc which is almost an order-of-magnitude larger than the distance of M87, a galaxy known to possess a BH of this mass [135].

What is the most massive BH that can be detected dynamically in a local galaxy redshift survey? SDSS probes a volume of $\sim 1\,{\rm Gpc}^3$ out to a distance ~ 30 times that of M87. At the peak of quasar activity at $z \sim 3$, the density of the brightest quasars implies that there should be ~ 100 BHs with masses of 3×10^{10} M$_\odot$ per Gpc^3, the nearest of which will be at a distance

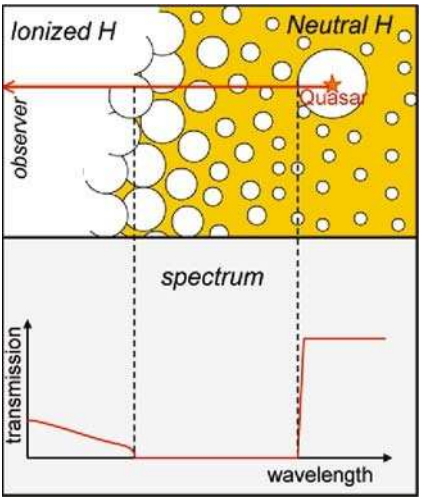

Fig. 41. Quasars serve as probes of the end of reionization. The measured size of the H II regions around *SDSS* quasars can be used Wyithe & Loeb (2004) [395], Mesinger & Haiman (2004) [250] to demonstrate that a significant fraction of the intergalactic hydrogen was neutral at $z \sim 6.3$ or else the inferred size of the quasar H II regions would have been much larger than observed (assuming typical quasar lifetimes Martini 2003 [247]). Also, quasars can be used to measure the redshift at which the intergalactic medium started to transmit Lyα photons White et al. (2003) [380], Wyithe & Loeb (2005) [399]. The upper panel illustrates how the line-of-sight towards a quasar intersects this transition redshift. The resulting Lyα transmission of the intrinsic quasar spectrum is shown schematically in the lower panel

$d_{\rm bh} \sim 130\,{\rm Mpc}$, or ~ 7 times the distance to M87. The radius of gravitational influence of the BH scales as $M_{\rm bh}/v_{\rm c}^2 \propto M_{\rm bh}^{3/5}$. We find that for the nearest 3×10^9 M$_\odot$ and 3×10^{10} M$_\odot$ BHs, the angular radius of influence should be similar. Thus, the dynamical signature of $\sim 3 \times 10^{10}$ M$_\odot$ BHs on their stellar host should be detectable.

6.3 What Seeded the Growth of the Supermassive Black Holes?

The BHs powering the bright *SDSS* quasars possess a mass of a few $\times 10^9$ M$_\odot$, and reside in galaxies with a velocity dispersion of $\sim 500\,{\rm km\,s^{-1}}$ [24]. A quasar radiating at its Eddington limiting luminosity, $L_{\rm E} = 1.4 \times 10^{46}\,{\rm erg\,s^{-1}}\,(M_{\rm bh}/10^8$ M$_\odot$), with a radiative efficiency, $\epsilon_{\rm rad} = L_{\rm E}/\dot{M}c^2$ would grow exponentially in mass as a function of time t, $M_{\rm bh} = M_{\rm seed}\exp\{t/t_{\rm E}\}$ on a time scale, $t_{\rm E} = 4.1 \times 10^7\,{\rm yr}(\epsilon_{\rm rad}/0.1)$. Thus, the required growth time in units of the Hubble time $t_{\rm hubble} = 9 \times 10^8\,{\rm yr}[(1+z)/7]^{-3/2}$ is

$$\frac{t_{\rm growth}}{t_{\rm hubble} = 0.7}\left(\frac{\epsilon_{\rm rad}}{10\%}\right)\left(\frac{1+z}{7}\right)^{3/2}\ln\left(\frac{M_{\rm bh}/10^9\ {\rm M}_\odot}{M_{\rm seed}/100\ {\rm M}_\odot}\right)\ . \qquad (115)$$

The age of the Universe at $z \sim 6$ provides just sufficient time to grow an *SDSS* BH with $M_{\rm bh} \sim 10^9$ $M_\odot$ out of a stellar mass seed with $\epsilon_{\rm rad} = 10\%$ [173]. The growth time is shorter for smaller radiative efficiencies, as expected if the seed originates from the optically-thick collapse of a supermassive star (in which case $M_{\rm seed}$ in the logarithmic factor is also larger).

What was the mass of the initial BH seeds? Were they planted in early dwarf galaxies through the collapse of massive, metal free (Pop-III) stars (leading to $M_{\rm seed}$ of hundreds of solar masses) or through the collapse of even more massive, i.e. supermassive, stars [219]? Bromm & Loeb [63] have shown through a hydrodynamical simulation (see Fig. 42) that supermassive stars were likely to form in early galaxies at $z \sim 10$ in which the virial temperature was close to the cooling threshold of atomic hydrogen, $\sim 10^4$ K. The gas in these galaxies condensed into massive $\sim 10^6$ $M_\odot$ clumps (the progenitors of supermassive stars), rather than fragmenting into many small clumps (the progenitors of stars), as it does in environments that are much hotter than the cooling threshold. This formation channel requires that a galaxy be close to its cooling threshold and immersed in a UV background that dissociates molecular hydrogen in it. These requirements should make this channel sufficiently rare, so as not to overproduce the cosmic mass density of supermassive BH.

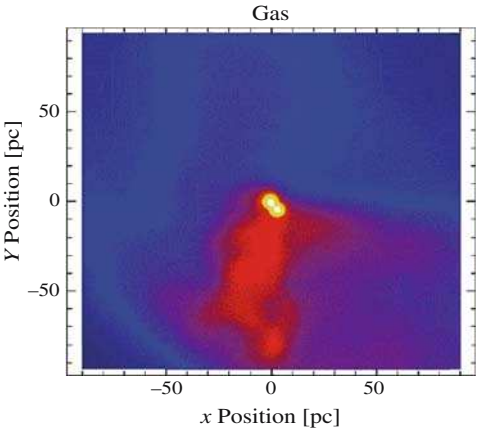

Fig. 42. SPH simulation of the collapse of an early dwarf galaxy with a virial temperature just above the cooling threshold of atomic hydrogen and no H_2 (from Bromm & Loeb 2003 [63]). The image shows a snapshot of the gas density distribution at $z \approx 10$, indicating the formation of two compact objects near the center of the galaxy with masses of 2.2×10^6 $M_\odot$ and 3.1×10^6 $M_\odot$, respectively, and radii < 1 pc. Sub-fragmentation into lower mass clumps is inhibited as long as molecular hydrogen is dissociated by a background UV flux. These circumstances lead to the formation of supermassive stars Loeb & Rasio (1994) [219] that inevitably collapse and trigger the birth of supermassive black holes Loeb & Rasio (1994) [219], Saijo (2002, 2004) [308]. The box size is 200 pc

The minimum seed BH mass can be identified observationally through the detection of gravitational waves from BH binaries with *Advanced LIGO* [394] or with *LISA* [392]. Most of the mHz binary coalescence events originate at $z > 7$ if the earliest galaxies included BHs that obey the $M_{\rm bh}$–$v_{\rm c}$ relation in (113). The number of *LISA* sources per unit redshift per year should drop substantially after reionization, when the minimum mass of galaxies increased due to photo-ionization heating of the intergalactic medium. Studies of the highest redshift sources among the few hundred detectable events per year, will provide unique information about the physics and history of BH growth in galaxies [326].

The early BH progenitors can also be detected as unresolved point sources, using the future *James Webb Space Telescope* (*JWST*). Unfortunately, the spectrum of metal-free massive and supermassive stars is the same, since their surface temperature $\sim 10^5$ K is independent of mass [59]. Hence, an unresolved cluster of massive early stars would show the same spectrum as a supermassive star of the same total mass.

It is difficult to ignore the possible environmental impact of quasars on *anthropic* selection. One may wonder whether it is not a coincidence that our Milky-Way Galaxy has a relatively modest BH mass of only a few million solar masses in that the energy output from a much more massive (e.g. $\sim 10^9$ $M_\odot$) black hole would have disrupted the evolution of life on our planet. A proper calculation remains to be done (as in the context of nearby Gamma-Ray Bursts [315]) in order to demonstrate any such link.

7 Radiative Feedback from the First Sources of Light

7.1 Escape of Ionizing Radiation from Galaxies

The intergalactic ionizing radiation field, a key ingredient in the development of reionization, is determined by the amount of ionizing radiation escaping from the host galaxies of stars and quasars. The value of the escape fraction as a function of redshift and galaxy mass remains a major uncertainty in all current studies, and could affect the cumulative radiation intensity by orders of magnitude at any given redshift. Gas within halos is far denser than the typical density of the IGM, and in general each halo is itself embedded within an overdense region, so the transfer of the ionizing radiation must be followed in the densest regions in the Universe. Reionization simulations are limited in resolution and often treat the sources of ionizing radiation and their immediate surroundings as unresolved point sources within the large-scale intergalactic medium (see, e.g., Gnedin 2000a [152]). The escape fraction is highly sensitive to the three-dimensional distribution of the UV sources relative to the geometry of the absorbing gas within the host galaxy (which may allow escape routes for photons along particular directions but not others).

The escape of ionizing radiation ($h\nu > 13.6\,\mathrm{eV}$, $\lambda < 912$ Å) from the disks of present-day galaxies has been studied in recent years in the context of explaining the extensive diffuse ionized gas layers observed above the disk in the Milky Way [299] and other galaxies [294, 181]. Theoretical models predict that of order 3–14% of the ionizing luminosity from O and B stars escapes the Milky Way disk [111, 110]. A similar escape fraction of $f_{\rm esc} = 6\%$ was determined by Bland-Hawthorn & Maloney (1999) [46] based on Hα measurements of the Magellanic Stream. From *Hopkins Ultraviolet Telescope* observations of four nearby starburst galaxies (Leitherer et al. 1995 [215]; Hurwitz et al. 1997 [183]), the escape fraction was estimated to be in the range $3\% < f_{\rm esc} < 57\%$. If similar escape fractions characterize high redshift galaxies, then stars could have provided a major fraction of the background radiation that reionized the IGM [235, 237]. However, the escape fraction from high-redshift galaxies, which formed when the Universe was much denser ($\rho \propto (1+z)^3$), may be significantly lower than that predicted by models ment to describe present-day galaxies. Current reionization calculations assume that galaxies are isotropic point sources of ionizing radiation and adopt escape fractions in the range $5\% < f_{\rm esc} < 60\%$ [152].

Clumping is known to have a significant effect on the penetration and escape of radiation from an inhomogeneous medium [49, 384, 268, 171, 42]. The inclusion of clumpiness introduces several unknown parameters into the calculation, such as the number and overdensity of the clumps, and the spatial correlation between the clumps and the ionizing sources. An additional complication may arise from hydrodynamic feedback, whereby part of the gas mass is expelled from the disk by stellar winds and supernovae (Sect. 8).

Wood & Loeb (2000) [386] used a three-dimensional radiation transfer code to calculate the steady-state escape fraction of ionizing photons from disk galaxies as a function of redshift and galaxy mass. The gaseous disks were assumed to be isothermal, with a sound speed $c_{\rm s} \sim 10\,\mathrm{km\,s^{-1}}$, and radially exponential, with a scale-length based on the characteristic spin parameter and virial radius of their host halos. The corresponding temperature of $\sim 10^4\,\mathrm{K}$ is typical for a gas which is continuousely heated by photo-ionization from stars. The sources of radiation were taken to be either stars embedded in the disk, or a central quasar. For stellar sources, the predicted increase in the disk density with redshift resulted in a strong decline of the escape fraction with increasing redshift. The situation is different for a central quasar. Due to its higher luminosity and central location, the quasar tends to produce an ionization channel in the surrounding disk through which much of its ionizing radiation escapes from the host. In a steady state, only recombinations in this ionization channel must be balanced by ionizations, while for stars there are many ionization channels produced by individual star-forming regions and the total recombination rate in these channels is very high. Escape fractions $\gtrsim 10\%$ were achieved for stars at $z \sim 10$ only if $\sim 90\%$ of the gas was expelled from the disks or if dense clumps removed the gas from the vast majority ($\gtrsim 80\%$) of the disk volume (see Fig. 43). This analysis applies only to halos with virial

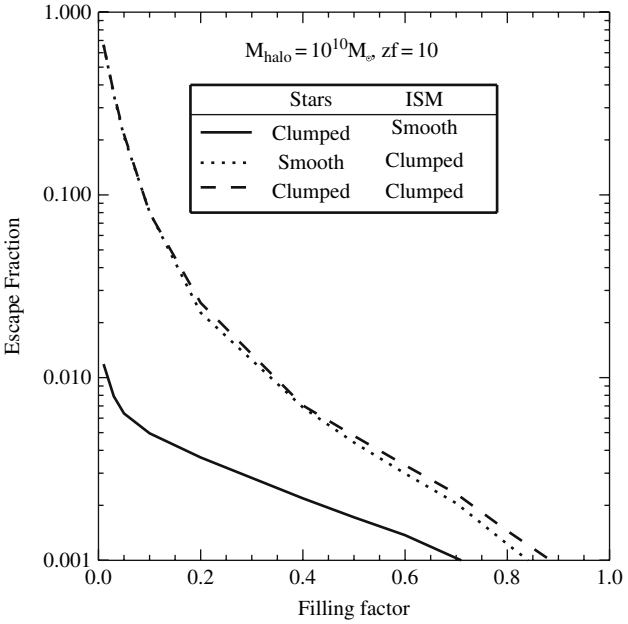

Fig. 43. Escape fractions of stellar ionizing photons from a gaseous disk embedded within a 10^{10} $M_\odot$ halo which has formed at $z = 10$ (from Wood & Loeb 2000 [386]). The curves show three different cases of clumpiness within the disk. The volume filling factor refers to either the ionizing emissivity, the gas clumps, or both, depending on the case. The escape fraction is substantial ($z \gtrsim 1\%$) only if the gas distribution is highly clumped

temperatures $\gtrsim 10^4$ K. Ricotti & Shull (2000) [301] reached similar conclusions but for a quasi-spherical configuration of stars and gas. They demonstrated that the escape fraction is substantially higher in low-mass halos with a virial temperature $\lesssim 10^4$ K. However, the formation of stars in such halos depends on their uncertain ability to cool via the efficient production of molecular hydrogen.

The main uncertainty in the above predictions involves the distribution of the gas inside the host galaxy, as the gas is exposed to the radiation released by stars and the mechanical energy deposited by supernovae. Given the fundamental role played by the escape fraction, it is desirable to calibrate its value observationally. Steidel et al. (2000) [351] reported a detection of significant Lyman continuum flux in the composite spectrum of 29 Lyman break galaxies (LBG) with redshifts in the range $z = 3.40 \pm 0.09$. They co-added the spectra of these galaxies in order to be able to measure the low flux. Another difficulty in the measurement comes from the need to separate the Lyman-limit break caused by the interstellar medium from that already produced in the stellar atmospheres. After correcting for intergalactic absorption, Steidel et al. [351] inferred a ratio between the emergent flux density at 1500 Å and 900 Å (rest

frame) of 4.6 ± 1.0. Taking into account the fact that the stellar spectrum should already have an intrinsic Lyman discontinuity of a factor of $\sim$3–5, but that only $\sim$15–20% of the 1500 Å photons escape from typical LBGs without being absorbed by dust (Pettini et al. 1998 [289]; Adelberger et al. 2000 [6]), the inferred 900 Å escape fraction is $f_{\rm esc} \sim$10–20%. Although the galaxies in this sample were drawn from the bluest quartile of the LBG spectral energy distributions, the measurement implies that this quartile may itself dominate the hydrogen-ionizing background relative to quasars at $z \sim 3$.

7.2 Propagation of Ionization Fronts in the IGM

The radiation output from the first stars ionizes hydrogen in a growing volume, eventually encompassing almost the entire IGM within a single H II bubble. In the early stages of this process, each galaxy produces a distinct H II region, and only when the overall H II filling factor becomes significant do neighboring bubbles begin to overlap in large numbers, ushering in the "overlap phase" of reionization. Thus, the first goal of a model of reionization is to describe the initial stage, when each source produces an isolated expanding H II region.

We assume a spherical ionized volume V, separated from the surrounding neutral gas by a sharp ionization front. Indeed, in the case of a stellar ionizing spectrum, most ionizing photons are just above the hydrogen ionization threshold of 13.6 eV, where the absorption cross-section is high and a very thin layer of neutral hydrogen is sufficient to absorb all the ionizing photons. On the other hand, an ionizing source such as a quasar produces significant numbers of higher energy photons and results in a thicker transition region.

In the absence of recombinations, each hydrogen atom in the IGM would only have to be ionized once, and the ionized proper volume $V_{\rm p}$ would simply be determined by

$$\bar{n}_{\rm H} V_{\rm p} = N_\gamma \ , \tag{116}$$

where $\bar{n}_{\rm H}$ is the mean number density of hydrogen and N_γ is the total number of ionizing photons produced by the source. However, the increased density of the IGM at high redshift implies that recombinations cannot be neglected. Indeed, in the case of a steady ionizing source (and neglecting the cosmological expansion), a steady-state volume would be reached corresponding to the Strömgren sphere, with recombinations balancing ionizations:

$$\alpha_{\rm B} \bar{n}_{\rm H}^2 V_{\rm p} = \frac{{\rm d}\, N_\gamma}{{\rm d}t} \ , \tag{117}$$

where the recombination rate depends on the square of the density and on the case B recombination coefficient $\alpha_{\rm B} = 2.6 \times 10^{-13}\,{\rm cm^3\,s^{-1}}$ for hydrogen at $T = 10^4\,{\rm K}$. The exact evolution for an expanding H II region, including a non-steady ionizing source, recombinations, and cosmological expansion, is given by (Shapiro & Giroux 1987 [328])

$$\bar{n}_{\rm H}\left(\frac{{\rm d}V_{\rm p}}{{\rm d}t}-3HV_{\rm p}\right)=\frac{{\rm d},N_\gamma}{{\rm d}t}-\alpha_{\rm B}\left\langle n_{\rm H}^2\right\rangle V_{\rm p}\ . \tag{118}$$

In this equation, the mean density $\bar{n}_{\rm H}$ varies with time as $1/a^3(t)$. A critical physical ingredient is the dependence of recombination on the square of the density. This means that if the IGM is not uniform, but instead the gas which is being ionized is mostly distributed in high-density clumps, then the recombination time is very short. This is often dealt with by introducing a volume-averaged clumping factor C (in general time-dependent), defined by[8]

$$C=\left\langle n_{\rm H}^2\right\rangle/\bar{n}_{\rm H}^2. \tag{119}$$

If the ionized volume is large compared to the typical scale of clumping, so that many clumps are averaged over, then (118) can be solved by supplementing it with (119) and specifying C. Switching to the comoving volume V, the resulting equation is

$$\frac{{\rm d}V}{{\rm d}t}=\frac{1}{\bar{n}_{\rm H}^0}\frac{{\rm d}\,N_\gamma}{{\rm d}t}-\alpha_{\rm B}\frac{C}{a^3}\bar{n}_{\rm H}^0V\ , \tag{120}$$

where the present number density of hydrogen is

$$\bar{n}_{\rm H}^0=1.88\times10^{-7}\left(\frac{\Omega_{\rm b}h^2}{0.022}\right)\ {\rm cm}^{-3}\ . \tag{121}$$

This number density is lower than the total number density of baryons $\bar{n}_{\rm b}^0$ by a factor of $\sim$0.76, corresponding to the primordial mass fraction of hydrogen. The solution for $V(t)$ (generalized from Shapiro & Giroux 1987 [328]) around a source which turns on at $t=t_{\rm i}$ is

$$V(t)=\int_{t_{\rm i}}^{t}\frac{1}{\bar{n}_{\rm H}^0}\frac{{\rm d}\,N_\gamma}{{\rm d}t'}\,e^{F(t',t)}{\rm d}t'\ , \tag{122}$$

where

$$F(t',t)=-\alpha_{\rm B}\bar{n}_{\rm H}^0\int_{t'}^{t}\frac{C(t'')}{a^3(t'')}\,{\rm d}t''\ . \tag{123}$$

At high redshift (when $(1+z)\gg|{\Omega_{\rm m}}^{-1}-1|$), the scale factor varies as

$$a(t)\simeq\left(\frac{3}{2}\sqrt{\Omega_{\rm m}}H_0t\right)^{2/3}\ , \tag{124}$$

and with the additional assumption of a constant C the function F simplifies as follows. Defining

[8] The recombination rate depends on the number density of electrons, and in using (119) we are neglecting the small contribution caused by partially or fully ionized helium.

$$f(t) = a(t)^{-3/2} \ , \tag{125}$$

we derive

$$F(t',t) = -\frac{2}{3}\frac{\alpha_{\rm B}\bar{n}_{\rm H}^0}{\sqrt{\Omega_{\rm m}}H_0}\,C\,[f(t') - f(t)] = -0.262\,[f(t') - f(t)] \ , \tag{126}$$

where the last equality assumes $C = 10$ and our standard choice of cosmological parameters: $\Omega_{\rm m} = 0.3$, $\Omega_\Lambda = 0.7$, and $\Omega_{\rm b} = 0.045$. Although this expression for $F(t',t)$ is in general an accurate approximation at high redshift, in the particular case of the ΛCDM model (where $\Omega_{\rm m} + \Omega_\Lambda = 1$) we get the exact result by replacing (125) with

$$f(t) = \sqrt{\frac{1}{a^3} + \frac{1-\Omega_{\rm m}}{\Omega_{\rm m}}} \ . \tag{127}$$

The size of the resulting H II region depends on the halo which produces it. Consider a halo of total mass M and baryon fraction $\Omega_{\rm b}/\Omega_{\rm m}$. To derive a rough estimate, we assume that baryons are incorporated into stars with an efficiency of $f_{\rm star} = 10\%$, and that the escape fraction for the resulting ionizing radiation is also $f_{\rm esc} = 10\%$. If the stellar IMF is similar to the one measured locally [314], then $N_\gamma \approx 4000$ ionizing photons are produced per baryon in stars (for a metallicity equal to 1/20 of the solar value). We define a parameter which gives the overall number of ionizations per baryon,

$$N_{\rm ion} \equiv N_\gamma\, f_{\rm star}\, f_{\rm esc} \ . \tag{128}$$

If we neglect recombinations then we obtain the maximum comoving radius of the region which the halo of mass M can ionize,

$$r_{\rm max} = \left(\frac{3}{4\pi}\,\frac{N_\gamma}{\bar{n}_{\rm H}^0}\right)^{1/3} = \left(\frac{3}{4\pi}\,\frac{N_{\rm ion}}{\bar{n}_{\rm H}^0}\,\frac{\Omega_b}{\Omega_{\rm m}}\,\frac{M}{m_{\rm p}}\right)^{1/3} = 680\,{\rm kpc}\left(\frac{N_{\rm ion}}{40}\,\frac{M}{10^8\ {\rm M}_\odot}\right)^{1/3} \ , \tag{129}$$

for our standard set of parameters. The actual radius never reaches this size if the recombination time is shorter than the lifetime of the ionizing source. For an instantaneous starburst with the Scalo (1998) [314] IMF, the production rate of ionizing photons can be approximated as

$$\frac{{\rm d}\,N_\gamma}{{\rm d}t} = \frac{\alpha-1}{\alpha}\,\frac{N_\gamma}{t_{\rm s}} \times \begin{cases} 1 & \text{if } t < t_{\rm s}, \\ \left(\frac{t}{t_{\rm s}}\right)^{-\alpha} & \text{otherwise,} \end{cases} \tag{130}$$

where $N_\gamma = 4000$, $\alpha = 4.5$, and the most massive stars fade away with the characteristic timescale $t_{\rm s} = 3 \times 10^6\,{\rm yr}$. In Fig. 44 we show the time evolution of the volume ionized by such a source, with the volume shown in units of the maximum volume $V_{\rm max}$ which corresponds to $r_{\rm max}$ in (129). We consider a source turning on at $z = 10$ (solid curves) or $z = 15$ (dashed curves), with three cases for each: no recombinations, $C = 1$, and $C = 10$, in order from

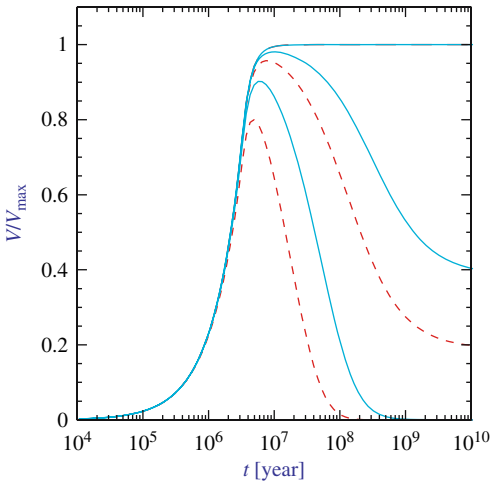

Fig. 44. Expanding H II region around an isolated ionizing source (from Barkana & Loeb 2001 [23]). The comoving ionized volume V is expressed in units of the maximum possible volume, $V_{\rm max} = 4\pi r_{\rm max}^3/3$ [with $r_{\rm max}$ given in (129)], and the time is measured after an instantaneous starburst which produces ionizing photons according to (130). We consider a source turning on at $z = 10$ (*solid curves*) or $z = 15$ (*dashed curves*), with three cases for each: no recombinations, $C = 1$, and $C = 10$, **in order from top to bottom**. The no-recombination curve is identical for the different source redshifts

top to bottom (Note that the result is independent of redshift in the case of no recombinations). When recombinations are included, the volume rises and reaches close to $V_{\rm max}$ before dropping after the source turns off. At large t recombinations stop due to the dropping density, and the volume approaches a constant value (although $V \ll V_{\rm max}$ at large t if $C = 10$).

We obtain a similar result for the size of the H II region around a galaxy if we consider a mini-quasar rather than stars. For the typical quasar spectrum (Elvis et al. 1994 [122]), roughly 11,000 ionizing photons are produced per baryon incorporated into the black hole, assuming a radiative efficiency of $\sim$ 6%. The efficiency of incorporating the baryons in a galaxy into a central black hole is low ($\lesssim 0.6\%$ in the local Universe, e.g. Magorrian et al. 1998 [242]), but the escape fraction for quasars is likely to be close to unity, i.e., an order of magnitude higher than for stars (see previous sub-section). Thus, for every baryon in galaxies, up to $\sim$65 ionizing photons may be produced by a central black hole and ~ 40 by stars, although both of these numbers for $N_{\rm ion}$ are highly uncertain. These numbers suggest that in either case the typical size of H II regions before reionization may be $\lesssim 1\,$Mpc or $\sim 10\,$Mpc, depending on whether 10^8 $M_\odot$ halos or 10^{12} $M_\odot$ halos dominate.

The ionization front around a bright transient source like a quasar expands at early times at nearly the speed of light. This occurs when the H II region

is sufficiently small so that the production rate of ionizing photons by the central source exceeds their consumption rate by hydrogen atoms within this volume. It is straightforward to do the accounting for these rates (including recombinations) taking the light propagation delay into account. This was done by Wyithe & Loeb [395] [see also White et al. (2003) [380]] who derived the general equation for the relativistic expansion of the *comoving* radius [$r = (1+z)r_{\rm p}$] of the quasar H II region in an IGM with a neutral filling fraction $x_{\rm HI}$ (fixed by other ionizing sources) as,

$$\frac{{\rm d}r}{{\rm d}t} = c(1+z)\left[\frac{\dot{N}_\gamma - \alpha_{\rm B} C x_{\rm HI}\left(\bar{n}_{\rm H}^0\right)^2 (1+z)^3 \left(\frac{4\pi}{3} r^3\right)}{\dot{N}_\gamma + 4\pi r^2 (1+z)\, c x_{\rm HI} \bar{n}_{\rm H}^0}\right], \qquad (131)$$

where c is the speed of light, C is the clumping factor, $\alpha_{\rm B} = 2.6\times10^{-13}\,{\rm cm^3\,s^{-1}}$ is the case-B recombination coefficient at the characteristic temperature of $10^4\,{\rm K}$, and $\dot{N}_\gamma$ is the rate of ionizing photons crossing a shell at the radius of the H II region at time t. Indeed, for $\dot{N}_\gamma \to \infty$ the propagation speed of the proper radius of the H II region $r_{\rm p} = r/(1+z)$ approaches the speed of light in the above expression, $({\rm d}r_{\rm p}/{\rm d}t) \to c$. The actual size of the H II region along the line-of-sight to a quasar can be inferred from the extent of the spectral gap between the quasar's rest-frame Lyα wavelength and the start of Lyα absorption by the IGM in the observed spectrum. Existing data from the SDSS quasars [395, 250, 400] provide typical values of $r_{\rm p} \sim 5\,{\rm Mpc}$ and indicate for plausible choices of the quasar lifetimes that $x_{\rm HI} > 0.1$ at $z > 6$. These ionized bubbles could be imaged directly by future 21 cm maps of the regions around the highest-redshift quasars [366, 396, 389].

The profile of the Lyα emission line of galaxies has also been suggested as a probe of the ionization state of the IGM [222, 313, 81, 175, 239, 226, 245]. If the IGM is neutral, then the damping wing of the Gunn-Peterson trough in (108) is modified since Lyα absorption starts only from the near edge of the ionized region along the line-of-sight to the source [81, 239]. Rhoads & Malhotra [245] showed that the observed abundance of galaxies with Lyα emission at $z \sim 6.5$ indicates that a substantial fraction (tens of percent) of the IGM must be ionized in order to allow transmission of the observed Lyα photons. However, if these galaxies reside in groups, then galaxies with peculiar velocities away from the observer will preferentially Doppler-shift the emitted Lyα photons to the red wing of the Lyα resonance and reduce the depression of the line profile [226, 85]. Additional uncertainties in the intrinsic line profile based on the geometry and the stellar or gaseous contents of the source galaxy [226, 313], as well as the clustering of galaxies which ionize their immediate environment in groups [399, 145], limits this method from reaching robust conclusions. Imaging of the expected halos of scattered Lyα radiation around galaxies embedded in a neutral IGM [222, 306] provide a more definitive test of the neutrality of the IGM, but is more challenging observationally.

7.3 Reionization of Hydrogen

In this section we summarize recent progress, both analytic and numerical, made toward elucidating the basic physics of reionization and the way in which the characteristics of reionization depend on the nature of the ionizing sources and on other input parameters of cosmological models.

The process of the reionization of hydrogen involves several distinct stages. The initial, "pre-overlap" stage (using the terminology of Gnedin [152]) consists of individual ionizing sources turning on and ionizing their surroundings. The first galaxies form in the most massive halos at high redshift, and these halos are biased and are preferentially located in the highest-density regions. Thus the ionizing photons which escape from the galaxy itself (see Sect. 7.1) must then make their way through the surrounding high-density regions, which are characterized by a high recombination rate. Once they emerge, the ionization fronts propagate more easily into the low-density voids, leaving behind pockets of neutral, high-density gas. During this period the IGM is a two-phase medium characterized by highly ionized regions separated from neutral regions by ionization fronts. Furthermore, the ionizing intensity is very inhomogeneous even within the ionized regions, with the intensity determined by the distance from the nearest source and by the ionizing luminosity of this source.

The central, relatively rapid "overlap" phase of reionization begins when neighboring H II regions begin to overlap. Whenever two ionized bubbles are joined, each point inside their common boundary becomes exposed to ionizing photons from both sources. Therefore, the ionizing intensity inside H II regions rises rapidly, allowing those regions to expand into high-density gas which had previously recombined fast enough to remain neutral when the ionizing intensity had been low. Since each bubble coalescence accelerates the process of reionization, the overlap phase has the character of a phase transition and is expected to occur rapidly, over less than a Hubble time at the overlap redshift. By the end of this stage most regions in the IGM are able to see several unobscured sources, and therefore the ionizing intensity is much higher than before overlap and it is also much more homogeneous. An additional ingredient in the rapid overlap phase results from the fact that hierarchical structure formation models predict a galaxy formation rate that rises rapidly with time at the relevant redshift range. This process leads to a state in which the low-density IGM has been highly ionized and ionizing radiation reaches everywhere except for gas located inside self-shielded, high-density clouds. This marks the end of the overlap phase, and this important landmark is most often referred to as the "moment of reionization".

Some neutral gas does, however, remain in high-density structures which correspond to Lyman Limit systems and damped Lyα systems seen in absorption at lower redshifts. The high-density regions are gradually ionized as galaxy formation proceeds, and the mean ionizing intensity also grows with time. The ionizing intensity continues to grow and to become more uniform

as an increasing number of ionizing sources is visible to every point in the IGM. This "post-overlap" phase continues indefinitely, since collapsed objects retain neutral gas even in the present Universe. The IGM does, however, reach another milestone at $z \sim 1.6$, the breakthrough redshift [238]. Below this redshift, all ionizing sources are visible to each other, while above this redshift absorption by the Lyα forest implies that only sources in a small redshift range are visible to a typical point in the IGM.

Semi-analytic models of the pre-overlap stage focus on the evolution of the H II filling factor, i.e., the fraction of the volume of the Universe which is filled by H II regions. We distinguish between the naive filling factor $F_{\rm H\ II}$ and the actual filling factor or porosity $Q_{\rm H\ II}$. The naive filling factor equals the number density of bubbles times the average volume of each, and it may exceed unity since when bubbles begin to overlap the overlapping volume is counted multiple times. However, as explained below, in the case of reionization the linearity of the physics means that $F_{\rm H\ II}$ is a very good approximation to $Q_{\rm H\ II}$ up to the end of the overlap phase of reionization.

The model of individual H II regions presented in the previous section can be used to understand the development of the total filling factor. Starting with (120), if we assume a common clumping factor C for all H II regions then we can sum each term of the equation over all bubbles in a given large volume of the Universe, and then divide by this volume. Then V is replaced by the filling factor and N_γ by the total number of ionizing photons produced up to some time t, per unit volume. The latter quantity equals the mean number of ionizing photons per baryon times the mean density of baryons $\bar{n}_{\rm b}$. Following the arguments leading to (129), we find that if we include only stars then

$$\frac{\bar{n}_\gamma}{\bar{n}_{\rm b}} = N_{\rm ion} F_{\rm col} \ , \tag{132}$$

where the collapse fraction $F_{\rm col}$ is the fraction of all the baryons in the Universe which are in galaxies, i.e., the fraction of gas which settles into halos and cools efficiently inside them. In writing (132) we are assuming instantaneous production of photons, i.e., that the timescale for the formation and evolution of the massive stars in a galaxy is short compared to the Hubble time at the formation redshift of the galaxy. In a model based on (120), the near-equality between $F_{\rm H\ II}$ and $Q_{\rm H\ II}$ results from the linearity of this equation. First, the total number of ionizations equals the total number of ionizing photons produced by stars, i.e., all ionizing photons contribute regardless of the spatial distribution of sources; and second, the total recombination rate is proportional to the total ionized volume, regardless of its topology. Thus, even if two or more bubbles overlap the model remains an accurate approximation for $Q_{\rm H\ II}$ (at least until $Q_{\rm H\ II}$ becomes nearly equal to 1). Note, however, that there still are a number of important simplifications in the model, including the assumption of a homogeneous (though possibly time-dependent) clumping factor, and the neglect of feedback whereby the formation of one galaxy may suppress further galaxy formation in neighboring

regions. These complications are discussed in detail below and in Sects. 7.5 and 8.

Under these assumptions we convert (120), which describes individual H II regions, to an equation which statistically describes the transition from a neutral Universe to a fully ionized one (compare to Madau et al. 1999 [238] and Haiman & Loeb 1997 [169]):

$$\frac{dQ_{\rm H\ II}}{dt} = \frac{N_{\rm ion}}{0.76}\frac{{\rm d}F_{\rm col}}{{\rm d}t} - \alpha_{\rm B}\frac{C}{a^3}\bar{n}_{\rm H}^0 Q_{\rm H\ II}\ , \tag{133}$$

where we assumed a primordial mass fraction of hydrogen of 0.76. The solution (in analogy with (122)) is

$$Q_{\rm H\ II}(t) = \int_0^t \frac{N_{\rm ion}}{0.76}\frac{{\rm d}F_{\rm col}}{{\rm d}t'}\, e^{F(t',t)}{\rm d}t'\ , \tag{134}$$

where $F(t',t)$ is determined by (123), (124), (125), (126), (127).

A simple estimate of the collapse fraction at high redshift is the mass fraction (given by (91) in the Press-Schechter model) in halos above the cooling threshold, which is the minimum mass of halos in which gas can cool efficiently. Assuming that only atomic cooling is effective during the redshift range of reionization, the minimum mass corresponds roughly to a halo of virial temperature $T_{\rm vir} = 10^4\,{\rm K}$, which can be converted to a mass using (86). With this prescription we derive (for $N_{\rm ion} = 40$) the reionization history shown in Fig. 45 for the case of a constant clumping factor C. The solid curves show $Q_{\rm H\ II}$ as a function of redshift for a clumping factor $C = 0$ (no recombinations), $C = 1$, $C = 10$, and $C = 30$, in order from left to right. Note that if $C \sim 1$ then recombinations are unimportant, but if $C \gtrsim 10$ then recombinations significantly delay the reionization redshift (for a fixed star-formation history). The dashed curve shows the collapse fraction $F_{\rm col}$ in this model. For comparison, the vertical dotted line shows the $z = 5.8$ observational lower limit (Fan et al. 2000 [124]) on the reionization redshift.

Clearly, star-forming galaxies in CDM hierarchical models are capable of ionizing the Universe at $z \sim 6$–15 with reasonable parameter choices. This has been shown by a large number of theoretical, semi-analytic calculations [138, 329, 169, 372, 89, 92, 391, 83, 370] as well as numerical simulations [79, 148, 152, 2, 295, 95, 341, 202, 184]. Similarly, if a small fraction ($\lesssim 1\%$) of the gas in each galaxy accretes onto a central black hole, then the resulting mini-quasars are also able to reionize the Universe, as has also been shown using semi-analytic models [138, 170, 372, 391].

Although many models yield a reionization redshift around 7–12, the exact value depends on a number of uncertain parameters affecting both the source term and the recombination term in (133). The source parameters include the formation efficiency of stars and quasars and the escape fraction of ionizing photons produced by these sources. The formation efficiency of low mass galaxies may also be reduced by feedback from galactic outflows. These

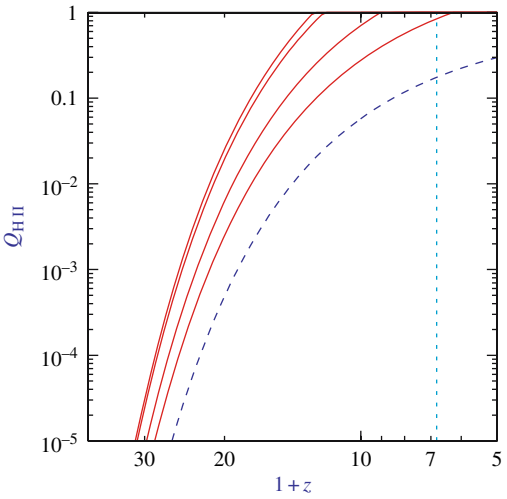

Fig. 45. Semi-analytic calculation of the reionization of the IGM (for $N_{\rm ion} = 40$), showing the redshift evolution of the filling factor $Q_{\rm H\ II}$ (from Barkana & Loeb 2001 [23]). Solid curves show $Q_{\rm H\ II}$ for a clumping factor $C = 0$ (no recombinations), $C = 1$, $C = 10$, and $C = 30$, in order from left to right. The dashed curve shows the collapse fraction $F_{\rm col}$, and the vertical dotted line shows the $z = 5.8$ observational lower limit (Fan et al. 2000 [124]) on the reionization redshift

parameters affecting the sources are discussed elsewhere in this review (see Sects. 7.1 and 8). Even when the clumping is inhomogeneous, the recombination term in (133) is generally valid if C is defined as in (119), where we take a global volume average of the square of the density inside ionized regions (since neutral regions do not contribute to the recombination rate). The resulting mean clumping factor depends on the density and clustering of sources, and on the distribution and topology of density fluctuations in the IGM. Furthermore, the source halos should tend to form in overdense regions, and the clumping factor is affected by this cross-correlation between the sources and the IGM density.

Miralda-Escudé et al. (2000) [255] presented a simple model for the distribution of density fluctuations, and more generally they discussed the implications of inhomogeneous clumping during reionization. They noted that as ionized regions grow, they more easily extend into low-density regions, and they tend to leave behind high-density concentrations, with these neutral islands being ionized only at a later stage. They therefore argued that, since at high-redshift the collapse fraction is low, most of the high-density regions, which would dominate the clumping factor if they were ionized, will in fact remain neutral and occupy only a tiny fraction of the total volume. Thus, the development of reionization through the end of the overlap phase should occur almost exclusively in the low-density IGM, and the effective clumping factor

during this time should be ~ 1, making recombinations relatively unimportant (see Fig. 45). Only in the post-reionization phase, Miralda-Escudé et al. (2000) [255] argued, do the high density clouds and filaments become gradually ionized as the mean ionizing intensity further increases.

The complexity of the process of reionization is illustrated by the numerical simulation of Gnedin [152] of stellar reionization (in ΛCDM with $\Omega_{\rm m} = 0.3$). This simulation uses a formulation of radiative transfer which relies on several rough approximations; although it does not include the effect of shadowing behind optically-thick clumps, it does include for each point in the IGM the effects of an estimated local optical depth around that point, plus a local optical depth around each ionizing source. This simulation helps to understand the advantages of the various theoretical approaches, while pointing to the complications which are not included in the simple models. Figures 46 and 47, taken from Fig. 3 in [152], show the state of the simulated Universe just before and just after the overlap phase, respectively. They show a thin ($15\,h^{-1}$ comoving kpc) slice through the box, which is $4\,h^{-1}$ Mpc on a side, achieves a spatial resolution of $1\,h^{-1}$ kpc, and uses 128^3 each of dark matter particles and baryonic particles (with each baryonic particle having a mass of 5×10^5 $M_\odot$). The figures show the redshift evolution of the mean ionizing intensity J_{21} (upper right panel), and visually the logarithm of the neutral

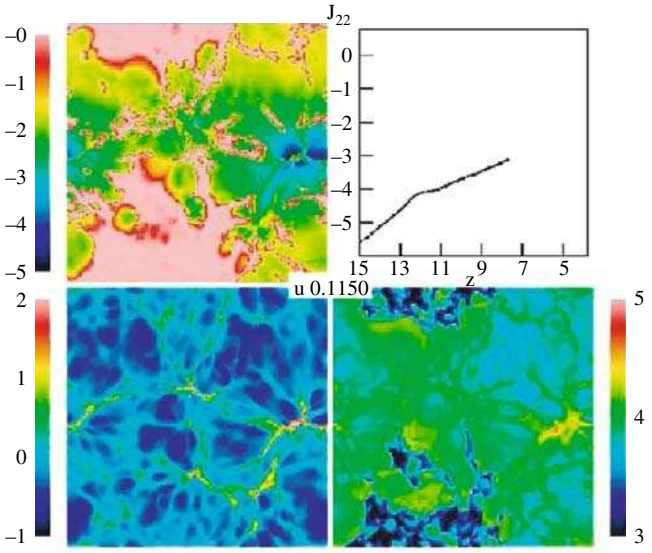

Fig. 46. Visualization at $z = 7.7$ of a numerical simulation of reionization, adopted from Fig. 3c of Gnedin (2000a) [152]. The panels display the logarithm of the neutral hydrogen fraction (**upper left**), the gas density (**lower left**), and the gas temperature (**lower right**). Also shown is the redshift evolution of the logarithm of the mean ionizing intensity (**upper right**). Note the periodic boundary conditions

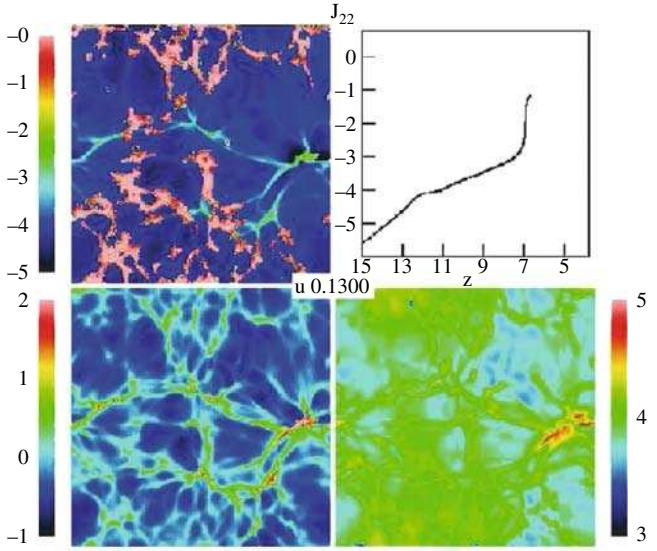

Fig. 47. Visualization at $z = 6.7$ of a numerical simulation of reionization, adopted from Fig. 3e of Gnedin (2000a) [152]. The panels display the logarithm of the neutral hydrogen fraction (**upper left**), the gas density (**lower left**), and the gas temperature (**lower right**). Also shown is the redshift evolution of the logarithm of the mean ionizing intensity (**upper right**). Note the periodic boundary conditions

hydrogen fraction (upper left panel), the gas density (lower left panel), and the gas temperature (lower right panel). Note the obvious features resulting from the periodic boundary conditions assumed in the simulation. Also note that the intensity J_{21} is defined as the intensity at the Lyman limit, expressed in units of $10^{-21}\,\mathrm{erg\,cm^{-2}\,s^{-1}\,sr^{-1}Hz^{-1}}$. For a given source emission, the intensity inside H II regions depends on absorption and radiative transfer through the IGM (e.g., Haardt & Madau 1996 [166]; Abel & Haehnelt 1999 [1]).

Figure 46 shows the two-phase IGM at $z = 7.7$, with ionized bubbles emanating from one main concentration of sources (located at the right edge of the image, vertically near the center; note the periodic boundary conditions). The bubbles are shown expanding into low density regions and beginning to overlap at the center of the image. The topology of ionized regions is clearly complex: While the ionized regions are analogous to islands in an ocean of neutral hydrogen, the islands themselves contain small lakes of dense neutral gas. One aspect which has not been included in theoretical models of clumping is clear from the figure. The sources themselves are located in the highest density regions (these being the sites where the earliest galaxies form) and must therefore ionize the gas in their immediate vicinity before the radiation can escape into the low density IGM. For this reason, the effective clumping factor is of order 100 in the simulation and also, by the overlap redshift,

roughly ten ionizing photons have been produced per baryon. Figure 47 shows that by $z = 6.7$ the low density regions have all become highly ionized along with a rapid increase in the ionizing intensity. The only neutral islands left are the highest density regions (compare the two panels on the left). However, we emphasize that the quantitative results of this simulation must be considered preliminary, since the effects of increased resolution and a more accurate treatment of radiative transfer are yet to be explored. Methods are being developed for incorporating a more complete treatment of radiative transfer into three dimensional cosmological simulations (e.g., [2, 295, 95, 341, 202, 184]).

Gnedin et al. (2000) [151] investigated an additional effect of reionization. They showed that the Biermann battery in cosmological ionization fronts inevitably generates coherent magnetic fields of an amplitude $\sim 10^{-19}$ Gauss. These fields form as a result of the breakout of the ionization fronts from galaxies and their propagation through the H I filaments in the IGM. Although the fields are too small to directly affect galaxy formation, they could be the seeds for the magnetic fields observed in galaxies and X-ray clusters today.

If quasars contribute substantially to the ionizing intensity during reionization then several aspects of reionization are modified compared to the case of pure stellar reionization. First, the ionizing radiation emanates from a single, bright point-source inside each host galaxy, and can establish an escape route (H II funnel) more easily than in the case of stars which are smoothly distributed throughout the galaxy (Sect. 7.1). Second, the hard photons produced by a quasar penetrate deeper into the surrounding neutral gas, yielding a thicker ionization front. Finally, the quasar X-rays catalyze the formation of H_2 molecules and allow stars to keep forming in very small halos.

Oh (1999) [269] showed that star-forming regions may also produce significant X-rays at high redshift. The emission is due to inverse Compton scattering of CMB photons off relativistic electrons in the ejecta, as well as thermal emission by the hot supernova remnant. The spectrum expected from this process is even harder than for typical quasars, and the hard photons photoionize the IGM efficiently by repeated secondary ionizations. The radiation, characterized by roughly equal energy per logarithmic frequency interval, would produce a uniform ionizing intensity and lead to gradual ionization and heating of the entire IGM. Thus, if this source of emission is indeed effective at high redshift, it may have a crucial impact in changing the topology of reionization. Even if stars dominate the emission, the hardness of the ionizing spectrum depends on the initial mass function. At high redshift it may be biased toward massive, efficiently ionizing stars, but this remains very much uncertain.

Semi-analytic as well as numerical models of reionization depend on an extrapolation of hierarchical models to higher redshifts and lower-mass halos than the regime where the models have been compared to observations (see e.g. [391, 83, 370]). These models have the advantage that they are based on the current CDM paradigm which is supported by a variety of observations of large-scale structure, galaxy clustering, and the CMB. The disadvantage is

that the properties of high-redshift galaxies are derived from those of their host halos by prescriptions which are based on low redshift observations, and these prescriptions will only be tested once abundant data is available on galaxies which formed during the reionization era (see [391] for the sensitivity of the results to model parameters). An alternative approach to analyzing the possible ionizing sources which brought about reionization is to extrapolate from the observed populations of galaxies and quasars at currently accessible redshifts. This has been attempted, e.g., by Madau et al. (1999) [238] and Miralda-Escudé et al. (2000) [255]. The general conclusion is that a high-redshift source population similar to the one observed at $z = 3$–4 would produce roughly the needed ionizing intensity for reionization. However, Dijkstra et al. (2004) [107] constrained the role of quasars in reionizing the Universe based on the unresolved flux of the X-ray background. At any event, a precise conclusion remains elusive because of the same kinds of uncertainties as those found in the models based on CDM: The typical escape fraction, and the faint end of the luminosity function, are both not well determined even at $z = 3$–4, and in addition the clumping factor at high redshift must be known in order to determine the importance of recombinations. Future direct observations of the source population at redshifts approaching reionization may help resolve some of these questions.

7.4 Photo-evaporation of Gaseous Halos After Reionization

The end of the reionization phase transition resulted in the emergence of an intense UV background that filled the Universe and heated the IGM to temperatures of ~ 1–2×10^4 K (see the previous section). After ionizing the rarefied IGM in the voids and filaments on large scales, the cosmic UV background penetrated the denser regions associated with the virialized gaseous halos of the first generation of objects. A major fraction of the collapsed gas had been incorporated by that time into halos with a virial temperature $\lesssim 10^4$ K, where the lack of atomic cooling prevented the formation of galactic disks and stars or quasars. Photoionization heating by the cosmic UV background could then evaporate much of this gas back into the IGM. The photo-evaporating halos, as well as those halos which did retain their gas, may have had a number of important consequences just after reionization as well as at lower redshifts.

In this section we focus on the process by which gas that had already settled into virialized halos by the time of reionization was evaporated back into the IGM due to the cosmic UV background. This process was investigated by Barkana & Loeb (1999) [22] using semi-analytic methods and idealized numerical calculations. They first considered an isolated spherical, centrally-concentrated dark matter halo containing gas. Since most of the photo-evaporation occurs at the end of overlap, when the ionizing intensity builds up almost instantaneously, a sudden illumination by an external ionizing background may be assumed. Self-shielding of the gas implies that the halo interior sees a reduced intensity and a harder spectrum, since the outer

gas layers preferentially block photons with energies just above the Lyman limit. It is useful to parameterize the external radiation field by a specific intensity per unit frequency, ν,

$$J_\nu = 10^{-21}\, J_{21} \left(\frac{\nu}{\nu_{\rm L}}\right)^{-\alpha} {\rm erg\,cm^{-2}s^{-1}sr^{-1}\ Hz^{-1}}\ , \qquad (135)$$

where $\nu_{\rm L}$ is the Lyman limit frequency, and J_{21} is the intensity at $\nu_{\rm L}$ expressed in units of $10^{-21}\,{\rm erg\,cm^{-2}\,s^{-1}\,sr^{-1}Hz^{-1}}$. The intensity is normalized to an expected post–reionization value of around unity for the ratio of ionizing photon density to the baryon density. Different power laws can be used to represent either quasar spectra ($\alpha \sim 1.8$) or stellar spectra ($\alpha \sim 5$).

Once the gas is heated throughout the halo, some fraction of it acquires a sufficiently high temperature that it becomes unbound. This gas expands due to the resulting pressure gradient and eventually evaporates back into the IGM. The pressure gradient force (per unit volume) $k\nabla(T\rho/\mu m_{\rm p})$ competes with the gravitational force of $\rho GM/r^2$. Due to the density gradient, the ratio between the pressure force and the gravitational force is roughly equal to the ratio between the thermal energy $\sim kT$ and the gravitational binding energy $\sim \mu m_{\rm p} GM/r$ (which is $\sim kT_{\rm vir}$ at the virial radius $r_{\rm vir}$) per particle. Thus, if the kinetic energy exceeds the potential energy (or roughly if $T > T_{\rm vir}$), the repulsive pressure gradient force exceeds the attractive gravitational force and expels the gas on a dynamical time (or faster for halos with $T \gg T_{\rm vir}$).

The left panel of Fig. 48 (adopted from Fig. 3 of Barkana & Loeb 1999 [22]) shows the fraction of gas within the virial radius which becomes unbound after reionization, as a function of the total halo circular velocity, with halo masses at $z = 8$ indicated at the top. The two pairs of curves correspond to spectral index $\alpha = 5$ (solid) or $\alpha = 1.8$ (dashed). In each pair, a calculation which assumes an optically-thin halo leads to the upper curve, but including radiative transfer and self-shielding modifies the result to the one shown by the lower curve. In each case self-shielding lowers the unbound fraction, but it mostly affects only a neutral core containing $\sim 30\%$ of the gas. Since high energy photons above the Lyman limit penetrate deep into the halo and heat the gas efficiently, a flattening of the spectral slope from $\alpha = 5$ to $\alpha = 1.8$ raises the unbound gas fraction. This figure is essentially independent of redshift if plotted in terms of circular velocity, but the conversion to a corresponding mass does vary with redshift. The characteristic circular velocity where most of the gas is lost is $\sim$10–15 $\rm km\,s^{-1}$, but clearly the effect of photo-evaporation is gradual, going from total gas removal down to no effect over a range of a factor of ~ 100 in halo mass.

Given the values of the unbound gas fraction in halos of different masses, the Press-Schechter mass function (Sect. 4.1) can be used to calculate the total fraction of the IGM which goes through the process of accreting onto a halo and then being recycled into the IGM at reionization. The low-mass cutoff in this sum over halos is given by the lowest mass halo in which gas has assembled by the reionization redshift. This mass can be estimated by the

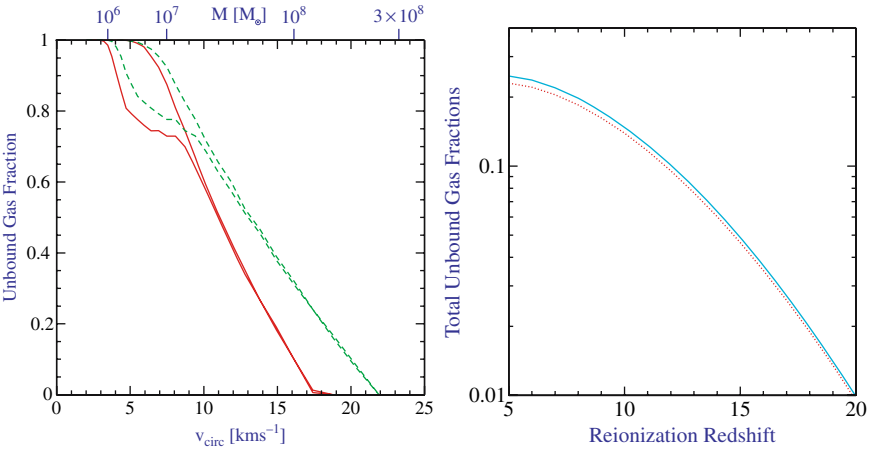

Fig. 48. Effect of photo-evaporation on individual halos and on the overall halo population. The left panel shows the unbound gas fraction (within the virial radius) versus total halo velocity dispersion or mass, adopted from Fig. 3 of Barkana & Loeb (1999) [22]. The two pairs of curves correspond to spectral index $\alpha = 5$ (*solid*) or $\alpha = 1.8$ (*dashed*), in each case at $z = 8$. In each pair, assuming an optically-thin halo leads to the upper curve, while the lower curve shows the result of including radiative transfer and self shielding. The right panel shows the total fraction of gas in the Universe which evaporates from halos at reionization, versus the reionization redshift, adopted from Fig. 7 of Barkana & Loeb (1999) [22]. The solid line assumes a spectral index $\alpha = 1.8$, and the dotted line assumes $\alpha = 5$

linear Jeans mass M_J in (62). The Jeans mass does not in general precisely equal the limiting mass for accretion (see the discussion in the next section). Indeed, at a given redshift some gas can continue to fall into halos of lower mass than the Jeans mass at that redshift. On the other hand, the larger Jeans mass at higher redshifts means that a time-averaged Jeans mass may be more appropriate, as indicated by the filtering mass. In practice, the Jeans mass is sufficiently accurate since at $z \sim 10$–20 it agrees well with the values found in the numerical spherical collapse calculations of Haiman et al. (1996a) [167].

The right panel of Fig. 48 (adopted from Fig. 7 of Barkana & Loeb 1999 [22]) shows the total fraction of gas in the Universe which evaporates from halos at reionization, versus the reionization redshift. The solid line assumes a spectral index $\alpha = 1.8$, and the dotted line assumes $\alpha = 5$, showing that the result is insensitive to the spectrum. Even at high redshift, the amount of gas which participates in photo-evaporation is significant, which suggests a number of possible implications as discussed below. The gas fraction shown in the figure represents most (~ 60–80% depending on the redshift) of the collapsed fraction before reionization, although some gas does remain in more massive halos.

The photo-evaporation of gas out of large numbers of halos may have interesting implications. First, gas which falls into halos and is expelled at reionization attains a different entropy than if it had stayed in the low-density IGM. The resulting overall reduction in the entropy is expected to be small – the same as would be produced by reducing the temperature of the entire IGM by a factor of ~ 1.5 – but localized effects near photo-evaporating halos may be more significant. Furthermore, the resulting $\sim 20\,\mathrm{km\,s^{-1}}$ outflows induce small-scale fluctuations in peculiar velocity and temperature. These outflows are usually well below the resolution limit of most numerical simulations, but some outflows were resolved in the simulation of Bryan et al. (1998) [70]. The evaporating halos may consume a significant number of ionizing photons in the post-overlap stage of reionization [172, 184], but a definitive determination requires detailed simulations which include the three-dimensional geometry of source halos and sink halos.

Although gas is quickly expelled out of the smallest halos, photo- evaporation occurs more gradually in larger halos which retain some of their gas. These surviving halos initially expand but they continue to accrete dark matter and to merge with other halos. These evaporating gas halos could contribute to the high column density end of the Lyα forest [51]. Abel & Mo (1998) [3] suggested that, based on the expected number of surviving halos, a large fraction of the Lyman limit systems at $z \sim 3$ may correspond to mini-halos that survived reionization. Surviving halos may even have identifiable remnants in the present Universe. These ideas thus offer the possibility that a population of halos which originally formed prior to reionization may correspond almost directly to several populations that are observed much later in the history of the Universe. However, the detailed dynamics of photo-evaporating halos are complex, and detailed simulations are required to confirm these ideas. Photo-evaporation of a gas cloud has been followed in a two dimensional simulation with radiative transfer, by Shapiro & Raga (2000) [330]. They found that an evaporating halo would indeed appear in absorption as a damped Lyα system initially, and as a weaker absorption system subsequently. Future simulations [184] will clarify the contribution to quasar absorption lines of the entire population of photo-evaporating halos.

7.5 Suppression of the Formation of Low Mass Galaxies

At the end of overlap, the cosmic ionizing background increased sharply, and the IGM was heated by the ionizing radiation to a temperature $\gtrsim 10^4\,\mathrm{K}$. Due to the substantial increase in the IGM temperature, the intergalactic Jeans mass increased dramatically, changing the minimum mass of forming galaxies [298, 117, 148, 254].

Gas infall depends sensitively on the Jeans mass. When a halo more massive than the Jeans mass begins to form, the gravity of its dark matter overcomes the gas pressure. Even in halos below the Jeans mass, although the gas is initially held up by pressure, once the dark matter collapses its increased

gravity pulls in some gas [167]. Thus, the Jeans mass is generally higher than the actual limiting mass for accretion. Before reionization, the IGM is cold and neutral, and the Jeans mass plays a secondary role in limiting galaxy formation compared to cooling. After reionization, the Jeans mass is increased by several orders of magnitude due to the photoionization heating of the IGM, and hence begins to play a dominant role in limiting the formation of stars. Gas infall in a reionized and heated Universe has been investigated in a number of numerical simulations. Thoul & Weinberg (1996) [362] inferred, based on a spherically-symmetric collapse simulation, a reduction of $\sim 50\%$ in the collapsed gas mass due to heating, for a halo of circular velocity $V_{\rm c} \sim 50\,{\rm km\,s^{-1}}$ at $z = 2$, and a complete suppression of infall below $V_{\rm c} \sim 30\,{\rm km\,s^{-1}}$. Kitayama & Ikeuchi (2000) [199] also performed spherically-symmetric simulations but included self-shielding of the gas, and found that it lowers the circular velocity thresholds by $\sim 5\,{\rm km\,s^{-1}}$. Three dimensional numerical simulations [293, 377, 266] found a significant suppression of gas infall in even larger halos ($V_{\rm c} \sim 75\,{\rm km\,s^{-1}}$), but this was mostly due to a suppression of late infall at $z \lesssim 2$.

When a volume of the IGM is ionized by stars, the gas is heated to a temperature $T_{\rm IGM} \sim 10^4\,{\rm K}$. If quasars dominate the UV background at reionization, their harder photon spectrum leads to $T_{\rm IGM} > 2 \times 10^4\,{\rm K}$. Including the effects of dark matter, a given temperature results in a linear Jeans mass corresponding to a halo circular velocity of

$$V_J = 81 \left(\frac{T_{\rm IGM}}{1.5 \times 10^4\,{\rm K}} \right)^{1/2} \left[\frac{1}{\Omega_{\rm m}^z} \, \frac{\Delta_{\rm c}}{18\pi^2} \right]^{1/6} {\rm km\ s^{-1}}, \qquad (136)$$

where we used (85 and 86) and assumed $\mu = 0.6$. In halos with $V_{\rm c} > V_{\rm J}$, the gas fraction in infalling gas equals the universal mean of $\Omega_{\rm b}/\Omega_{\rm m}$, but gas infall is suppressed in smaller halos. Even for a small dark matter halo, once it collapses to a virial overdensity of $\Delta_{\rm c}/\Omega_{\rm m}^z$ relative to the mean, it can pull in additional gas. A simple estimate of the limiting circular velocity, below which halos have essentially no gas infall, is obtained by substituting the virial overdensity for the mean density in the definition of the Jeans mass. The resulting estimate is

$$V_{\rm lim} = 34 \left(\frac{T_{\rm IGM}}{1.5 \times 10^4\,{\rm K}} \right)^{1/2} {\rm km\ s^{-1}}. \qquad (137)$$

This value is in rough agreement with the numerical simulations mentioned before. A more recent study by Dijkstra et al. (2004) [107] indicates that at the high redshifts of $z > 10$ gas could nevertheless assemble into halos with circular velocities as low as $v_{\rm c} \sim 10\,{\rm km\,s^{-1}}$, even in the presence of a UV background.

Although the Jeans mass is closely related to the rate of gas infall at a given time, it does not directly yield the total gas residing in halos at a given time. The latter quantity depends on the entire history of gas accretion onto

halos, as well as on the merger histories of halos, and an accurate description must involve a time-averaged Jeans mass. Gnedin [153] showed that the gas content of halos in simulations is well fit by an expression which depends on the filtering mass, a particular time-averaged Jeans mass (Gnedin & Hui 1998 [150]). Gnedin [153] calculated the Jeans and filtering masses using the mean temperature in the simulation to define the sound speed, and found the following fit to the simulation results:

$$\bar{M}_{\rm g} = \frac{f_{\rm b} M}{\left[1 + \left(2^{1/3} - 1\right) M_{\rm C}/M\right]^3}, \tag{138}$$

where $\bar{M}_{\rm g}$ is the average gas mass of all objects with a total mass M, $f_{\rm b} = \Omega_{\rm b}/\Omega_{\rm m}$ is the universal baryon fraction, and the characteristic mass $M_{\rm C}$ is the total mass of objects which on average retain 50% of their gas mass. The characteristic mass was well fit by the filtering mass at a range of redshifts from $z = 4$ up to $z \sim 15$.

The reionization process was not perfectly synchronized throughout the Universe. Large-scale regions with a higher density than the mean tend to form galaxies first and reionize earlier than underdense regions (see detailed discussion in Haiman et al. (1996a) [167]). The suppression of low-mass galaxies by reionization will therefore be modulated by the fluctuations in the timing of reionization. Babich & Loeb (2005) [14] considered the effect of inhomogeneous reionization on the power-spectrum of low-mass galaxies. They showed that the shape of the high redshift galaxy power spectrum on small scales in a manner which depends on the details of epoch of reionization. This effect is significantly larger than changes in the galaxy power spectrum due to the current uncertainty in the inflationary parameters, such as the tilt of the scalar power spectrum n and the running of the tilt α. Therefore, future high redshift galaxies surveys hoping to constrain inflationary parameters must properly model the effects of reionization, but conversely they will also be sensitive to the thermal history of the high redshift intergalactic medium.

8 Feedback from Galactic Outflows

8.1 Propagation of Supernova Outflows in the IGM

Star formation is accompanied by the violent death of massive stars in supernova explosions. In general, if each halo has a fixed baryon fraction and a fixed fraction of the baryons turns into massive stars, then the total energy in supernovae outflows is proportional to the halo mass. The binding energy of the gas in the halo is proportional to the halo mass squared. Thus, outflows are expected to escape more easily out of low-mass galaxies, and to expel a greater fraction of the gas from dwarf galaxies. At high redshifts, most galaxies form in relatively low-mass halos, and the high halo merger rate leads to

vigorous star formation. Thus, outflows may have had a great impact on the earliest generations of galaxies, with consequences that may include metal enrichment of the IGM and the disruption of dwarf galaxies. In this subsection we present a simple model for the propagation of individual supernova shock fronts in the IGM. We discuss some implications of this model, but we defer to the following subsection the brunt of the discussion of the cosmological consequences of outflows.

For a galaxy forming in a given halo, the supernova rate is related to the star formation rate. In particular, for a Scalo (1998) [314] initial stellar mass function, if we assume that a supernova is produced by each $M > 8$ $M_\odot$ star, then on average one supernova explodes for every 126 $M_\odot$ of star formation, expelling an ejecta mass of ~ 3 $M_\odot$ including ~ 1 $M_\odot$ of heavy elements. We assume that the individual supernovae produce expanding hot bubbles which merge into a single overall region delineated by an outwardly moving shock front. We assume that most of the baryons in the outflow lie in a thin shell, while most of the thermal energy is carried by the hot interior. The total ejected mass equals a fraction $f_{\rm gas}$ of the total halo gas which is lifted out of the halo by the outflow. This gas mass includes a fraction $f_{\rm eject}$ of the mass of the supernova ejecta itself (with $f_{\rm eject} \leqslant 1$ since some metals may be deposited in the disk and not ejected). Since at high redshift most of the halo gas is likely to have cooled onto a disk, we assume that the mass carried by the outflow remains constant until the shock front reaches the halo virial radius. We assume an average supernova energy of $10^{51} E_{51}$ erg, a fraction $f_{\rm wind}$ of which remains in the outflow after it escapes from the disk. The outflow must overcome the gravitational potential of the halo, which we assume to have a Navarro et al. (1997) [265] density profile [NFW; see (88)]. Since the entire shell mass must be lifted out of the halo, we include the total shell mass as well as the total injected energy at the outset. This assumption is consistent with the fact that the burst of star formation in a halo is typically short compared to the total time for which the corresponding outflow expands.

The escape of an outflow from an NFW halo depends on the concentration parameter $c_{\rm N}$ of the halo. Simulations by Bullock et al. (2000) [72] indicate that the concentration parameter decreases with redshift, and their results may be extrapolated to our regime of interest (i.e., to smaller halo masses and higher redshifts) by assuming that

$$c_{\rm N} = \left(\frac{M}{10^9 \ M_\odot} \right)^{-0.1} \frac{25}{(1+z)}. \tag{139}$$

Although we calculate below the dynamics of each outflow in detail, it is also useful to estimate which halos can generate large-scale outflows by comparing the kinetic energy of the outflow to the potential energy needed to completely escape (i.e., to infinite distance) from an NFW halo. We thus find that the outflow can escape from its originating halo if the circular velocity is below a critical value given by

$$V_{\rm crit} = 200 \sqrt{\frac{E_{51} f_{\rm wind} (\eta/0.1)}{f_{\rm gas}\ g(c_{\rm N})}}\, {\rm km\,s^{-1}}, \tag{140}$$

where the efficiency η is the fraction of baryons incorporated in stars, and

$$g(x) = \frac{x^2}{(1+x)\ln(1+x) - x}. \tag{141}$$

Note that the contribution to $f_{\rm gas}$ of the supernova ejecta itself is $0.024 \eta f_{\rm eject}$, so the ejecta mass is usually negligible unless $f_{\rm gas} \lesssim 1\%$. Equation (140) can also be used to yield the maximum gas fraction $f_{\rm gas}$ which can be ejected from halos, as a function of their circular velocity. Although this equation is most general, if we assume that the parameters $f_{\rm gas}$ and $f_{\rm wind}$ are independent of M and z then we can normalize them based on low-redshift observations. If we specify $c_{\rm N} \sim 10$ (with $g(10) = 6.1$) at $z = 0$, then setting $E_{51} = 1$ and $\eta = 10\%$ yields the required energy efficiency as a function of the ejected halo gas fraction:

$$f_{\rm wind} = 1.5 f_{\rm gas} \left[\frac{V_{\rm crit}}{100\,{\rm km\,s^{-1}}} \right]^2. \tag{142}$$

A value of $V_{\rm crit} \sim 100\,{\rm km\,s^{-1}}$ is suggested by several theoretical and observational arguments which are discussed in the next subsection. However, these arguments are not conclusive, and $V_{\rm crit}$ may differ from this value by a large factor, especially at high redshift (where outflows are observationally unconstrained at present). Note the degeneracy between $f_{\rm gas}$ and $f_{\rm wind}$ which remains even if $V_{\rm crit}$ is specified. Thus, if $V_{\rm crit} \sim 100\,{\rm km\,s^{-1}}$ then a high efficiency $f_{\rm wind} \sim 1$ is required to eject most of the gas from all halos with $V_c < V_{\rm crit}$, but only $f_{\rm wind} \sim 10\%$ is required to eject 5–10% of the gas. The evolution of the outflow does depend on the value of $f_{\rm wind}$ and not just the ratio $f_{\rm wind}/f_{\rm gas}$, since the shell accumulates material from the IGM which eventually dominates over the initial mass carried by the outflow.

We solve numerically for the spherical expansion of a galactic outflow, elaborating on the basic approach of Tegmark et al. (1993) [357]. We assume that most of the mass m carried along by the outflow lies in a thin, dense, relatively cool shell of proper radius R. The interior volume, while containing only a fraction $f_{\rm int} \ll 1$ of the mass m, carries most of the thermal energy in a hot, isothermal plasma of pressure $p_{\rm int}$ and temperature T. We assume a uniform exterior gas, at the mean density of the Universe (at each redshift), which may be neutral or ionized, and may exert a pressure $p_{\rm ext}$ as indicated below. We also assume that the dark matter distribution follows the NFW profile out to the virial radius, and is at the mean density of the Universe outside the halo virial radius. Note that in reality an overdense distribution of gas as well as dark matter may surround each halo due to secondary infall.

The shell radius R in general evolves as follows:

$$m\frac{\mathrm{d}^2R}{\mathrm{d}t^2} = 4\pi R^2\delta p - \left(\frac{\mathrm{d}R}{\mathrm{d}t} - HR\right)\frac{\mathrm{d}m}{\mathrm{d}t} - \frac{Gm}{R^2}\left(M(R) + \frac{1}{2}m\right) + \frac{8}{3}\pi GRm\rho_\Lambda, \tag{143}$$

where the right-hand-side includes forces due to pressure, sweeping up of additional mass, gravity, and a cosmological constant, respectively. The shell is accelerated by internal pressure and decelerated by external pressure, i.e., $\delta p = p_{\rm int} - p_{\rm ext}$. In the gravitational force, $M(R)$ is the total enclosed mass, not including matter in the shell, and $(1/2)m$ is the effective contribution of the shell mass in the thin-shell approximation [278]. The interior pressure is determined by energy conservation, and evolves according to [357]:

$$\frac{\mathrm{d}p_{\rm int}}{\mathrm{d}t} = \frac{L}{2\pi R^3} - 5\,\frac{p_{\rm int}}{R}\,\frac{\mathrm{d}R}{\mathrm{d}t}, \tag{144}$$

where the luminosity L incorporates heating and cooling terms. We include in L the supernova luminosity $L_{\rm sn}$ (during a brief initial period of energy injection), cooling terms $L_{\rm cool}$, ionization $L_{\rm ion}$, and dissipation $L_{\rm diss}$. For simplicity, we assume ionization equilibrium for the interior plasma, and a primordial abundance of hydrogen and helium. We include in $L_{\rm cool}$ all relevant atomic cooling processes in hydrogen and helium, i.e., collisional processes, Bremsstrahlung emission, and Compton cooling off the CMB. Compton scattering is the dominant cooling process for high-redshift outflows. We include in $L_{\rm ion}$ only the power required to ionize the incoming hydrogen upstream, at the energy cost of 13.6 eV per hydrogen atom. The interaction between the expanding shell and the swept-up mass dissipates kinetic energy. The fraction $f_{\rm d}$ of this energy which is re-injected into the interior depends on complex processes occurring near the shock front, including turbulence, non-equilibrium ionization and cooling, and so (following Tegmark et al. 1993 [357]) we let

$$L_{\rm diss} = \frac{1}{2}f_{\rm d}\frac{\mathrm{d}m}{\mathrm{d}t}\left(\frac{\mathrm{d}R}{\mathrm{d}t} - HR\right)^2, \tag{145}$$

where we set $f_{\rm d} = 1$ and compare below to the other extreme of $f_{\rm d} = 0$.

In an expanding Universe, it is preferable to describe the propagation of outflows in terms of comoving coordinates since, e.g., the critical result is the maximum *comoving* size of each outflow, since this size yields directly the total IGM mass which is displaced by the outflow and injected with metals. Specifically, we apply the following transformation [327, 373]:

$$\mathrm{d}\hat{t} = a^{-2}\mathrm{d}t, \quad \hat{R} = a^{-1}R, \quad \hat{p} = a^5 p, \quad \hat{\rho} = a^3\rho. \tag{146}$$

For $\Omega_\Lambda = 0$, Voit (1996) [373] obtained (with the time origin $\hat{t} = 0$ at redshift z_1):

$$\hat{t} = \frac{2}{\Omega_{\rm m}H_0}\left[\sqrt{1+\Omega_{\rm m}z_1} - \sqrt{1+\Omega_{\rm m}z}\right], \tag{147}$$

while for $\Omega_{\rm m}+\Omega_\Lambda=1$ there is no simple analytic expression. We set $\beta=\hat{R}/\hat{r}_{\rm vir}$, in terms of the virial radius $r_{\rm vir}$ (84) of the source halo. We define α_S^1 as the ratio of the shell mass m to $\frac{4}{3}\pi\hat{\rho}_{\rm b}\,\hat{r}_{\rm vir}^3$, where $\hat{\rho}_{\rm b}=\rho_{\rm b}(z=0)$ is the mean baryon density of the Universe at $z=0$. More generally, we define

$$\alpha_S(\beta)\equiv\frac{m}{\frac{4}{3}\pi\hat{\rho}_{\rm b}\,\hat{\rho}^3}=\begin{cases}\alpha_{\rm S}^1/\beta^3 & {\rm if}\beta<1\\ 1+\left(\alpha_{\rm S}^1-1\right)/\beta^3 & {\rm otherwise.}\end{cases}\tag{148}$$

Here we assumed, as noted above, that the shell mass is constant until the halo virial radius is reached, at which point the outflow begins to sweep up material from the IGM. We thus derive the following equations:

$$\frac{{\rm d}^2\hat{R}}{{\rm d}\hat{t}^2}=\begin{cases}\frac{3}{\alpha_S(\beta)}\frac{\hat{p}}{\hat{\rho}_{\rm b}\,\hat{R}}-\frac{a}{2}\hat{R}H_0^2\Omega_{\rm m}\bar{\delta}(\beta) & {\rm if}\beta<1\\ \frac{3}{\alpha_S(\beta)\hat{R}}\left[\frac{\hat{p}}{\hat{\rho}_{\rm b}}-\left(\frac{{\rm d}\hat{R}}{{\rm d}\hat{t}}\right)^2\right]-\frac{a}{2}\hat{R}H_0^2\Omega_{\rm m}\bar{\delta}(\beta)+\frac{a}{4}\hat{R}H_0^2\Omega_{\rm b}\alpha_S(\beta) & {\rm otherwise,}\end{cases}\tag{149}$$

along with

$$\frac{\rm d}{{\rm d}\hat{t}}\left(\hat{R}^5\hat{p}_{\rm int}\right)=\frac{a^4}{2\pi}L\hat{R}^2\ .\tag{150}$$

In the evolution equation for $\hat{R}$, for $\beta<1$ we assume for simplicity that the baryons are distributed in the same way as the dark matter, since in any case the dark matter halo dominates the gravitational force. For $\beta>1$, however, we correct (via the last term on the right-hand side) for the presence of mass in the shell, since at $\beta\gg1$ this term may become important. The $\beta>1$ equation also includes the braking force due to the swept-up IGM mass. The enclosed mean overdensity for the NFW profile (88) surrounded by matter at the mean density is

$$\bar{\delta}(\beta)=\begin{cases}\frac{\Delta_{\rm c}}{\Omega_{\rm m}^z\beta^3}\frac{\ln(1+c_{\rm N}\beta)-c_{\rm N}\beta/(1+c_{\rm N}\beta)}{\ln(1+c)-c/(1+c)} & {\rm if}\beta<1\\ \left(\frac{\Delta_{\rm c}}{\Omega_{\rm m}^z}-1\right)\frac{1}{\beta^3} & {\rm otherwise.}\end{cases}\tag{151}$$

The physics of supernova shells is discussed in Ostriker & McKee (1988) [278] along with a number of analytical solutions. The propagation of cosmological blast waves has also been computed by Ostriker & Cowie (1981) [277], Bertschinger (1985) [40] and Carr & Ikeuchi (1985) [74]. Voit (1996) [373] derived an exact analytic solution to the fluid equations which, although of limited validity, is nonetheless useful for understanding roughly how the outflow size depends on several of the parameters. The solution requires an idealized case of an outflow which at all times expands into a homogeneous IGM. Peculiar gravitational forces, and the energy lost in escaping from the host halo, are neglected, cooling and ionization losses are also assumed to be negligible, and the external pressure is not included. The dissipated energy is assumed to be retained, i.e., $f_{\rm d}$ is set equal to unity. Under these conditions, the standard

Sedov self-similar solution [323, 324] generalizes to the cosmological case as follows [373]:

$$\hat{R} = \left(\frac{\xi \hat{E}_0}{\hat{\rho}_\mathrm{b}} \right)^{1/5} \hat{t}^{\,2/5}, \tag{152}$$

where $\xi = 2.026$ and $\hat{E}_0 = E_0/(1+z_1)^2$ in terms of the initial (i.e., at $t = \hat{t} = 0$ and $z = z_1$) energy E_0. Numerically, the comoving radius is

$$\hat{R} = 280 \left(\frac{0.022}{\Omega_\mathrm{b} h^2} \frac{E_0}{10^{56}\mathrm{erg}} \right)^{1/5} \left(\frac{10}{1+z_1} \frac{\hat{t}}{10^{10}\ \mathrm{yr}} \right)^{2/5} \ \mathrm{kpc}. \tag{153}$$

In solving the equations described above, we assume that the shock front expands into a pre-ionized region which then recombines after a time determined by the recombination rate. Thus, the external pressure is included initially, it is turned off after the pre-ionized region recombines, and it is then switched back on at a lower redshift when the Universe is reionized. When the ambient IGM is neutral and the pressure is off, the shock loses energy to ionization. In practice we find that the external pressure is unimportant during the initial expansion, although it *is* generally important after reionization. Also, at high redshift ionization losses are much smaller than losses due to Compton cooling. In the results shown below, we assume an instantaneous reionization at $z = 9$.

Figure 49 shows the results for a starting redshift $z = 15$, for a halo of mass 5.4×10^7 $\mathrm{M}_\odot$, stellar mass 8.0×10^5 $\mathrm{M}_\odot$, comoving $\hat{r}_\mathrm{vir} = 12\,\mathrm{kpc}$, and circular velocity $V_\mathrm{c} = 20\,\mathrm{km\ s^{-1}}$. We show the shell comoving radius in units of the virial radius of the source halo (top panel), and the physical peculiar velocity of the shock front (bottom panel). Results are shown (solid curve) for the standard set of parameters $f_\mathrm{int} = 0.1$, $f_\mathrm{d} = 1$, $f_\mathrm{wind} = 75\%$, and $f_\mathrm{gas} = 50\%$. For comparison, we show several cases which adopt the standard parameters except for no cooling (dotted curve), no reionization (short-dashed curve), $f_\mathrm{d} = 0$ (long-dashed curve), or $f_\mathrm{wind} = 15\%$ and $f_\mathrm{gas} = 10\%$ (dot-short dashed curve). When reionization is included, the external pressure halts the expanding bubble. We freeze the radius at the point of maximum expansion (where $\mathrm{d}\hat{R}/\mathrm{d}\hat{t} = 0$), since in reality the shell will at that point begin to spread and fill out the interior volume due to small-scale velocities in the IGM. For the chosen parameters, the bubble easily escapes from the halo, but when f_wind and f_gas are decreased the accumulated IGM mass slows down the outflow more effectively. In all cases the outflow reaches a size of 10–20 times $\hat{r}_\mathrm{vir}$, i.e., 100–200 comoving kpc. If all the metals are ejected (i.e., $f_\mathrm{eject} = 1$), then this translates to an average metallicity in the shell of ~ 1–5×10^{-3} in units of the solar metallicity (which is 2% by mass). The asymptotic size of the outflow varies roughly as $f_\mathrm{wind}^{1/5}$, as predicted by the simple solution in (152), but the asymptotic size is rather insensitive to f_gas (at a fixed f_wind) since the outflow mass becomes dominated by the swept-up IGM mass once $\hat{R} \gtrsim 4\hat{r}_\mathrm{vir}$. With the standard parameter values (i.e., those corresponding to the solid curve), Fig. 49 also shows (dot-long dashed curve) the Voit (1996) [373] solution of

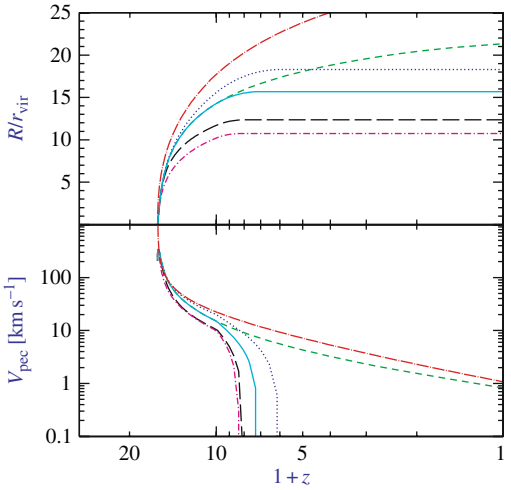

Fig. 49. Evolution of a supernova outflow from a $z = 15$ halo of circular velocity $V_c = 20\,\mathrm{km\ s^{-1}}$ (from Barkana & Loeb 2001 [23]). Plotted are the shell comoving radius in units of the virial radius of the source halo (**top panel**), and the physical peculiar velocity of the shock front (**bottom panel**). Results are shown for the standard parameters $f_{\rm int} = 0.1$, $f_d = 1$, $f_{\rm wind} = 75\%$, and $f_{\rm gas} = 50\%$ (*solid curve*). Also shown for comparison are the cases of no cooling (*dotted curve*), no reionization (*short-dashed curve*), $f_d = 0$ (*long-dashed curve*), or $f_{\rm wind} = 15\%$ and $f_{\rm gas} = 10\%$ (*dot-short dashed curve*), as well as the simple Voit (1996) [373] solution of (152) for the standard parameter set (*dot-long dashed curve*). In cases where the outflow halts, we freeze the radius at the point of maximum expansion

(152). The Voit solution behaves similarly to the no-reionization curve at low redshift, although it overestimates the shock radius by $\sim 30\%$, and the overestimate is greater compared to the more realistic case which does include reionization.

Figure 50 shows different curves than Fig. 49 but on an identical layout. A single curve starting at $z = 15$ (solid curve) is repeated from Fig. 49, and it is compared here to outflows with the same parameters but starting at $z = 20$ (dotted curve), $z = 10$ (short-dashed curve), and $z = 5$ (long-dashed curve). A $V_c = 20\,\mathrm{km\ s^{-1}}$ halo, with a stellar mass equal to 1.5% of the total halo mass, is chosen at the three higher redshifts, but at $z = 5$ a $V_c = 42$ halo is assumed. Because of the suppression of gas infall after reionization, we assume that the $z = 5$ outflow is produced by supernovae from a stellar mass equal to only 0.3% of the total halo mass (with a similarly reduced initial shell mass), thus leading to a relatively small final shell radius. The main conclusion from both figures is the following: In all cases, the outflow undergoes a rapid initial expansion over a fractional redshift interval $\delta z/z \sim 0.2$, at which point the shell has slowed down to $\sim 10\,\mathrm{km\ s^{-1}}$ from an initial $300\,\mathrm{km\ s^{-1}}$. The rapid deceleration is due to the accumulating IGM mass. External pressure from the reionized IGM completely halts all high-redshift outflows, and even without

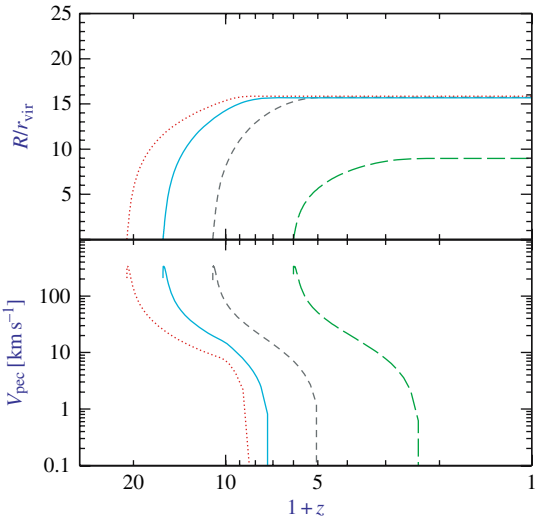

Fig. 50. Evolution of supernova outflows at different redshifts (from Barkana & Loeb 2001 [23]). The top and bottom panels are arranged similarly to Fig. 49. The $z = 15$ outflow (*solid curve*) is repeated from Fig. 49, and it is compared here to outflows with the same parameters but starting at $z = 20$ (*dotted curve*), $z = 10$ (*short-dashed curve*), and $z = 5$ (*long-dashed curve*). A $V_c = 20\,\mathrm{km\ s^{-1}}$ halo is assumed except for $z = 5$, in which case a $V_c = 42\,\mathrm{km\ s^{-1}}$ halo is assumed to produce the outflow (see text)

this effect most outflows would only move at $\sim 10\,\mathrm{km\ s^{-1}}$ after the brief initial expansion. Thus, it may be possible for high-redshift outflows to pollute the Lyman alpha forest with metals without affecting the forest hydrodynamically at $z \lesssim 4$. While the bulk velocities of these outflows may dissipate quickly, the outflows do sweep away the IGM and create empty bubbles. The resulting effects on observations of the Lyman alpha forest should be studied in detail (some observational signatures of feedback have been suggested recently by Theuns et al. 2000 [361]).

Furlanetto & Loeb (2003) [141] derived the evolution of the characteristic scale and filling fraction of supernova-driven bubbles based on a refinement of this formalism (see also their 2001 paper for quasar-driven outflows [139]). The role of metal-rich outflows in smearing the transition epoch between Pop-III (metal-free) and Pop II (metal-enriched) stars, was also analysed by Furlanetto & Loeb (2005) [144], who concluded that a double-reionization history in which the ionization fraction goes through two (or more) peaks is unlikely.

8.2 Effect of Outflows on Dwarf Galaxies and on the IGM

Galactic outflows represent a complex feedback process which affects the evolution of cosmic gas through a variety of phenomena. Outflows inject

hydrodynamic energy into the interstellar medium of their host galaxy. As shown in the previous subsection, even a small fraction of this energy suffices to eject most of the gas from a dwarf galaxy, perhaps quenching further star formation after the initial burst. At the same time, the enriched gas in outflows can mix with the interstellar medium and with the surrounding IGM, allowing later generations of stars to form more easily because of metal-enhanced cooling. On the other hand, the expanding shock waves may also strip gas in surrounding galaxies and suppress star formation.

Dekel & Silk (1986) [104] attempted to explain the different properties of diffuse dwarf galaxies in terms of the effect of galactic outflows. They noted the observed trends whereby lower-mass dwarf galaxies have a lower surface brightness and metallicity, but a higher mass-to-light ratio, than higher mass galaxies. They argued that these trends are most naturally explained by substantial gas removal from an underlying dark matter potential. Galaxies lying in small halos can eject their remaining gas after only a tiny fraction of the gas has turned into stars, while larger galaxies require more substantial star formation before the resulting outflows can expel the rest of the gas. Assuming a wind efficiency $f_{\rm wind} \sim 100\%$, Dekel & Silk showed that outflows in halos below a circular velocity threshold of $V_{\rm crit} \sim 100\,{\rm km\ s^{-1}}$ have sufficient energy to expel most of the halo gas. Furthermore, cooling is very efficient for the characteristic gas temperatures associated with $V_{\rm crit} \lesssim 100\,{\rm km\ s^{-1}}$ halos, but it becomes less efficient in more massive halos. As a result, this critical velocity is expected to signify a dividing line between bright galaxies and diffuse dwarf galaxies. Although these simple considerations may explain a number of observed trends, many details are still not conclusively determined. For instance, even in galaxies with sufficient energy to expel the gas, it is possible that this energy gets deposited in only a small fraction of the gas, leaving the rest almost unaffected.

Since supernova explosions in an inhomogeneous interstellar medium lead to complicated hydrodynamics, in principle the best way to determine the basic parameters discussed in the previous subsection ($f_{\rm wind}$, $f_{\rm gas}$, and $f_{\rm eject}$) is through detailed numerical simulations of individual galaxies. Mac Low & Ferrara (1999) [234] simulated a gas disk within a $z = 0$ dark matter halo. The disk was assumed to be azimuthally symmetric and initially smooth. They represented supernovae by a central source of energy and mass, assuming a constant luminosity which is maintained for 50 million years. They found that the hot, metal-enriched ejecta can in general escape from the halo much more easily than the colder gas within the disk, since the hot gas is ejected in a tube perpendicular to the disk without displacing most of the gas in the disk. In particular, most of the metals were expelled except for the case with the most massive halo considered (with 10^9 $M_\odot$ in gas) and the lowest luminosity ($10^{37}\,{\rm erg\ s^{-1}}$, or a total injection of $2 \times 10^{52}\,{\rm erg}$). On the other hand, only a small fraction of the total gas mass was ejected except for the least massive halo (with 10^6 $M_\odot$ in gas), where a luminosity of $10^{38}\,{\rm erg\ s^{-1}}$ or more expelled most of the gas. We note that beyond the standard issues of numerical

resolution and convergence, there are several difficulties in applying these results to high-redshift dwarf galaxies. Clumping within the expanding shells or the ambient interstellar medium may strongly affect both the cooling and the hydrodynamics. Also, the effect of distributing the star formation throughout the disk is unclear since in that case several characteristics of the problem will change; many small explosions will distribute the same energy over a larger gas volume than a single large explosion [as in the Sedov (1959) [323] solution; see, e.g., (152)], and the geometry will be different as each bubble tries to dig its own escape route through the disk. Also, high-redshift disks should be denser by orders of magnitude than $z = 0$ disks, due to the higher mean density of the Universe at early times. Thus, further numerical simulations of this process are required in order to assess its significance during the reionization epoch.

Some input on these issues also comes from observations. Martin (1999) [246] showed that the hottest extended X-ray emission in galaxies is characterized by a temperature of $\sim 10^{6.7}$ K. This hot gas, which is lifted out of the disk at a rate comparable to the rate at which gas goes into new stars, could escape from galaxies with rotation speeds of $\lesssim 130\,\mathrm{km\ s^{-1}}$. However, these results are based on a small sample which includes only the most vigorous star-forming local galaxies, and the mass-loss rate depends on assumptions about the poorly understood transfer of mass and energy among the various phases of the interstellar medium.

Many authors have attempted to estimate the overall cosmological effects of outflows by combining simple models of individual outflows with the formation rate of galaxies, obtained via semi-analytic methods [98, 357, 373, 264, 129, 316] or numerical simulations [148, 149, 80, 9]. The main goal of these calculations is to explain the characteristic metallicities of different environments as a function of redshift. For example, the IGM is observed to be enriched with metals at redshifts $z \lesssim 5$. Identification of C IV, Si IV an O VI absorption lines which correspond to Lyα absorption lines in the spectra of high-redshift quasars has revealed that the low-density IGM has been enriched to a metal abundance (by mass) of $Z_{\rm IGM} \sim 10^{-2.5(\pm 0.5)}$ $\mathrm{Z}_\odot$, where $\mathrm{Z}_\odot = 0.019$ is the solar metallicity [251, 371, 346, 228, 99, 345, 121]. The metal enrichment has been clearly identified down to H I column densities of $\sim 10^{14.5}\,\mathrm{cm^{-2}}$. The detailed comparison of cosmological hydrodynamic simulations with quasar absorption spectra has established that the forest of Lyα absorption lines is caused by the smoothly-fluctuating density of the neutral component of the IGM [84, 404, 178]. The simulations show a strong correlation between the H I column density and the gas overdensity $\delta_{\rm gas}$ [102], implying that metals were dispersed into regions with an overdensity as low as $\delta_{\rm gas} \sim 3$ or possibly even lower.

In general, dwarf galaxies are expected to dominate metal enrichment at high-redshift for several reasons. As noted above and in the previous subsection, outflows can escape more easily out of the potential wells of dwarfs. Also, at high redshift, massive halos are rare and dwarf halos are much more

common. Finally, as already noted, the Sedov (1959) [323] solution or (152) implies that for a given total energy and expansion time, multiple small outflows fill large volumes more effectively than would a smaller number of large outflows. Note, however, that the strong effect of feedback in dwarf galaxies may also quench star formation rapidly and reduce the efficiency of star formation in dwarfs below that found in more massive galaxies.

Cen & Ostriker (1999) [80] showed via numerical simulation that metals produced by supernovae do not mix uniformly over cosmological volumes. Instead, at each epoch the highest density regions have much higher metallicity than the low-density IGM. They noted that early star formation occurs in the most overdense regions, which therefore reach a high metallicity (of order a tenth of the solar value) by $z \sim 3$, when the IGM metallicity is lower by 1–2 orders of magnitude. At later times, the formation of high-temperature clusters in the highest-density regions suppresses star formation there, while lower-density regions continue to increase their metallicity. Note, however, that the spatial resolution of the hydrodynamic code of Cen & Ostriker is a few hundred kpc, and anything occurring on smaller scales is inserted directly via simple parametrized models. Scannapieco & Broadhurst (2000) [316] implemented expanding outflows within a numerical scheme which, while not a full gravitational simulation, did include spatial correlations among halos. They showed that winds from low-mass galaxies may also strip gas from nearby galaxies (see also Scannapieco, Ferrara, & Broadhurst 2000 [317]), thus suppressing star formation in a local neighborhood and substantially reducing the overall abundance of galaxies in halos below a mass of $\sim 10^{10}$ $M_\odot$. Although quasars do not produce metals, they may also affect galaxy formation in their vicinity via energetic outflows [116, 15, 338, 262].

Gnedin & Ostriker (1997) [148] and Gnedin (1998) [149] identified another mixing mechanism which, they argued, may be dominant at high redshift ($z \gtrsim 4$). In a collision between two protogalaxies, the gas components collide in a shock and the resulting pressure force can eject a few percent of the gas out of the merger remnant. This is the merger mechanism, which is based on gravity and hydrodynamics rather than direct stellar feedback. Even if supernovae inject most of their metals in a local region, larger-scale mixing can occur via mergers. Note, however, that Gnedin's (1998) [149] simulation assumed a comoving star formation rate at $z \gtrsim 5$ of ~ 1 $M_\odot$ per year per comoving Mpc^3, which is 5–10 times larger than the observed rate at redshift 3–4. Aguirre et al. [9] used outflows implemented in simulations to conclude that winds of $\sim 300\,\mathrm{km\ s^{-1}}$ at $z \lesssim 6$ can produce the mean metallicity observed at $z \sim 3$ in the Lyα forest. In a separate paper Aguirre et al. [10] explored another process, where metals in the form of dust grains are driven to large distances by radiation pressure, thus producing large-scale mixing without displacing or heating large volumes of IGM gas. The success of this mechanism depends on detailed microphysics such as dust grain destruction and the effect of magnetic fields. The scenario, though, may be directly testable because it leads to significant ejection only of elements which solidify as grains.

Feedback from galactic outflows encompasses a large variety of processes and influences. The large range of scales involved, from stars or quasars embedded in the interstellar medium up to the enriched IGM on cosmological scales, make possible a multitude of different, complementary approaches, promising to keep galactic feedback an active field of research.

9 The Frontier of 21 cm Cosmology

9.1 Mapping Hydrogen Before Reionization

The small residual fraction of free electrons after cosmological recombination coupled the temperature of the cosmic gas to that of the cosmic microwave background (CMB) down to a redshift, $z \sim 200$ [283]. Subsequently, the gas temperature dropped adiabatically as $T_{\rm gas} \propto (1+z)^2$ below the CMB temperature $T_\gamma \propto (1+z)$. The gas heated up again after being exposed to the photo-ionizing ultraviolet light emitted by the first stars during the *reionization epoch* at $z \lesssim 20$. Prior to the formation of the first stars, the cosmic neutral hydrogen must have resonantly absorbed the CMB flux through its spin-flip 21 cm transition (see Fig. 51) [131, 322, 366, 403]. The linear density fluctuations at that time should have imprinted anisotropies on the CMB sky at an observed wavelength of $\lambda = 21.12[(1+z)/100]$ m. We discuss these early 21 cm fluctuations mainly for pedagogical purposes. Detection of the earliest 21 cm signal will be particularly challenging because the foreground sky brightness rises as $\lambda^{2.5}$ at long wavelengths in addition to the standard $\sqrt{\lambda}$ scaling of the detector noise temperature for a given integration time and fractional bandwidth. The discussion in this section follows Loeb & Zaldarriaga (2004) [225].

We start by calculating the history of the spin temperature, $T_{\rm s}$, defined through the ratio between the number densities of hydrogen atoms in the excited and ground state levels, $n_1/n_0 = (g_1/g_0)\exp\{-T_\star/T_{\rm s}\}$,

$$\frac{n_1}{n_0} = \frac{g_1}{g_0}\exp\left\{-\frac{T_\star}{T_{\rm s}}\right\}, \tag{154}$$

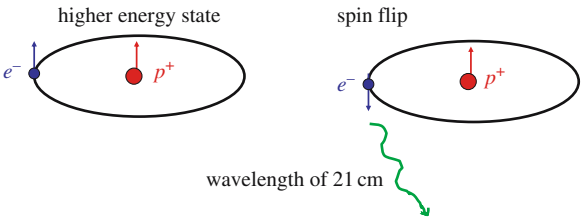

Fig. 51. The 21 cm transition of hydrogen. The higher energy level the spin of the electron (e−) is aligned with that of the proton (p+). A spin flip results in the emission of a photon with a wavelength of 21 cm (or a frequency of 1420 MHz)

where subscripts 1 and 0 correspond to the excited and ground state levels of the 21 cm transition, $(g_1/g_0) = 3$ is the ratio of the spin degeneracy factors of the levels, $n_{\rm H} = (n_0 + n_1) \propto (1+z)^3$ is the total hydrogen density, and $T_\star = 0.068\,{\rm K}$ is the temperature corresponding to the energy difference between the levels. The time evolution of the density of atoms in the ground state is given by,

$$\left(\frac{\partial}{\partial t} + 3\frac{\dot{a}}{a}\right) n_0 = -n_0\left(C_{01} + B_{01}I_\nu\right) + n_1\left(C_{10} + A_{10} + B_{10}I_\nu\right), \quad (155)$$

where $a(t) = (1+z)^{-1}$ is the cosmic scale factor, A's and B's are the Einstein rate coefficients, C's are the collisional rate coefficients, and $I_{\rm \nu}$ is the blackbody intensity in the Rayleigh-Jeans tail of the CMB, namely $I_\nu = 2kT_\gamma/\lambda^2$ with $\lambda = 21\,{\rm cm}$ [305]. Here a dot denotes a time-derivative. The $0 \rightarrow 1$ transition rates can be related to the $1 \rightarrow 0$ transition rates by the requirement that in thermal equilibrium with $T_{\rm s} = T_\gamma = T_{\rm gas}$, the right-hand-side of (155) should vanish with the collisional terms balancing each other separately from the radiative terms. The Einstein coefficients are $A_{10} = 2.85 \times 10^{-15}\,{\rm s}^{-1}$, $B_{10} = (\lambda^3/2hc)A_{10}$ and $B_{01} = (g_1/g_0)B_{10}$ [131, 305]. The collisional de-excitation rates can be written as $C_{10} = \frac{4}{3}\kappa(1-0)n_{\rm H}$, where $\kappa(1-0)$ is tabulated as a function of $T_{\rm gas}$ [11, 405].

Equation (155) can be simplified to the form,

$$\frac{{\rm d}\Upsilon}{{\rm d}z} = -\left[H(1+z)\right]^{-1}\left[-\Upsilon(C_{01} + B_{01}I_\nu) + (1-\Upsilon)(C_{10} + A_{10} + B_{10}I_\nu)\right], \quad (156)$$

where $\Upsilon \equiv n_0/n_{\rm H}$, $H \approx H_0\sqrt{\Omega_{\rm m}}(1+z)^{3/2}$ is the Hubble parameter at high redshifts (with a present-day value of H_0), and $\Omega_{\rm m}$ is the density parameter of matter. The upper panel of Fig. 52 shows the results of integrating (156). Both the spin temperature and the kinetic temperature of the gas track the CMB temperature down to $z \sim 200$. Collisions are efficient at coupling $T_{\rm s}$ and $T_{\rm gas}$ down to $z \sim 70$ and so the spin temperature follows the kinetic temperature around that redshift. At much lower redshifts, the Hubble expansion makes the collision rate subdominant relative the radiative coupling rate to the CMB, and so $T_{\rm s}$ tracks T_γ again. Consequently, there is a redshift window between $30 \lesssim z \lesssim 200$, during which the cosmic hydrogen absorbs the CMB flux at its resonant 21 cm transition. Coincidentally, this redshift interval precedes the appearance of collapsed objects [23] and so its signatures are not contaminated by nonlinear density structures or by radiative or hydrodynamic feedback effects from stars and quasars, as is the case at lower redshifts [403].

During the period when the spin temperature is smaller than the CMB temperature, neutral hydrogen atoms absorb CMB photons. The resonant 21 cm absorption reduces the brightness temperature of the CMB by,

$$T_{\rm b} = \tau\left(T_{\rm s} - T_\gamma\right)/(1+z), \quad (157)$$

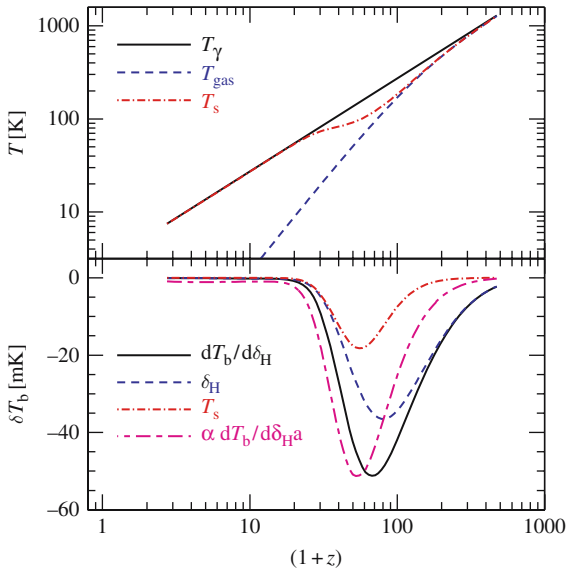

Fig. 52. Upper panel: Evolution of the gas, CMB and spin temperatures with redshift (from Loeb & Zaldarriaga 2004 [225]). **Lower panel:** $\mathrm{d}T_\mathrm{b}/\mathrm{d}\delta_\mathrm{H}$ as function of redshift. The separate contributions from fluctuations in the density and the spin temperature are depicted. We also show $\mathrm{d}T_\mathrm{b}/\mathrm{d}\delta_\mathrm{H} a \propto \mathrm{d}T_\mathrm{b}/\mathrm{d}\delta_\mathrm{H} \times \delta_\mathrm{H}$, with an arbitrary normalization

where the optical depth for resonant 21 cm absorption is,

$$\tau = \frac{3c\lambda^2 h A_{10} n_\mathrm{H}}{32\pi k T_\mathrm{s} H(z)}. \tag{158}$$

Small inhomogeneities in the hydrogen density $\delta_\mathrm{H} \equiv (n_\mathrm{H} - \bar{n}_\mathrm{H})/\bar{n}_\mathrm{H}$ result in fluctuations of the 21 cm absorption through two separate effects. An excess of neutral hydrogen directly increases the optical depth and also alters the evolution of the spin temperature. For now, we ignore the additional effects of peculiar velocities (Bharadwaj & Ali 2004 [41]; Barkana & Loeb 2004 [27]) as well as fluctuations in the gas kinetic temperature due to the adiabatic compression (rarefaction) in overdense (underdense) regions [29]. Under these approximations, we can write an equation for the resulting evolution of Υ fluctuations,

$$\begin{aligned}\frac{\mathrm{d}\delta\Upsilon}{\mathrm{d}z} = [H(1+z)]^{-1} \{&[C_{10} + C_{01} + (B_{01} + B_{10})I_\nu]\delta\Upsilon \\ &+ [C_{01}\Upsilon - C_{10}(1-\Upsilon)]\,\delta_\mathrm{H}\}\,,\end{aligned} \tag{159}$$

leading to spin temperature fluctuations,

$$\frac{\delta T_\mathrm{s}}{\bar{T}_\mathrm{s}} = -\frac{1}{\ln[3\Upsilon/(1-\Upsilon)]}\frac{\delta\Upsilon}{\Upsilon(1-\Upsilon)}. \tag{160}$$

The resulting brightness temperature fluctuations can be related to the derivative,

$$\frac{\delta T_{\rm b}}{\bar{T}_{\rm b}} = \delta_{\rm H} + \frac{T_\gamma}{(\bar{T}_{\rm s} - T_\gamma)} \frac{\delta T_{\rm s}}{\bar{T}_{\rm s}}. \tag{161}$$

The spin temperature fluctuations $\delta T_{\rm s}/T_{\rm s}$ are proportional to the density fluctuations and so we define,

$$\frac{{\rm d}T_{\rm b}}{{\rm d}\delta_{\rm H}} \equiv \bar{T}_{\rm b} + \frac{T_\gamma \bar{T}_{\rm b}}{(\bar{T}_{\rm s} - T_\gamma)} \frac{\delta T_{\rm s}}{\bar{T}_{\rm s}\delta_{\rm H}}, \tag{162}$$

through $\delta T_{\rm b} = ({\rm d}T_{\rm b}/{\rm d}\delta_{\rm H})\delta_{\rm H}$. We ignore fluctuations in $C_{\rm ij}$ due to fluctuations in $T_{\rm gas}$ which are very small [11]. Figure 52 shows ${\rm d}T_b/{\rm d}\delta_{\rm H}$ as a function of redshift, including the two contributions to ${\rm d}T_{\rm b}/{\rm d}\delta_{\rm H}$, one originating directly from density fluctuations and the second from the associated changes in the spin temperature [322]. Both contributions have the same sign, because an increase in density raises the collision rate and lowers the spin temperature and so it allows $T_{\rm s}$ to better track $T_{\rm gas}$. Since $\delta_{\rm H}$ grows with time as $\delta_{\rm H} \propto a$, the signal peaks at $z \sim 50$, a slightly lower redshift than the peak of ${\rm d}T_{\rm b}/{\rm d}\delta_{\rm H}$.

Next we calculate the angular power spectrum of the brightness temperature on the sky, resulting from density perturbations with a power spectrum $P_\delta(k)$,

$$\langle \delta_{\rm H}(\mathbf{k}_1)\delta_{\rm H}(\mathbf{k}_2)\rangle = (2\pi)^3 \delta^D(\mathbf{k}_1 + \mathbf{k}_2) P_\delta(k_1). \tag{163}$$

where $\delta_{\rm H}(\mathbf{k})$ is the Fourier tansform of the hydrogen density field, $\mathbf{k}$ is the comoving wavevector, and $\langle \cdots \rangle$ denotes an ensemble average (following the formalism described in [403]). The 21 cm brightness temperature observed at a frequency ν corresponding to a distance r along the line of sight, is given by

$$\delta T_{\rm b}(\mathbf{n}, \nu) = \int {\rm d}r W_\nu(r) \, \frac{{\rm d}T_{\rm b}}{{\rm d}\delta_{\rm H}} \delta_{\rm H}(\mathbf{n}, r), \tag{164}$$

where $\mathbf{n}$ denotes the direction of observation, $W_{\rm v}(r)$ is a narrow function of r that peaks at the distance corresponding to ν. The details of this function depend on the characteristics of the experiment. The brightness fluctuations in 164 can be expanded in spherical harmonics with expansion coefficients $a_{\rm lm}(\nu)$. The angular power spectrum of map $C_{\rm l}(\nu) = \langle |a_{\rm lm}(\nu)|^2 \rangle$ can be expressed in terms of the 3D power spectrum of fluctuations in the density $P_\delta(k)$,

$$C_{\rm l}(\nu) = 4\pi \int \frac{{\rm d}^3 k}{(2\pi)^3} P_\delta(k) \alpha_{\rm l}^2(k, \nu)$$

$$\alpha_{\rm l}(k, \nu) = \int {\rm d}r W_{\rm r_0}(r) \frac{{\rm d}T_{\rm b}}{{\rm d}\delta_{\rm H}}(r) j_l(kr). \tag{165}$$

Our calculation ignores inhomogeneities in the hydrogen ionization fraction, since they freeze at the earlier recombination epoch ($z \sim 10^3$) and so their amplitude is more than an order of magnitude smaller than $\delta_{\rm H}$ at $z \lesssim 100$.

The gravitational potential perturbations induce a redshift distortion effect that is of order $\sim (H/ck)^2$ smaller than $\delta_{\rm H}$ for the high–l modes of interest here.

Figure 53 shows the angular power spectrum at various redshifts. The signal peaks around $z \sim 50$ but maintains a substantial amplitude over the full range of $30 \lesssim z \lesssim 100$. The ability to probe the small scale power of density fluctuations is only limited by the Jeans scale, below which the dark matter inhomogeneities are washed out by the finite pressure of the gas. Interestingly, the cosmological Jeans mass reaches its minimum value, $\sim 3 \times 10^4$ $M_\odot$, within the redshift interval of interest here which corresponds to modes of angular scale $\sim$ arcsecond on the sky. During the epoch of reionization, photoionization heating raises the Jeans mass by several orders of magnitude and broadens spectral features, thus limiting the ability of other probes of the intergalactic medium, such as the Lyα forest, from accessing the same very low mass scales. The 21 cm tomography has the additional advantage of probing the majority of the cosmic gas, instead of the trace amount ($\sim 10^{-5}$) of neutral hydrogen probed by the Lyα forest after reionization. Similarly to the primary CMB anisotropies, the 21 cm signal is simply shaped by gravity, adiabatic cosmic expansion, and well-known atomic physics, and is not contaminated by complex astrophysical processes that affect the intergalactic medium at $z \lesssim 30$.

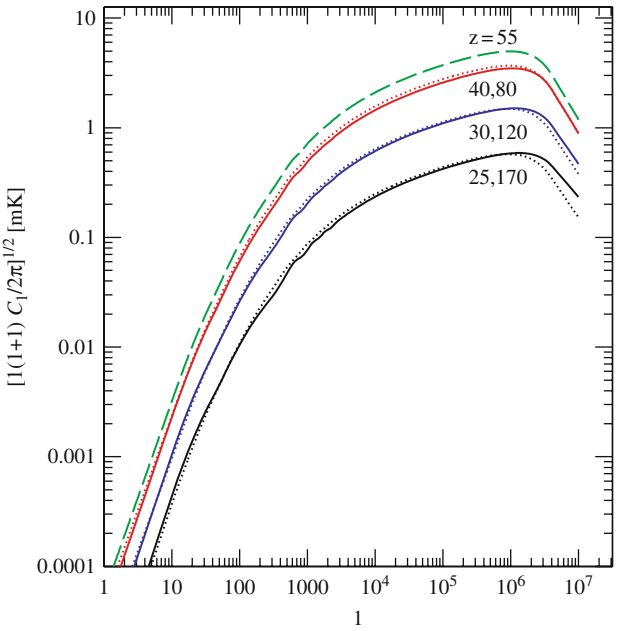

Fig. 53. Angular power spectrum of 21 cm anisotropies on the sky at various redshifts (from Loeb & Zaldarriaga 2004 [225]). From top to bottom, $z = 55, 40, 80, 30, 120, 25, 170$

Characterizing the initial fluctuations is one of the primary goals of observational cosmology, as it offers a window into the physics of the very early Universe, namely the epoch of inflation during which the fluctuations are believed to have been produced. In most models of inflation, the evolution of the Hubble parameter during inflation leads to departures from a scale-invariant spectrum that are of order $1/N_{\rm efold}$ with $N_{\rm efold} \sim 60$ being the number of e–folds between the time when the scale of our horizon was of order the horizon during inflation and the end of inflation [216]. Hints that the standard ΛCDM model may have too much power on galactic scales have inspired several proposals for suppressing the power on small scales. Examples include the possibility that the dark matter is warm and it decoupled while being relativistic so that its free streaming erased small-scale power [48], or direct modifications of inflation that produce a cut-off in the power on small scales [190]. An unavoidable collisionless component of the cosmic mass budget beyond CDM, is provided by massive neutrinos (see [196] for a review). Particle physics experiments established the mass splittings among different species which translate into a lower limit on the fraction of the dark matter accounted for by neutrinos of $f_\nu > 0.3\%$, while current constraints based on galaxies as tracers of the small scale power imply $f_\nu < 12\%$ [359].

Figure 54 shows the 21 cm power spectrum for various models that differ in their level of small scale power. It is clear that a precise measurement of the 21 cm power spectrum will dramatically improve current constraints on alternatives to the standard ΛCDM spectrum.

The 21 cm signal contains a wealth of information about the initial fluctuations. A full sky map at a single photon frequency measured up to $l_{\rm max}$, can probe the power spectrum up to $k_{\rm max} \sim (l_{\rm max}/10^4){\rm Mpc}^{-1}$. Such a map contains $l^2_{\rm max}$ independent samples. By shifting the photon frequency, one may obtain many independent measurements of the power. When measuring a mode l, which corresponds to a wavenumber $k \sim l/r$, two maps at different photon frequencies will be independent if they are separated in radial distance by $1/k$. Thus, an experiment that covers a spatial range Δr can probe a total of $k\Delta r \sim l\Delta r/r$ independent maps. An experiment that detects the 21 cm signal over a range $\Delta\nu$ centered on a frequency ν, is sensitive to $\Delta r/r \sim 0.5(\Delta\nu/\nu)(1+z)^{-1/2}$, and so it measures a total of $N_{21\,{\rm cm}} \sim 3\times 10^{16}(l_{\rm max}/10^6)^3(\Delta\nu/\nu)(z/100)^{-1/2}$ independent samples.

This detection capability cannot be reproduced even remotely by other techniques. For example, the primary CMB anisotropies are damped on small scales (through the so-called Silk damping), and probe only modes with $l \leqslant 3000 (k \leqslant 0.2\,{\rm Mpc}^{-1})$. The total number of modes available in the full sky is $N_{\rm cmb} = 2l^2_{\rm max} \sim 2\times 10^7 (l_{\rm max}/3000)^2$, including both temperature and polarization information.

The sensitivity of an experiment depends strongly on its particular design, involving the number and distribution of the antennae for an interferometer. Crudely speaking, the uncertainty in the measurement of $[l(l+1)C_{\rm l}/2\pi]^{1/2}$ is dominated by noise, N_ν, which is controlled by the sky brightness I_ν at the observed frequency ν [403],

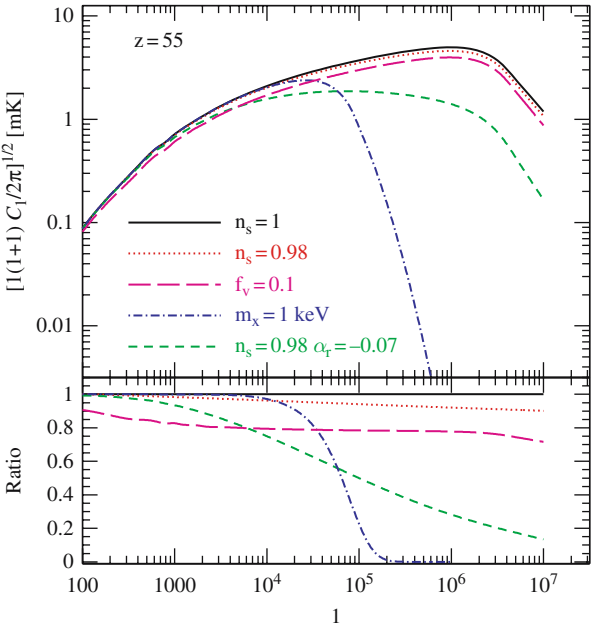

Fig. 54. Upper panel: Power spectrum of 21 cm anisotropies at $z = 55$ for a ΛCDM scale-invariant power spectrum, a model with $n = 0.98$, a model with $n = 0.98$ and $\alpha_r \equiv (\frac{1}{2})(d^2 \ln P/d \ln k^2) = -0.07$, a model of warm dark matter particles with a mass of 1 keV, and a model in which $f_\nu = 10\%$ of the matter density is in three species of massive neutrinos with a mass of 0.4 eV each. **Lower panel:** Ratios between the different power spectra and the scale-invariant spectrum (from Loeb & Zaldarriaga 2004 [225])

$$N_\nu \sim 0.4\,\mathrm{mK} \left(\frac{I_\nu[50\,\mathrm{MHz}]}{5 \times 10^5 \mathrm{Jy\ sr}^{-1}}\right) \left(\frac{l_{\min}}{35}\right) \left(\frac{5000}{l_{\max}}\right) \left(\frac{0.016}{f_{\mathrm{cover}}}\right) \times \left(\frac{1\ \mathrm{year}}{t_0}\right)^{1/2} \left(\frac{\Delta\nu}{50\mathrm{MHz}}\right)^{-1/2} \left(\frac{50\,\mathrm{MHz}}{\nu}\right)^{2.5}, \quad (166)$$

where $l_{\min}$ is the minimum observable l as determined by the field of view of the instruments, $l_{\max}$ is the maximum observable l as determined by the maximum separation of the antennae, f_{cover} is the fraction of the array area thats is covered by telescopes, t_0 is the observation time and $\Delta\nu$ is the frequency range over which the signal can be detected. Note that the assumed sky temperature of 0.7×10^4 K at $\nu = 50$ MHz (corresponding to $z \sim 30$) is more than six orders of magnitude larger than the signal. We have already included the fact that several independent maps can be produced by varying the observed frequency. The numbers adopted above are appropriate for the inner core of the *LOFAR* array (http://www.lofar.org), planned for initial operation in 2006. The predicted signal is ~ 1 mK, and so a year of integration or an

increase in the covering fraction are required to observe it with *LOFAR*. Other experiments whose goal is to detect 21 cm fluctuations from the subsequent epoch of reionization at $z \sim 6$–12 (when ionized bubbles exist and the fluctuations are larger) include the Mileura Wide-Field Array (MWA; Fig. 55 http://web.haystack.mit.edu/arrays/MWA/), the Primeval Structure Telescope (*PAST*; http://arxiv.org/abs/astro-ph/0502029), and in the more distant future the Square Kilometer Array (*SKA*; http://www.skatelescope.org). The main challenge in detecting the predicted signal from higher redshifts involves its appearance at low frequencies where the sky noise is high. Proposed space-based instruments [192] avoid the terrestrial radio noise and the increasing atmospheric opacity at $\nu < 20\,\mathrm{MHz}$ (corresponding to $z > 70$).

The 21 cm absorption is replaced by 21 cm emission from neutral hydrogen as soon as the intergalactic medium is heated above the CMB temperature by X-ray sources during the epoch of reionization [88]. This occurs long before reionization since the required heating requires only a modest amount of energy, $\sim 10^{-2}\,\mathrm{eV}[(1+z)/30]$, which is three orders of magnitude smaller than the amount necessary to ionize the Universe. As demonstrated by Chen & Miralda-Escude (2004) [88], heating due the recoil of atoms as they absorb Lyα photons [236] is not effective; the Lyα color temperature reaches equilibrium with the gas kinetic temperature and suppresses subsequent heating

Fig. 55. Prototype of the tile design for the *Mileura Wide-Field Array* (MWA) in western Australia, aimed at detecting redshifted 21 cm from the epoch of reionization. Each $4\mathrm{m} \times 4\mathrm{m}$ tile contains 16 dipole antennas operating in the frequency range of 80–300 MHz. Altogether the initial phase of MWA (the so-called "Low-Frequency Demostrator") will include 500 antenna tiles with a total collecting area of $8000\,\mathrm{m}^2$ at 150 MHz, scattered across a 1.5 km region and providing an angular resolution of a few arcminutes

before the level of heating becomes substantial. Once most of the cosmic hydrogen is reionized at $z_{\rm reion}$, the 21 cm signal is diminished. The optical depth for free-free absorption after reionization, $\sim 0.1[(1+z_{\rm reion})/20]^{5/2}$, modifies only slightly the expected 21 cm anisotropies. Gravitational lensing should modify the power spectrum [286] at high l, but can be separated as in standard CMB studies (see [325] and references therein). The 21 cm signal should be simpler to clean as it includes the same lensing foreground in independent maps obtained at different frequencies.

The large number of independent modes probed by the 21 cm signal would provide a measure of non-Gaussian deviations to a level of $\sim N_{21\,{\rm cm}}^{-1/2}$, constituting a test of the inflationary origin of the primordial inhomogeneities which are expected to possess deviations $\gtrsim 10^{-6}$ [244].

9.2 The Characteristic Observed Size of Ionized Bubbles at the End of Reionization

The first galaxies to appear in the Universe at redshifts $z \gtrsim 20$ created ionized bubbles in the intergalactic medium (IGM) of neutral hydrogen (HI) left over from the Big-Bang. It is thought that the ionized bubbles grew with time, surrounded clusters of dwarf galaxies[67, 143] and eventually overlapped quickly throughout the Universe over a narrow redshift interval near $z \sim 6$. This event signaled the end of the reionization epoch when the Universe was a billion years old. Measuring the unknown size distribution of the bubbles at their final overlap phase is a focus of forthcoming observational programs aimed at highly redshifted 21 cm emission from atomic hydrogen. In this subsection we follow Wyithe & Loeb (2004) [398] and show that the combined constraints of cosmic variance and causality imply an observed bubble size at the end of the overlap epoch of ~ 10 physical Mpc, and a scatter in the observed redshift of overlap along different lines-of-sight of ~ 0.15. This scatter is consistent with observational constraints from recent spectroscopic data on the farthest known quasars. This result implies that future radio experiments should be tuned to a characteristic angular scale of $\sim 0.5^\circ$ and have a minimum frequency band-width of ~ 8 MHz for an optimal detection of 21 cm flux fluctuations near the end of reionization.

During the reionization epoch, the characteristic bubble size (defined here as the spherically averaged mean radius of the H II regions that contain most of the ionized volume [143]) increased with time as smaller bubbles combined until their overlap completed and the diffuse IGM was reionized. However the largest size of isolated bubbles (fully surrounded by HI boundaries) that can be *observed* is finite, because of the combined phenomena of cosmic variance and causality. Figure 56 presents a schematic illustration of the geometry. There is a surface on the sky corresponding to the time along different lines-of-sight when the diffuse (uncollapsed) IGM was *most recently neutral*. We refer to it as the Surface of Bubble Overlap (SBO). There are two competing

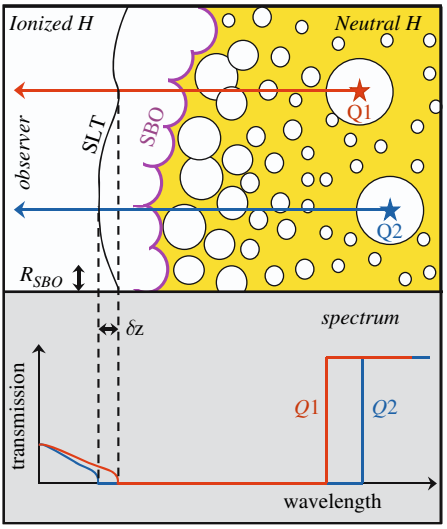

Fig. 56. The distances to the observed Surface of Bubble Overlap (SBO) and Surface of Lyα Transmission (SLT) fluctuate on the sky. The SBO corresponds to the first region of diffuse neutral IGM *observed* along a random line-of-sight. It fluctuates across a shell with a minimum width dictated by the condition that the light crossing time across the characteristic radius $R_{\rm SBO}$ of ionized bubbles equals the cosmic scatter in their formation times. Thus, *causality* and *cosmic variance* determine the characteristic scale of bubbles at the completion of bubble overlap. After some time delay the IGM becomes transparent to Lyα photons, resulting in a second surface, the SLT. The upper panel illustrates how the lines-of-sight towards two quasars ($Q1$ in red and $Q2$ in blue) intersect the SLT with a redshift difference δz. The resulting variation in the observed spectrum of the two quasars is shown in the lower panel. Observationally, the ensemble of redshifts down to which the Gunn-Peterson troughs are seen in the spectra of $z > 6.1$ quasars is drawn from the probability distribution $\mathrm{d}P/\mathrm{d}z_{\rm SLT}$ for the redshift at which the IGM started to allow Lyα transmission along random lines-of-sight. The observed values of $z_{\rm SLT}$ show a small scatter Fan et al. (2004) [127] in the SLT redshift around an average value of $\langle z_{\rm SLT}\rangle \approx 5.95$. Some regions of the IGM may have also become transparent to Lyα photons prior to overlap, resulting in windows of transmission inside the Gunn-Peterson trough (one such region may have been seen White et al. (2003) [380] in SDSS J1148+5251). In the existing examples, the portions of the Universe probed by the lower end of the Gunn-Peterson trough are located several hundred comoving Mpc away from the background quasar, and are therefore not correlated with the quasar host galaxy. The distribution $\mathrm{d}P/\mathrm{d}z_{\rm SLT}$ is also independent of the redshift distribution of the quasars. Moreover, lines-of-sight to these quasars are not causally connected at $z \sim 6$ and may be considered independent. (From Wyithe & Loeb 2004 [398])

sources for fluctuations in the SBO, each of which is dependent on the characteristic size, $R_{\rm SBO}$, of the ionized regions just before the final overlap. First, the finite speed of light implies that 21 cm photons observed from different

points along the curved boundary of an H II region must have been emitted at different times during the history of the Universe. Second, bubbles on a comoving scale R achieve reionization over a spread of redshifts due to cosmic variance in the initial conditions of the density field smoothed on that scale. The characteristic scale of H II bubbles grows with time, leading to a decline in the spread of their formation redshifts [67] as the cosmic variance is averaged over an increasing spatial volume. However the 21 cm light-travel time across a bubble rises concurrently. Suppose a signal 21 cm photon which encodes the presence of neutral gas, is emitted from the far edge of the ionizing bubble. If the adjacent region along the line-of-sight has not become ionized by the time this photon reaches the near side of the bubble, then the photon will encounter diffuse neutral gas. Other photons emitted at this lower redshift will therefore also encode the presence of diffuse neutral gas, implying that the first photon was emitted prior to overlap, and not from the SBO. Hence the largest observable scale of H II regions when their overlap completes, corresponds to the first epoch at which the light crossing time becomes larger than the spread in formation times of ionized regions. Only then will the signal photon leaving the far side of the H II region have the lowest redshift of any signal photon along that line-of-sight.

The observed spectra of some quasars (see Fig. 57) beyond $z \sim 6.1$ show a Gunn-Peterson trough [163, 127] (Fan et al. 2005 [128]), a blank spectral region at wavelengths shorter than Lyα at the quasar redshift, implying the presence of HI in the diffuse IGM. The detection of Gunn-Peterson troughs indicates a rapid change [126, 287, 380] in the neutral content of the IGM at $z \sim 6$, and hence a rapid change in the intensity of the background ionizing flux. This rapid change implies that overlap, and hence the reionization epoch, concluded near $z \sim 6$. The most promising observational probe[403, 258] of the reionization epoch is redshifted 21 cm emission from intergalactic HI . Future observations using low frequency radio arrays (e.g. LOFAR, MWA, and PAST) will allow a direct determination of the topology and duration of the phase of bubble overlap. In this section we determine the expected angular scale and redshift width of the 21 cm fluctuations at the SBO theoretically, and show that this determination is consistent with current observational constraints.

We start by quantifying the constraints of causality and cosmic variance. First suppose we have an HII region with a physical radius $R/(1 + \langle z \rangle)$. For a 21 cm photon, the light crossing time of this radius is

$$\langle \Delta z^2 \rangle^{1/2} = \left| \frac{\mathrm{d}z}{\mathrm{d}t} \right|_{\langle z \rangle} \frac{R}{c(1 + \langle z \rangle)}, \tag{167}$$

where at the high-redshifts of interest $(\mathrm{d}z/\mathrm{d}t) = -(H_0\sqrt{\Omega_m})(1+z)^{5/2}$. Here, c is the speed of light, H_0 is the present-day Hubble constant, Ω_m is the present day matter density parameter, and $\langle z \rangle$ is the mean redshift of the SBO. Note that when discussing this crossing time, we are referring to photons used to probe the ionized bubble (e.g. at 21 cm), rather than photons involved in the dynamics of the bubble evolution (Fig. 58).

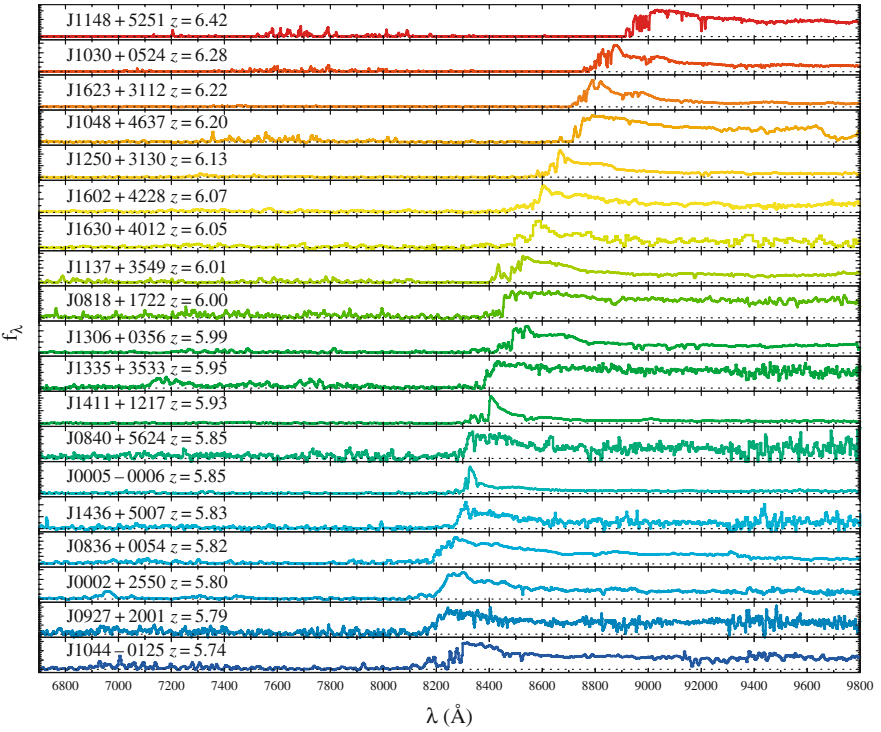

Fig. 57. Spectra of 19 quasars with redshifts $5.74 < z < 6.42$ from the *Sloan Digital Sky Survey* Far et al. (2005) [128]. For some of the highest-redshift quasars, the spectrum shows no transmitted flux shortward of the Lyα wavelength at the quasar redshift (the so-called "Gunn-Peterson trough"), indicating a non-negligible neutral fraction in the IGM (see the analysis of Fan et al. 2005 [128] for details)

Second, overlap would have occurred at different times in different regions of the IGM due to the cosmic scatter in the process of structure formation within finite spatial volumes [67]. Reionization should be completed within a region of comoving radius R when the fraction of mass incorporated into collapsed objects in this region attains a certain critical value, corresponding to a threshold number of ionizing photons emitted per baryon. The ionization state of a region is governed by the enclosed ionizing luminosity, by its over-density, and by dense pockets of neutral gas that are self shielding to ionizing radiation. There is an offset [67] δz between the redshift when a region of mean over-density $\bar{\delta}_{\rm R}$ achieves this critical collapsed fraction, and the redshift $\bar{z}$ when the Universe achieves the same collapsed fraction on average. This offset may be computed [67] from the expression for the collapsed fraction [52] $F_{\rm col}$ within a region of over-density $\bar{\delta}_{\rm R}$ on a comoving scale R,

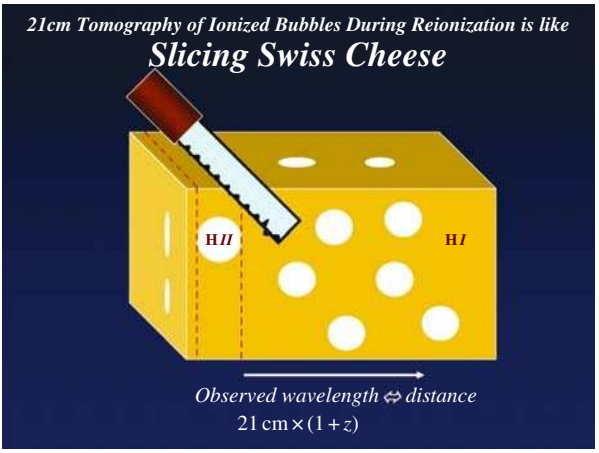

Fig. 58. 21 cm imaging of ionized bubbles during the epoch of reionization is analogous to slicing swiss cheese. The technique of slicing at intervals separated by the typical dimension of a bubble is optimal for revealing different pattens in each slice. (From Loeb 2007 [218])

$$F_{\rm col}(M_{\rm min}) = {\rm erfc}\left[\frac{\delta_{\rm c}-\bar{\delta}_{\rm R}}{\sqrt{2[\sigma^2_{\rm R_{min}}-\sigma^2_{\rm R}]}}\right] \rightarrow \frac{\delta z}{(1+\bar{z})} = \frac{\bar{\delta}_{\rm R}}{\delta_{\rm c}(\bar{z})} - \left[1-\sqrt{1-\frac{\sigma^2_{\rm R}}{\sigma^2_{\rm R_{min}}}}\right], \tag{168}$$

where $\delta_{\rm c}(\bar{z}) \propto (1+\bar{z})$ is the collapse threshold for an over-density at a redshift $\bar{z}$; $\sigma_{\rm R}$ and $\sigma_{R_{\rm min}}$ are the variances in the power-spectrum linearly extrapolated to $z = 0$ on comoving scales corresponding to the region of interest and to the minimum galaxy mass $M_{\rm min}$, respectively. The offset in the ionization redshift of a region depends on its linear over-density, $\bar{\delta}_{\rm R}$. As a result, the distribution of offsets, and therefore the scatter in the SBO may be obtained directly from the power spectrum of primordial inhomogeneities. As can be seen from (168), larger regions have a smaller scatter due to their smaller cosmic variance.

Note that (168) is independent of the critical value of the collapsed fraction required for reionization. Moreover, our numerical constraints are very weakly dependent on the minimum galaxy mass, which we choose to have a virial temperature of 10^4 K corresponding to the cooling threshold of primordial atomic gas. The growth of an H II bubble around a cluster of sources requires that the mean-free-path of ionizing photons be of order the bubble radius or larger. Since ionizing photons can be absorbed by dense pockets of neutral gas inside the H II region, the necessary increase in the mean-free-path with time implies that the critical collapsed fraction required to ionize a region of size R increases as well. This larger collapsed fraction affects the redshift at which the region becomes ionized, but not the scatter in redshifts from place to place which is the focus of this sub-section. Our results are therefore independent of assumptions about unknown quantities such as the star formation efficiency

and the escape fraction of ionizing photons from galaxies, as well as unknown processes of feedback in galaxies and clumping of the IGM.

Figure 59 displays the above two fundamental constraints. The causality constraint (167) is shown as the blue line, giving a longer crossing time for a larger bubble size. This contrasts with the constraint of cosmic variance (168), indicated by the red line, which shows how the scatter in formation times decreases with increasing bubble size. The scatter in the SBO redshift and the corresponding fluctuation scale of the SBO are given by the intersection of these curves. We find that the thickness of the SBO is $\langle \Delta z^2 \rangle^{1/2} \sim 0.13$, and that the bubbles which form the SBO have a characteristic comoving size of $\sim 60\,\mathrm{Mpc}$ (equivalent to 8.6 physical Mpc). At $z \sim 6$ this size corresponds to angular scales of $\theta_{\rm SBO} \sim 0.4$ degrees on the sky.

A scatter of ~ 0.15 in the SBO is somewhat larger than the value extracted from existing numerical simulations [152, 401]. The difference is most likely due to the limited size of the simulated volumes; while the simulations appropriately describe the reionization process within limited regions of the Universe, they are not sufficiently large to describe the global properties of the overlap phase [67]. The scales over which cosmological radiative transfer has been simulated are smaller than the characteristic extent of the SBO, which we find to be $R_{\rm SBO} \sim 70$ comoving Mpc.

We can constrain the scatter in the SBO redshift observationally using the spectra of the highest redshift quasars. Since only a trace amount of neutral hydrogen is needed to absorb Lyα photons, the time where the IGM becomes Lyα transparent need not coincide with bubble overlap. Following overlap the IGM was exposed to ionizing sources in all directions and the ionizing intensity rose rapidly. After some time the ionizing background flux was sufficiently high that the HI fraction fell to a level at which the IGM allowed transmission of resonant Lyα photons. This is shown schematically in Fig. 56. The lower wavelength limit of the Gunn-Peterson trough corresponds to the Lyα wavelength at the redshift when the IGM started to allow transmission of Lyα photons *along that particular line-of-sight*. In addition to the SBO we therefore also define the Surface of Lyα Transmission (hereafter SLT) as the redshift along different lines-of-sight when the diffuse IGM became transparent to Lyα photons.

The scatter in the SLT redshift is an observable which we would like to compare with the scatter in the SBO redshift. The variance of the density field on large scales results in the biased clustering of sources [67]. H II regions grow in size around these clusters of sources. In order for the ionizing photons produced by a cluster to advance the walls of the ionized bubble around it, the mean-free-path of these photons must be of order the bubble size or larger. After bubble overlap, the ionizing intensity at any point grows until the ionizing photons have time to travel across the scale of the new mean-free-path, which represents the horizon out to which ionizing sources are visible. Since the mean-free-path is larger than $R_{\rm SBO}$, the ionizing intensity at the SLT averages the cosmic scatter over a larger volume than at the SBO. This

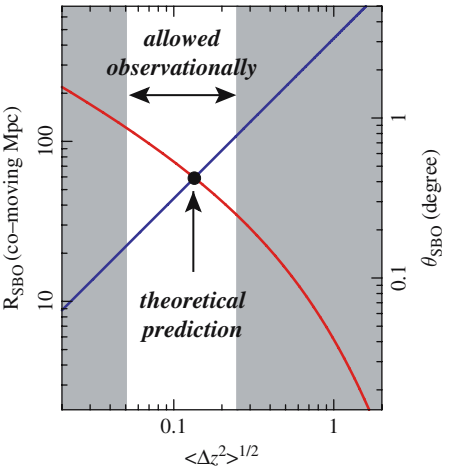

Fig. 59. Constraints on the scatter in the SBO redshift and the characteristic size of isolated bubbles at the final overlap stage, $R_{\rm SBO}$ (see Fig. 1). The characteristic size of H II regions grows with time. The SBO is observed for the bubble scale at which the light crossing time (*blue line*) first becomes smaller than the cosmic scatter in bubble formation times (*red line*). At $z \sim 6$, the implied scale $R_{\rm SBO} \sim 60$ comoving Mpc (or ~ 8.6 physical Mpc), corresponds to a characteristic angular radius of $\theta_{\rm SLT} \sim 0.4$ degrees on the sky. After bubble overlap, the ionizing intensity grows to a level at which the IGM becomes transparent to Lyα photons. The collapsed fraction required for Lyα transmission within a region of a certain size will be larger than required for its ionization. However, the scatter in (168) is not sensitive to the collapsed fraction, and so may be used for both the SBO and SLT. The scatter in the SLT is smaller than the cosmic scatter in the structure formation time on the scale of the mean-free-path for ionizing photons. This mean-free-path must be longer than $R_{\rm SBO} \sim 60$ Mpc, an inference which is supported by analysis of the Lyα forest at $z \sim 4$ where the mean-free-path is estimated Miralda-Escuclé (2003) [256] to be $\sim$120 comoving Mpc at the Lyman limit (and longer at higher frequencies). If it is dominated by cosmic variance, then the scatter in the SLT redshift provides a lower limit to the SBO scatter. The three known quasars at $z > 6.1$ have Lyα transmission redshifts of White et al. (2003) [380], Far et al. (2004) [127] $z_{\rm SLT} = 5.9$, 5.95 and 5.98, implying that the scatter in the SBO must be $\gtrsim 0.05$ (this scatter may become better known from follow-up spectroscopy of Gamma Ray Burst afterglows at $z > 6$ that might be discovered by the *SWIFT* satellite Barkana & Loeb (2004) [26], Bromm & Loeb (2002) [61]). The observed scatter in the SLT redshift is somewhat smaller than the predicted SBO scatter, confirming the expectation that cosmic variance is smaller at the SLT. The scatter in the SBO redshift must also be $\lesssim 0.25$ because the lines-of-sight to the two highest redshift quasars have a redshift of Lyα transparency at $z \sim 6$, but a neutral fraction that is known from the *proximity effect* Wyithe & Loeb (2004) [395] to be substantial at $z > 6.2$–6.3. The excluded regions of scatter for the SBO are shown in gray. (From Whithe & Loeb (2004) [398])

constraint implies that the cosmic variance in the SLT redshift must be smaller than the scatter in the SBO redshift. However, it is possible that opacity from small-scale structure contributes additional scatter to the SLT redshift.

If cosmic variance dominates the observed scatter in the SLT redshift, then based on the spectra of the three $z > 6.1$ quasars[127, 380] we would expect the scatter in the SBO redshift to satisfy $\langle \Delta z^2 \rangle_{\rm obs}^{1/2} \gtrsim 0.05$. In addition, analysis of the *proximity effect* for the size of the H II regions around the two highest redshift quasars[395, 250] implies a neutral fraction that is of order unity (i.e. pre-overlap) at $z \sim 6.2$–6.3, while the transmission of Lyα photons at $z \lesssim 6$ implies that overlap must have completed by that time. This restricts the scatter in the SBO to be $\langle \Delta z^2 \rangle_{\rm obs}^{1/2} \lesssim 0.25$. The constraints on values for the scatter in the SBO redshift are shaded gray in Fig. 59. It is reassuring that the theoretical prediction for the SBO scatter of $\langle \Delta z^2 \rangle_{\rm obs}^{1/2} \sim 0.15$, with a characteristic scale of ~ 70 comoving Mpc, is bounded by these constraints.

The possible presence of a significantly neutral IGM just beyond the redshift of overlap[395, 250] is encouraging for upcoming 21 cm studies of the reionization epoch as it results in emission near an observed frequency of 200 MHz where the signal is most readily detectable. Future observations of redshifted 21 cm line emission at $6 \lesssim z \lesssim 6.5$ with instruments such as LOFAR, MWA, and PAST, will be able to map the three-dimensional distribution of HI at the end of reionization. The intergalactic H II regions will imprint a "knee" in the power-spectrum of the 21 cm anisotropies on a characteristic angular scale corresponding to a typical isolated H II region[403]. Our results suggest that this characteristic angular scale is large at the end of reionization, $\theta_{\rm SBO} \sim 0.5$ degrees, motivating the construction of compact low frequency arrays. An SBO thickness of $\langle \Delta z^2 \rangle^{1/2} \sim 0.15$ suggests a minimum frequency band-width of ~ 8 MHz for experiments aiming to detect anisotropies in 21 cm emission just prior to overlap. These results will help guide the design of the next generation of low-frequency radio observatories in the search for 21 cm emission at the end of the reionization epoch.

The full size distribution of ionized bubbles has to be calculated from a numerical cosmological simulation that includes gas dynamics and radiative transfer. The simulation box needs to be sufficiently large for it to sample an unbiased volume of the Universe with little cosmic variance, but at the same time one must resolve the scale of individual dwarf galaxies which provide (as well as consume) ionizing photons (see discussion at the last section of this review). Until a reliable simulation of this magnitude exists, one must adopt an approximate analytic approach to estimate the bubble size distribution. Below we describe an example for such a method, developed by Furlanetto et al. (2004) [143].

The criterion for a region to be ionized is that galaxies inside of it produce a sufficient number of ionizing photons per baryon. This condition can be translated to the requirement that the collapsed fraction of mass in halos above some threshold mass $M_{\rm min}$ will exceed some threshold, namely $F_{\rm col} > \zeta^{-1}$.

The minimum halo mass most likely corresponds to a virial temperature of 10^4 K relating to the threshold for atomic cooling (assuming that molecular hydrogen cooling is suppressed by the UV background in the Lyman-Werner band). We would like to find the largest region around every point that satisfies the above condition on the collapse fraction and then calculate the abundance of ionized regions of this size. Different regions have different values of $F_{\rm col}$ because their mean density is different. In the extended Press-Schechter model (Bond et al. 1991 [52]; Lacey & Cole 1993 [210]), the collapse fraction in a region of mean overdensity $\delta_{\rm M}$ is

$$F_{\rm col} = {\rm erfc}\left(\frac{\delta_{\rm c} - \delta_{\rm M}}{\sqrt{2[\sigma^2_{\rm min} - \sigma^2(M,z)]}}\right). \tag{169}$$

where $\sigma^2(M,z)$ is the variance of density fluctuations on mass scale M, $\sigma^2_{\rm min} \equiv \sigma^2(M_{\rm min},z)$, and $\delta_{\rm c}$ is the collapse threshold. This equation can be used to derive the condition on the mean overdensity within a region of mass M in order for it to be ionized,

$$\delta_{\rm M} > \delta_{\rm B}(M,z) \equiv \delta_{\rm c} - \sqrt{2}K(\zeta)[\sigma^2_{\rm min} - \sigma^2(M,z)]^{1/2}, \tag{170}$$

where $K(\zeta) = {\rm erfc}^{-1}(1-\zeta^{-1})$. Furlanetto et al. [143] showed how to construct the mass function of ionized regions from $\delta_{\rm B}$ in analogy with the halo mass function (Press & Schechter 1974 [290]; Bond et al. 1991 [52]). The barrier in (170) is well approximated by a linear dependence on σ^2,

$$\delta_{\rm B} \approx B(M) = B_0 + B_1\sigma^2(M), \tag{171}$$

in which case the mass function has an analytic solution (Sheth 1998 [331]),

$$n(M) = \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}}{M^2}\left|\frac{{\rm d}\ln\sigma}{{\rm d}\ln M}\right|\frac{B_0}{\sigma(M)}\exp\left[-\frac{B^2(M)}{2\sigma^2(M)}\right], \tag{172}$$

where $\bar{\rho}$ is the mean mass density. This solution provides the comoving number density of ionized bubbles with mass in the range of $(M, M+{\rm d}M)$. The main difference of this result from the Press-Schechter mass function is that the barrier in this case becomes more difficult to cross on smaller scales because $\delta_{\rm B}$ is a decreasing function of mass M. This gives bubbles a characteristic size. The size evolves with redshift in a way that depends only on ζ and $M_{\rm min}$.

One limitation of the above analytic model is that it ignores the non-local influence of sources on distant regions (such as voids) as well as the possible shadowing effect of intervening gas. Radiative transfer effects in the real Universe are inherently three-dimensional and cannot be fully captured by spherical averages as done in this model. Moreover, the value of $M_{\rm min}$ is expected to increase in regions that were already ionized, complicating the expectation of whether they will remain ionized later. The history of reionization could be complicated and non monotonic in individual regions, as described

by Furlanetto & Loeb (2005) [144]. Finally, the above analytic formalism does not take the light propagation delay into account as we have done above in estimating the characteristic bubble size at the end of reionization. Hence this formalism describes the observed bubbles only as long as the characteristic bubble size is sufficiently small, so that the light propagation delay can be neglected compared to cosmic variance. The general effect of the light propagation delay on the power-spectrum of 21 cm fluctuations was quantified by Barkana & Loeb (2005) [29].

9.3 Separating the "Physics" from the "Astrophysics" of the Reionization Epoch with 21 cm Fluctuations

The 21 cm signal can be seen from epochs during which the cosmic gas was largely neutral and deviated from thermal equilibrium with the cosmic microwave background (CMB). The signal vanished at redshifts $z \gtrsim 200$, when the residual fraction of free electrons after cosmological recombination kept the gas kinetic temperature, $T_{\rm k}$, close to the CMB temperature, T_γ (see Fig. 52). But during $200 \gtrsim z \gtrsim 30$ the gas cooled adiabatically and atomic collisions kept the spin temperature of the hyperfine level population below T_γ, so that the gas appeared in absorption [322, 225]. As the Hubble expansion continued to rarefy the gas, radiative coupling of $T_{\rm s}$ to T_γ began to dominate and the 21 cm signal faded. When the first galaxies formed, the UV photons they produced between the Lyα and Lyman limit wavelengths propagated freely through the Universe, redshifted into the Lyα resonance, and coupled $T_{\rm s}$ and $T_{\rm k}$ once again through the Wouthuysen-Field [387, 131] effect by which the two hyperfine states are mixed through the absorption and re-emission of a Lyα photon [236, 96]. Emission above the Lyman limit by the same galaxies initiated the process of reionization by creating ionized bubbles in the neutral cosmic gas, while X-ray photons propagated farther and heated $T_{\rm k}$ above T_γ throughout the Universe. Once $T_{\rm s}$ grew larger than T_γ, the gas appeared in 21 cm emission. The ionized bubbles imprinted a knee in the power spectrum of 21 cm fluctuations [403], which traced the H I topology until the process of reionization was completed [143].

The various effects that determine the 21 cm fluctuations can be separated into two classes. The density power spectrum probes basic cosmological parameters and inflationary initial conditions, and can be calculated exactly in linear theory. However, the radiation from galaxies, both Lyα radiation and ionizing photons, involves the complex, non-linear physics of galaxy formation and star formation. If only the sum of all fluctuations could be measured, then it would be difficult to extract the separate sources, and in particular, the extraction of the power spectrum would be subject to systematic errors involving the properties of galaxies. Barkana & Loeb (2005) [28] showed that the unique three-dimensional properties of 21 cm measurements permit a separation of these distinct effects. Thus, 21 cm fluctuations can probe astrophysical (radiative) sources associated with the first galaxies, while at the same time

separately probing the physical (inflationary) initial conditions of the Universe. In order to affect this separation most easily, it is necessary to measure the three-dimensional power spectrum of 21 cm fluctuations. The discussion in this section follows Barkana & Loeb (2005) [28].

Spin Temperature History

As long as the spin-temperature $T_{\rm s}$ is smaller than the CMB temperature $T_\gamma = 2.725(1+z)\,{\rm K}$, hydrogen atoms absorb the CMB, whereas if $T_{\rm s} > T_\gamma$ they emit excess flux. In general, the resonant 21 cm interaction changes the brightness temperature of the CMB by [322, 236] $T_{\rm b} = \tau\,(T_{\rm s} - T_\gamma)\,/(1+z)$, where the optical depth at a wavelength $\lambda = 21\,{\rm cm}$ is

$$\tau = \frac{3c\lambda^2 h A_{10} n_{\rm H}}{32\pi k T_{\rm s}\,(1+z)\,({\rm d}v_{\rm r}/{\rm d}r)} x_{\rm HI}\ , \qquad (173)$$

where $n_{\rm H}$ is the number density of hydrogen, $A_{10} = 2.85\times10^{-15}\,{\rm s}^{-1}$ is the spontaneous emission coefficient, $x_{\rm HI}$ is the neutral hydrogen fraction, and ${\rm d}v_{\rm r}/{\rm d}r$ is the gradient of the radial velocity along the line of sight with $v_{\rm r}$ being the physical radial velocity and r the comoving distance; on average ${\rm d}v_{\rm r}/{\rm d}r = H(z)/(1+z)$ where H is the Hubble parameter. The velocity gradient term arises because it dictates the path length over which a 21 cm photon resonates with atoms before it is shifted out of resonance by the Doppler effect [340].

For the concordance set of cosmological parameters [347], the mean brightness temperature on the sky at redshift z is

$$T_{\rm b} = 28\,{\rm mK}\left(\frac{\Omega_{\rm b}h}{0.033}\right)\left(\frac{\Omega_{\rm m}}{0.27}\right)^{-\frac{1}{2}}\left[\frac{(1+z)}{10}\right]^{1/2}\left[\frac{(T_{\rm s}-T_\gamma)}{T_{\rm s}}\right]\bar{x}_{\rm HI}, \qquad (174)$$

where $\bar{x}_{\rm HI}$ is the mean neutral fraction of hydrogen. The spin temperature itself is coupled to $T_{\rm k}$ through the spin-flip transition, which can be excited by collisions or by the absorption of Lyα photons. As a result, the combination that appears in $T_{\rm b}$ becomes [131] $(T_{\rm s}-T_\gamma)/T_{\rm s} = [x_{\rm tot}/(1+x_{\rm tot})]\,(1-T_\gamma/T_{\rm k})$, where $x_{\rm tot} = x_\alpha + x_{\rm c}$ is the sum of the radiative and collisional threshold parameters. These parameters are $x_\alpha = 4P_\alpha T_\star/27A_{10}T_\gamma$ and $x_{\rm c} = 4\kappa_{1-0}(T_{\rm k})\,n_{\rm H}T_\star/3A_{10}T_\gamma$, where P_α is the Lyα scattering rate which is proportional to the Lyα intensity, and κ_{1-0} is tabulated as a function of $T_{\rm k}$ [11, 405]. The coupling of the spin temperature to the gas temperature becomes substantial when $x_{\rm tot} \gtrsim 1$.

Brightness Temperature Fluctuations

Although the mean 21 cm emission or absorption is difficult to measure due to bright foregrounds, the unique character of the fluctuations in $T_{\rm b}$ allows for a much easier extraction of the signal [154, 403, 258, 259, 313]. We adopt the

notation $\delta_{\rm A}$ for the fractional fluctuation in quantity A (with a lone δ denoting density perturbations). In general, the fluctuations in $T_{\rm b}$ can be sourced by fluctuations in gas density (δ), Lyα flux (through δ_{x_α}) neutral fraction ($\delta_{x_{\rm HI}}$), radial velocity gradient ($\delta_{d_{\rm r}v_{\rm r}}$), and temperature, so we find

$$\delta_{T_b} = \left(1+\frac{x_c}{\tilde{x}_{\rm tot}}\right)\delta + \frac{x_\alpha}{\tilde{x}_{\rm tot}}\delta_{x_\alpha} + \delta_{x_{\rm HI}} - \delta_{d_r v_r} + (\gamma_a - 1)\left[\frac{T_\gamma}{T_k - T_\gamma} + \frac{x_{\rm c}}{\tilde{x}_{\rm tot}}\frac{d\log(\kappa_{1-0})}{d\log(T_k)}\right]\delta \ , \quad (175)$$

where the adiabatic index is $\gamma_a = 1 + (\delta_{T_{\rm k}}/\delta)$, and we define $\tilde{x}_{\rm tot} \equiv (1 + x_{\rm tot})x_{\rm tot}$. Taking the Fourier transform, we obtain the power spectrum of each quantity; e.g., the total power spectrum $P_{T_{\rm b}}$ is defined by

$$\langle \tilde{\delta}_{T_{\rm b}}(\mathbf{k_1})\tilde{\delta}_{\mathbf{T}_{\rm b}}(\mathbf{k_2})\rangle = (\mathbf{2}\pi)^{\mathbf{3}}\delta^{\rm D}(\mathbf{k_1}+\mathbf{k_2})\mathbf{P}_{\mathbf{T}_{\rm b}}(\mathbf{k_1}) \ , \quad (176)$$

where $\tilde{\delta}_{T_{\rm b}}(\mathbf{k})$ is the Fourier transform of $\delta_{T_{\rm b}}$, $\mathbf{k}$ is the comoving wavevector, $\delta^{\rm D}$ is the Dirac delta function, and $\langle\cdots\rangle$ denotes an ensemble average. In this analysis, we consider scales much bigger than the characteristic bubble size and the early phase of reionization (when $\bar{\delta_{x_{\rm HI}}} << 1$), so that the fluctuations $\delta_{x_{\rm HI}}$ are also much smaller than unity. For a more general treatment, see McQuinn et al. (2005) [249].

The Separation of Powers

The fluctuation $\delta_{T_{\rm b}}$ consists of a number of isotropic sources of fluctuations plus the peculiar velocity term $-\delta_{d_{\rm r}v_{\rm r}}$. Its Fourier transform is simply proportional to that of the density field [189, 41],

$$\tilde{\delta}_{d_{\rm r}v_{\rm r}} = -\mu^2\tilde{\delta}, \quad (177)$$

where $\mu = \cos\theta_{\rm k}$ in terms of the angle $\theta_{\rm k}$ of $\mathbf{k}$ with respect to the line of sight. The μ^2 dependence in this equation results from taking the radial (i.e., line-of-sight) component ($\propto \mu$) of the peculiar velocity, and then the radial component ($\propto \mu$) of its gradient. Intuitively, a high-density region possesses a velocity infall towards the density peak, implying that a photon must travel further from the peak in order to reach a fixed relative redshift, compared with the case of pure Hubble expansion. Thus the optical depth is always increased by this effect in regions with $\delta > 0$. This phenomenon is most properly termed *velocity compression*.

We therefore write the fluctuation in Fourier space as

$$\tilde{\delta}_{T_{\rm b}}(\mathbf{k}) = \mu^{\mathbf{2}}\tilde{\delta}(\mathbf{k}) + \beta\tilde{\delta}(\mathbf{k}) + \tilde{\delta}_{\rm rad}(\mathbf{k}) \ , \quad (178)$$

where we have defined a coefficient β by collecting all terms $\propto \delta$ in (175), and have also combined the terms that depend on the radiation fields of Lyα photons and ionizing photons, respectively. We assume that these radiation fields

produce isotropic power spectra, since the physical processes that determine them have no preferred direction in space. The total power spectrum is

$$P_{T_{\rm b}}(\mathbf{k}) = \mu^4 P_\delta(k) + 2\mu^2[\beta P_\delta(k) + P_{\delta\cdot{\rm rad}}(k)] \\ +[\beta^2 P_\delta(k) + P_{\rm rad}(k) + 2\beta P_{\delta\cdot{\rm rad}}(k)] \ , \quad (179)$$

where we have defined the power spectrum $P_{\delta\cdot{\rm rad}}$ as the Fourier transform of the cross-correlation function,

$$\xi_{\delta\cdot{\rm rad}}(r) = \langle \delta(\mathbf{r_1})\, \delta_{\rm rad}(\mathbf{r_1} + \mathbf{r}) \rangle \ . \quad (180)$$

We note that a similar anisotropy in the power spectrum has been previously derived in a different context, i.e., where the use of galaxy redshifts to estimate distances changes the apparent line-of-sight density of galaxies in redshift surveys [189, 217, 176, 133]. However, galaxies are intrinsically complex tracers of the underlying density field, and in that case there is no analog to the method that we demonstrate below for separating in 21 cm fluctuations the effect of initial conditions from that of later astrophysical processes.

The velocity gradient term has also been examined for its global effect on the sky-averaged power and on radio visibilities [365, 41]. The other sources of 21 cm perturbations are isotropic and would produce a power spectrum $P_{T_{\rm b}}(k)$ that could be measured by averaging the power over spherical shells in $\mathbf{k}$ space. In the simple case where $\beta = 1$ and only the density and velocity terms contribute, the velocity term increases the total power by a factor of $\langle (1+\mu^2)^2 \rangle = 1.87$ in the spherical average. However, instead of averaging the signal, we can use the angular structure of the power spectrum to greatly increase the discriminatory power of 21 cm observations. We may break up each spherical shell in $\mathbf{k}$ space into rings of constant μ and construct the observed $P_{T_{\rm b}}(k, \mu)$. Considering (179) as a polynomial in μ, i.e., $\mu^4 P_{\mu^4} + \mu^2 P_{\mu^2} + P_{\mu^0}$, we see that the power at just three values of μ is required in order to separate out the coefficients of 1, μ^2, and μ^4 for each k.

If the velocity compression were not present, then only the μ-independent term (times $T_{\rm b}^2$) would have been observed, and its separation into the five components ($T_{\rm b}$, β, and three power spectra) would have been difficult and subject to degeneracies. Once the power has been separated into three parts, however, the μ^4 coefficient can be used to measure the density power spectrum directly, with no interference from any other source of fluctuations. Since the overall amplitude of the power spectrum, and its scaling with redshift, are well determined from the combination of the CMB temperature fluctuations and galaxy surveys, the amplitude of P_{μ^4} directly determines the mean brightness temperature $T_{\rm b}$ on the sky, which measures a combination of $T_{\rm s}$ and $\bar{x}_{\rm HI}$ at the observed redshift. McQuinn et al. (2005) [249] analysed in detail the parameters that can be constrained by upcoming 21 cm experiments in concert with future CMB experiments such as Planck (http://www.rssd.esa.int/index.php?project=PLANCK). Once $P_\delta(k)$ has been determined, the coefficients of the μ^2 term and the μ-independent

term must be used to determine the remaining unknowns, β, $P_{\delta\cdot\mathrm{rad}}(k)$, and $P_{\mathrm{rad}}(k)$. Since the coefficient β is independent of k, determining it and thus breaking the last remaining degeneracy requires only a weak additional assumption on the behavior of the power spectra, such as their asymptotic behavior at large or small scales. If the measurements cover N_{k} values of wavenumber k, then one wishes to determine $2N_{\mathrm{k}}+1$ quantities based on $2N_{\mathrm{k}}$ measurements, which should not cause significant degeneracies when $N_{\mathrm{k}} \gg 1$. Even without knowing β, one can probe whether some sources of $P_{\mathrm{rad}}(k)$ are uncorrelated with δ; the quantity $P_{\mathrm{un}-\delta}(k) \equiv P_{\mu^0} - P_{\mu^2}^2/(4P_{\mu^4})$ equals $P_{\mathrm{rad}} - P_{\delta\cdot\mathrm{rad}}^2/P_\delta$, which receives no contribution from any source that is a linear functional of the density distribution (see the next subsection for an example).

Specific Epochs

At $z \sim 35$, collisions are effective due to the high gas density, so one can measure the density power spectrum [225] and the redshift evolution of n_{HI}, T_γ, and T_{k}. At $z \lesssim 35$, collisions become ineffective but the first stars produce a cosmic background of Lyα photons (i.e. photons that redshift into the Lyα resonance) that couples T_{s} to T_{k}. During the period of initial Lyα coupling, fluctuations in the Lyα flux translate into fluctuations in the 21 cm brightness [30]. This signal can be observed from $z \sim 25$ until the Lyα coupling is completed (i.e., $x_{\mathrm{tot}} \gg 1$) at $z \sim 15$. At a given redshift, each atom sees Lyα photons that were originally emitted at earlier times at rest-frame wavelengths between Lyα and the Lyman limit. Distant sources are time retarded, and since there are fewer galaxies in the distant, earlier Universe, each atom sees sources only out to an apparent source horizon of $\sim$100 comoving Mpc at $z \sim 20$. A significant portion of the flux comes from nearby sources, because of the $1/r^2$ decline of flux with distance, and since higher Lyman series photons, which are degraded to Lyα photons through scattering, can only be seen from a small redshift interval that corresponds to the wavelength interval between two consecutive atomic levels.

There are two separate sources of fluctuations in the Lyα flux [30]. The first is density inhomogeneities. Since gravitational instability proceeds faster in overdense regions, the biased distribution of rare galactic halos fluctuates much more than the global dark matter density. When the number of sources seen by each atom is relatively small, Poisson fluctuations provide a second source of fluctuations. Unlike typical Poisson noise, these fluctuations are correlated between gas elements at different places, since two nearby elements see many of the same sources. Assuming a scale-invariant spectrum of primordial density fluctuations, and that $x_\alpha = 1$ is produced at $z = 20$ by galaxies in dark matter halos where the gas cools efficiently via atomic cooling, Fig. 60 shows the predicted observable power spectra. The figure suggests that β can be measured from the ratio P_{μ^2}/P_{μ^4} at $k \gtrsim 1\,\mathrm{Mpc}^{-1}$, allowing the density-induced fluctuations in flux to be extracted from P_{μ^2}, while only the Poisson

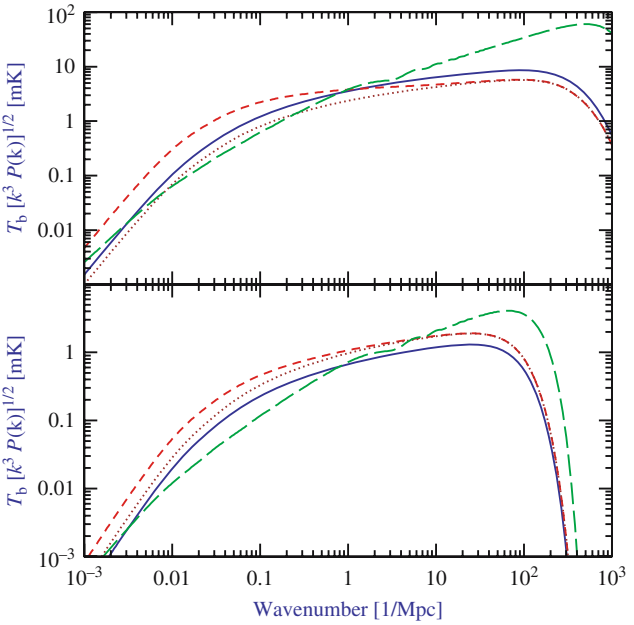

Fig. 60. Observable power spectra during the period of initial Lyα coupling. **Upper panel:** Assumes adiabatic cooling. **Lower panel:** Assumes pre-heating to 500 K by X-ray sources. Shown are $P_{\mu^4} = P_\delta$ (*solid curves*), P_{μ^2} (*short-dashed curves*), and $P_{\mathrm{un}-\delta}$ (*long-dashed curves*), as well as for comparison $2\beta P_\delta$ (*dotted curves*)

fluctuations contribute to $P_{\mathrm{un}-\delta}$. Each of these components probes the number density of galaxies through its magnitude, and the distribution of source distances through its shape. Measurements at $k \gtrsim 100\,\mathrm{Mpc}^{-1}$ can independently probe T_k because of the smoothing effects of the gas pressure and the thermal width of the 21 cm line.

After Lyα coupling and X-ray heating are both completed, reionization continues. Since $\beta = 1$ and $T_\mathrm{k} \gg T_\gamma$, the normalization of P_{μ^4} directly measures the mean neutral hydrogen fraction, and one can separately probe the density fluctuations, the neutral hydrogen fluctuations, and their cross-correlation.

Fluctuations on Large Angular Scales

Full-sky observations must normally be analyzed with an angular and radial transform [143, 313, 41], rather than a Fourier transform which is simpler and yields more directly the underlying 3D power spectrum [258, 259]. The 21 cm brightness fluctuations at a given redshift – corresponding to a comoving distance r_0 from the observer – can be expanded in spherical harmonics with expansion coefficients $a_{\mathrm{lm}}(\nu)$, where the angular power spectrum is

$$C_{\rm l}(r_0) = \langle |a_{\rm lm}(\nu)|^2 \rangle = 4\pi \int \frac{k^2 {\rm d}k}{2\pi^2} \Big[G_{\rm l}^2(kr_0) P_\delta(k) + 2P_{\delta\cdot{\rm rad}}(k) G_{\rm l}(kr_0) j_{\rm l}(kr_0) + P_{\rm rad}(k) j_{\rm l}^2(kr_0) \Big] , \qquad (181)$$

with $G_{\rm l}(x) \equiv J_{\rm l}(x) + (\beta - 1) j_{\rm l}(x)$ and $J_{\rm l}(x)$ being a linear combination of spherical Bessel functions [41].

In an angular transform on the sky, an angle of θ radians translates to a spherical multipole $l \sim 3.5/\theta$. For measurements on a screen at a comoving distance r_0, a multipole l normally measures 3D power on a scale of $k^{-1} \sim \theta r_0 \sim 35/l\,{\rm Gpc}$ for $l \gg 1$, since $r_0 \sim 10\,{\rm Gpc}$ at $z \gtrsim 10$. This estimate fails at $l \lesssim 100$, however, when we consider the sources of 21 cm fluctuations. The angular projection implied in C_l involves a weighted average (181) that favors large scales when l is small, but density fluctuations possess little large-scale power, and the $C_{\rm l}$ are dominated by power around the peak of $kP_\delta(k)$, at a few tens of comoving Mpc.

Figure 61 shows that for density and velocity fluctuations, even the $l = 1$ multipole is affected by power at $k^{-1} > 200\,{\rm Mpc}$ only at the 2% level. Due to the small number of large angular modes available on the sky, the expectation value of $C_{\rm l}$ cannot be measured precisely at small l. Figure 61 shows that

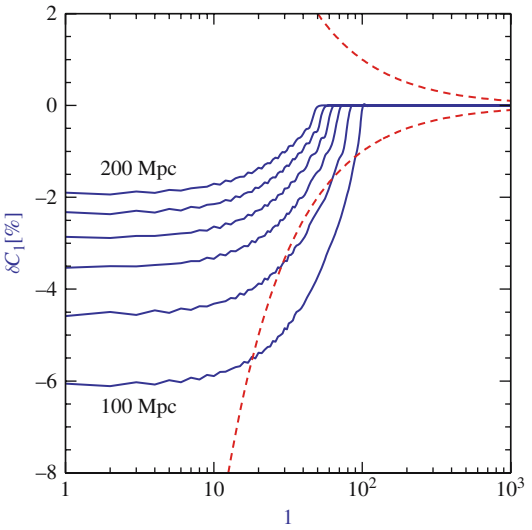

Fig. 61. Effect of large-scale power on the angular power spectrum of 21 cm anisotropies on the sky. This example shows the power from density fluctuations and velocity compression, assuming a warm IGM at $z = 12$ with $T_{\rm s} = T_{\rm k} \gg T_\gamma$. Shown is the % change in $C_{\rm l}$ if we were to cut off the power spectrum above $1/k$ of 200, 180, 160, 140, 120, and 100 Mpc (*top to bottom*). Also shown for comparison is the cosmic variance for averaging in bands of $\Delta l \sim l$ (*dashed lines*)

this precludes new information from being obtained on scales $k^{-1} \gtrsim 130\,\mathrm{Mpc}$ using angular structure at any given redshift. Fluctuations on such scales may be measurable using a range of redshifts, but the required $\Delta z \gtrsim 1$ at $z \sim 10$ implies significant difficulties with foreground subtraction and with the need to account for time evolution.

10 Major Challenge for Future Theoretical Research

Radiative transfer during reionization requires a large dynamic range, challenging the capabilities of existing simulation codes.

Observations of the cosmic microwave background [347] have confirmed the notion that the present large-scale structure in the Universe originated from small-amplitude density fluctuations at early cosmic times. Due to the natural instability of gravity, regions that were denser than average collapsed and formed bound halos, first on small spatial scales and later on larger and larger scales. At each snapshot of this cosmic evolution, the abundance of collapsed halos, whose masses are dominated by cold dark matter, can be computed from the initial conditions using numerical simulations and can be understood using approximate analytic models [291, 52]. The common understanding of galaxy formation is based on the notion that the constituent stars formed out of the gas that cooled and subsequently condensed to high densities in the cores of some of these halos [378].

The standard analytic model for the abundance of halos [291, 52] considers the small density fluctuations at some early, initial time, and attempts to predict the number of halos that will form at some later time corresponding to a redshift z. First, the fluctuations are extrapolated to the present time using the growth rate of linear fluctuations, and then the average density is computed in spheres of various sizes. Whenever the overdensity (i.e., the density perturbation in units of the cosmic mean density) in a sphere rises above a critical threshold $\delta_{\mathrm{c}}(z)$, the corresponding region is assumed to have collapsed by redshift z, forming a halo out of all the mass that had been included in the initial spherical region. In analyzing the statistics of such regions, the model separates the contribution of large-scale modes from that of small-scale density fluctuations. It predicts that galactic halos will form earlier in regions that are overdense on large scales [188, 19, 97, 257], since these regions already start out from an enhanced level of density, and small-scale modes need only supply the remaining perturbation necessary to reach $\delta_{\mathrm{c}}(z)$. On the other hand, large-scale voids should contain a reduced number of halos at high redshift. In this way, the analytic model describes the clustering of massive halos.

As gas falls into a dark matter halo, it can fragment into stars only if its virial temperature is above $10^4\,\mathrm{K}$ for cooling mediated by atomic transitions [or $\sim 500\,\mathrm{K}$ for molecular H_2 cooling; see Fig. 62]. The abundance of dark matter halos with a virial temperature above this cooling threshold declines

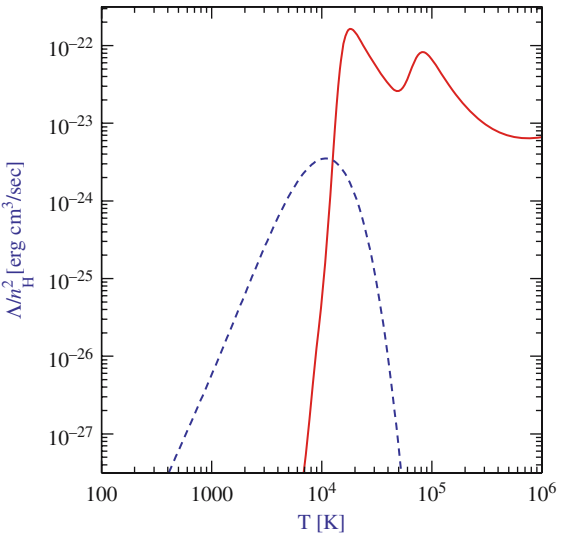

Fig. 62. Cooling rates as a function of temperature for a primordial gas composed of atomic hydrogen and helium, as well as molecular hydrogen, in the absence of any external radiation (from Barkana & Loeb 2001 [23]). We assume a hydrogen number density $n_{\rm H} = 0.045$ cm^{-3}, corresponding to the mean density of virialized halos at $z = 10$. The plotted quantity $\Lambda/n_{\rm H}^2$ is roughly independent of density (unless $n_{\rm H} \gtrsim 10$ cm^{-3}), where Λ is the volume cooling rate (in erg sec^{-1} cm^{-3}). The solid line shows the cooling curve for an atomic gas, with the characteristic peaks due to collisional excitation of H1 and He2. The dashed line shows the additional contribution of molecular cooling, assuming a molecular abundance equal to 1% of $n_{\rm H}$

sharply with increasing redshift due to the exponential cutoff in the abundance of massive halos at early cosmic times. Consequently, a small change in the collapse threshold of these rare halos, due to mild inhomogeneities on much larger spatial scales, can change the abundance of such halos dramatically. Barkana & Loeb (2004) [27] have shown that the modulation of galaxy formation by long wavelength modes of density fluctuations is therefore amplified considerably at high redshift; the discussion in this section follows their analysis.

10.1 Amplification of Density Fluctuations

Galaxies at high redshift are believed to form in all halos above some minimum mass $M_{\rm min}$ that depends on the efficiency of atomic and molecular transitions that cool the gas within each halo. This makes useful the standard quantity of the collapse fraction $F_{\rm col}(M_{\rm min})$, which is the fraction of mass in a given volume that is contained in halos of individual mass $M_{\rm min}$ or greater (see Fig. 63). If we set $M_{\rm min}$ to be the minimum halo mass in which efficient

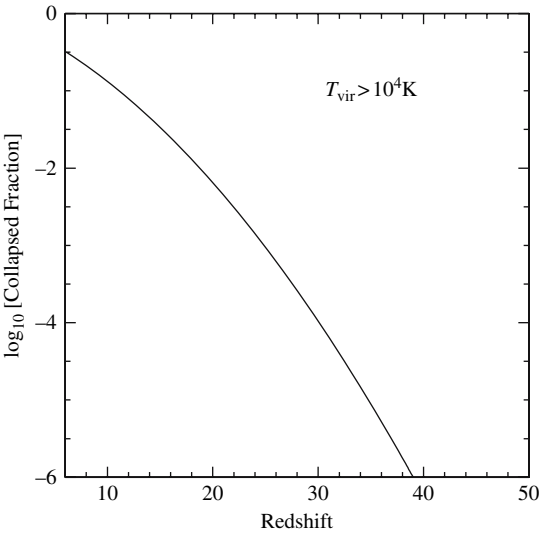

Fig. 63. Fraction of baryons that assembled into dark matter halos with a virial temperature of $T_{\rm vir} > 10^4\,{\rm K}$ as a function of redshift. These baryons are above the temperature threshold for gas cooling and fragmentation via atomic transitions. After reionization the temperature barrier for star formation in galaxies is raised because the photo-ionized intergalactic medium is already heated to $\sim 10^4\,{\rm K}$ and it can condense only into halos with $T_{\rm vir} > 10^5\,{\rm K}$

cooling processes are triggered, then $F_{\rm col}(M_{\rm min})$ is the fraction of all baryons that reside in galaxies. In a large-scale region of comoving radius R with a mean overdensity $\bar{\delta}_{\rm R}$, the standard result is

$$F_{\rm col}(M_{\rm min}) = {\rm erfc}\left[\frac{\delta_{\rm c}(z) - \bar{\delta}_{\rm R}}{\sqrt{2\,[S(R_{\rm min}) - S(R)]}}\right]\,, \tag{182}$$

where $S(R) = \sigma^2(R)$ is the variance of fluctuations in spheres of radius R, and $S(R_{\rm min})$ is the variance in spheres of radius $R_{\rm min}$ corresponding to the region at the initial time that contained a mass $M_{\rm min}$. In particular, the cosmic mean value of the collapse fraction is obtained in the limit of $R \to \infty$ by setting $\bar{\delta}_{\rm R}$ and $S(R)$ to zero in this expression. Throughout this section we shall adopt this standard model, known as the extended Press-Schechter model. Whenever we consider a cubic region, we will estimate its halo abundance by applying the model to a spherical region of equal volume. Note also that we will consistently quote values of comoving distance, which equals physical distance times a factor of $(1 + z)$.

At high redshift, galactic halos are rare and correspond to high peaks in the Gaussian probability distribution of initial fluctuations. A modest change in the overall density of a large region modulates the threshold for high peaks in the Gaussian density field, so that the number of galaxies is exponentially

sensitive to this modulation. This amplification of large-scale modes is responsible for the large statistical fluctuations that we find.

In numerical simulations, periodic boundary conditions are usually assumed, and this forces the mean density of the box to equal the cosmic mean density. The abundance of halos as a function of mass is then biased in such a box (see Fig. 64), since a similar region in the real Universe will have a distribution of different overdensities $\bar{\delta}_{\rm R}$. At high redshift, when galaxies correspond to high peaks, they are mostly found in regions with an enhanced large-scale density. In a periodic box, therefore, the total number of galaxies is artificially reduced, and the relative abundance of galactic halos with different masses is artificially tilted in favor of lower-mass halos. Let us illustrate these results for two sets of parameters, one corresponding to the first galaxies and early reionization ($z = 20$) and the other to the current horizon in observations of galaxies and late reionization ($z = 7$). Let us consider a resolution equal to that of state-of-the-art cosmological simulations that include gravity and

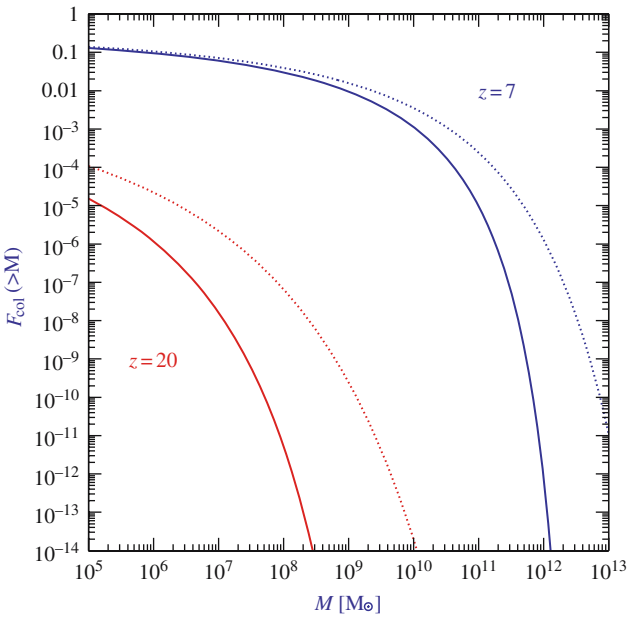

Fig. 64. Bias in the halo mass distribution in simulations. Shown is the amount of mass contained in all halos of individual mass $M_{\rm min}$ or greater, expressed as a fraction of the total mass in a given volume. This cumulative fraction $F_{\rm col}(M_{\rm min})$ is illustrated as a function of the minimum halo mass $M_{\rm min}$. We consider two cases of redshift and simulation box size, namely $z = 7$, $l_{\rm box} = 6\,{\rm Mpc}$ (*upper curves*), and $z = 20$, $l_{\rm box} = 1\,{\rm Mpc}$ (*lower curves*). At each redshift, we compare the true average distribution in the Universe (*dotted curve*) to the biased distribution (*solid curve*) that would be measured in a simulation box with periodic boundary conditions (for which $\bar{\delta}_{\rm R}$ is artificially set to zero)

gas hydrodynamics. Specifically, let us assume that the total number of dark matter particles in the simulation is $N = 324^3$, and that the smallest halo that can form a galaxy must be resolved into 500 particles; [348] showed that this resolution is necessary in order to determine the star formation rate in an individual halo reliably to within a factor of two. Therefore, if we assume that halos that cool via molecular hydrogen must be resolved at $z = 20$ (so that $M_{\rm min} = 7 \times 10^5\ {\rm M}_\odot$), and only those that cool via atomic transitions must be resolved at $z = 7$ (so that $M_{\rm min} = 10^8\ {\rm M}_\odot$), then the maximum box sizes that can currently be simulated in hydrodynamic comological simulations are $l_{\rm box} = 1\,{\rm Mpc}$ and $l_{\rm box} = 6\,{\rm Mpc}$ at these two redshifts, respectively.

At each redshift we only consider cubic boxes large enough so that the probability of forming a halo on the scale of the entire box is negligible. In this case, $\bar{\delta}_{\rm R}$ is Gaussian distributed with zero mean and variance $S(R)$, since the no-halo condition $\sqrt{S(R)} \ll \delta_{\rm c}(z)$ implies that at redshift z the perturbation on the scale R is still in the linear regime. We can then calculate the probability distribution of collapse fractions in a box of a given size (see Fig. 65). This distribution corresponds to a real variation in the fraction of

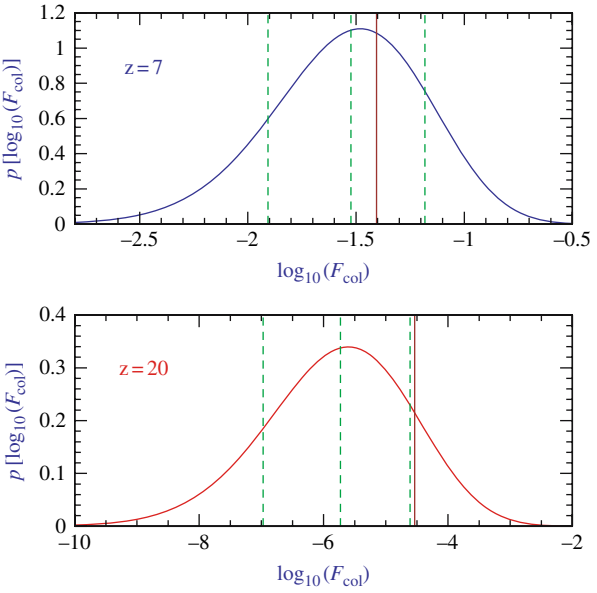

Fig. 65. Probability distribution within a small volume of the total mass fraction in galactic halos. The normalized distribution of the logarithm of this fraction $F_{\rm col}(M_{\rm min})$ is shown for two cases: $z = 7$, $l_{\rm box} = 6\,{\rm Mpc}$, $M_{\rm min} = 10^8\ {\rm M}_\odot$ (**upper panel**), and $z = 20$, $l_{\rm box} = 1\,{\rm Mpc}$, $M_{\rm min} = 7 \times 10^5\ {\rm M}_\odot$ (**bottom panel**). In each case, the value in a periodic box ($\bar{\delta}_{\rm R} = 0$) is shown along with the value that would be expected given a plus or minus $1 - \sigma$ fluctuation in the mean density of the box (*dashed vertical lines*). Also shown in each case is the mean value of $F_{\rm col}(M_{\rm min})$ averaged over large cosmological volumes (*solid vertical line*)

gas in galaxies within different regions of the Universe at a given time. In a numerical simulation, the assumption of periodic boundary conditions eliminates the large-scale modes that cause this cosmic scatter. Note that Poisson fluctuations in the number of halos within the box would only add to the scatter, although the variations we have calculated are typically the dominant factor. For instance, in our two standard examples, the mean expected number of halos in the box is 3 at $z = 20$ and 900 at $z = 7$, resulting in Poisson fluctuations of a factor of about 2 and 1.03, respectively, compared to the clustering-induced scatter of a factor of about 16 and 2 in these two cases.

Within the extended Press-Schechter model, both the numerical bias and the cosmic scatter can be simply described in terms of a shift in the redshift (see Fig. 66). In general, a region of radius R with a mean overdensity $\bar{\delta}_{\rm R}$ will contain a different collapse fraction than the cosmic mean value at a given redshift z. However, at some wrong redshift $z + \Delta z$ this small region will contain the cosmic mean collapse fraction at z. At high redshifts ($z > 3$), this shift in redshift was derived by Barkana & Loeb [27] from (182) [and was already mentioned in (168)]

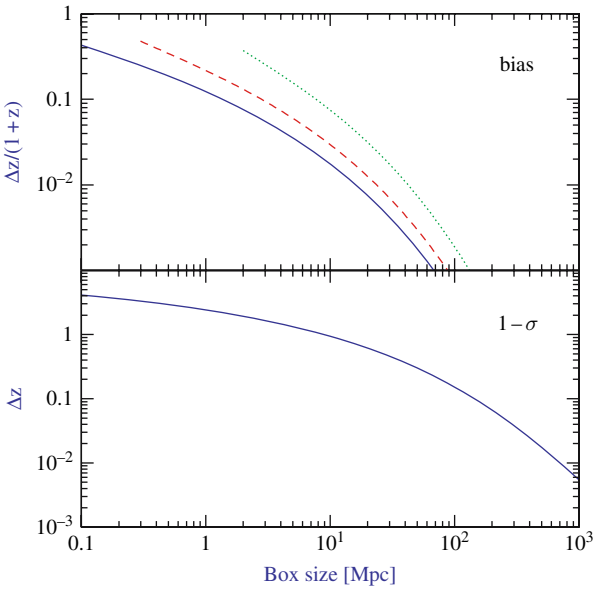

Fig. 66. Cosmic scatter and numerical bias, expressed as the change in redshift needed to get the correct cosmic mean of the collapse fraction. The plot shows the $1-\sigma$ scatter (about the biased value) in the redshift of reionization, or any other phenomenon that depends on the mass fraction in galaxies (**bottom panel**), as well as the redshift bias [expressed as a fraction of $(1+z)$] in periodic simulation boxes (**upper panel**). The bias is shown for $M_{\rm min} = 7 \times 10^5$ $M_\odot$ (*solid curve*), $M_{\rm min} = 10^8$ $M_\odot$ (*dashed curve*), and $M_{\rm min} = 3 \times 10^{10}$ $M_\odot$ (*dotted curve*). The bias is always negative, and the plot gives its absolute value. When expressed as a shift in redshift, the scatter is independent of $M_{\rm min}$

$$\Delta z = \frac{\bar{\delta}_{\rm R}}{\delta_0} - (1+z) \times \left[1 - \sqrt{1 - \frac{S(R)}{S(R_{\rm min})}} \right] \ , \tag{183}$$

where $\delta_0 \equiv \delta_{\rm c}(z)/(1+z)$ is approximately constant at high redshifts [282], and equals 1.28 for the standard cosmological parameters (with its deviation from the Einstein-de Sitter value of 1.69 resulting from the existence of a cosmological constant). Thus, in our two examples, the bias is -2.6 at $z = 20$ and -0.4 at $z = 7$, and the one-sided $1-\sigma$ scatter is 2.4 at $z = 20$ and 1.2 at $z = 7$.

10.2 Matching Numerical Simulations

Next we may develop an improved model that fits the results of numerical simulations more accurately. The model constructs the halo mass distribution (or mass function); cumulative quantities such as the collapse fraction or the total number of galaxies can then be determined from it via integration. We first define $f(\delta_{\rm c}(z), S)\,{\rm d}S$ to be the mass fraction contained at z within halos with mass in the range corresponding to S to $S + {\rm d}S$. As derived earlier, the Press-Schechter halo abundance is

$$\frac{{\rm d}n}{{\rm d}M} = \frac{\bar{\rho}_0}{M} \left| \frac{{\rm d}S}{{\rm d}M} \right| f(\delta_{\rm c}(z), S) \ , \tag{184}$$

where ${\rm d}n$ is the comoving number density of halos with masses in the range M to $M + {\rm d}M$, and

$$f_{\rm PS}(\delta_{\rm c}(z), S) = \frac{1}{\sqrt{2\pi}} \frac{\nu}{S} \exp\left[-\frac{\nu^2}{2} \right] \ , \tag{185}$$

where $\nu = \delta_{\rm c}(z)/\sqrt{S}$ is the number of standard deviations that the critical collapse overdensity represents on the mass scale M corresponding to the variance S.

However, the Press-Schechter mass function fits numerical simulations only roughly, and in particular it substantially underestimates the abundance of the rare halos that host galaxies at high redshift. The halo mass function of [332] [see also [333]] adds two free parameters that allow it to fit numerical simulations much more accurately [186]. These N-body simulations followed very large volumes at low redshift, so that cosmic scatter did not compromise their accuracy. The matching mass function is given by

$$f_{\rm ST}(\delta_{\rm c}(z), S) = A' \frac{\nu}{S} \sqrt{\frac{a'}{2\pi}} \left[1 + \frac{1}{(a'\nu^2)^{q'}} \right] \exp\left[-\frac{a'\nu^2}{2} \right] \ , \tag{186}$$

with best-fit parameters [334] $a' = 0.75$ and $q' = 0.3$, and where normalization to unity is ensured by taking $A' = 0.322$.

In order to calculate cosmic scatter we must determine the biased halo mass function in a given volume at a given mean density. Within the extended Press-Schechter model [52], the halo mass distribution in a region of comoving radius R with a mean overdensity $\bar{\delta}_R$ is given by

$$f_{\rm bias-PS}(\delta_c(z), \bar{\delta}_{\rm R}, R, S) = f_{\rm PS}(\delta_{\rm c}(z) - \bar{\delta}_{\rm R}, S - S(R)) \ . \tag{187}$$

The corresponding collapse fraction in this case is given simply by (182). Despite the relatively low accuracy of the Press-Schechter mass function, the *relative change* is predicted rather accurately by the extended Press-Schechter model. In other words, the prediction for the halo mass function in a given volume compared to the cosmic mean mass function provides a good fit to numerical simulations over a wide range of parameters [257, 77].

For the improved model (derived in [27]), we adopt a hybrid approach that combines various previous models with each applied where it has been found to closely match numerical simulations. We obtain the halo mass function within a restricted volume by starting with the Sheth-Torme formula for the cosmic mean mass function, and then adjusting it with a relative correction based on the extended Press-Schechter model. In other words, we set

$$\begin{aligned} &f_{\rm bias}(\delta_{\rm c}(z), \bar{\delta}_{\rm R}, R, S) = \\ &f_{\rm ST}(\delta_{\rm c}(z), S) \ \times \left[\frac{f_{\rm PS}(\delta_c(z) - \bar{\delta}_{\rm R}, S - S(R))}{f_{\rm PS}(\delta_c(z), S)} \right] \ . \end{aligned} \tag{188}$$

As noted, this model is based on fits to simulations at low redshifts, but we can check it at high redshifts as well. Figure 67 shows the number of galactic halos at $z \sim 15$–30 in two numerical simulations run by [401], and our predictions given the cosmological input parameters assumed by each simulation. The close fit to the simulated data (with no additional free parameters) suggests that our hybrid model (solid lines) improves on the extended Press-Schechter model (dashed lines), and can be used to calculate accurately the cosmic scatter in the number of galaxies at both high and low redshifts. The simulated data significantly deviate from the expected cosmic mean [(186), shown by the dotted line], due to the artificial suppression of large-scale modes outside the simulated box.

As an additional example, we consider the highest-resolution first star simulation [5], which used $l_{\rm box} = 128\,{\rm kpc}$ and $M_{\rm min} = 7 \times 10^5\ {\rm M}_\odot$. The first star forms within the simulated volume when the first halo of mass $M_{\rm min}$ or larger collapses within the box. To compare with the simulation, we predict the redshift at which the probability of finding at least one halo within the box equals 50%, accounting for Poisson fluctuations. We find that if the simulation formed a population of halos corresponding to the correct cosmic average [as given by (186)], then the first star should have formed already at $z = 24.0$. The first star actually formed in the simulation box only at $z = 18.2$ [5]. Using (188) we can account for the loss of large-scale modes beyond the periodic

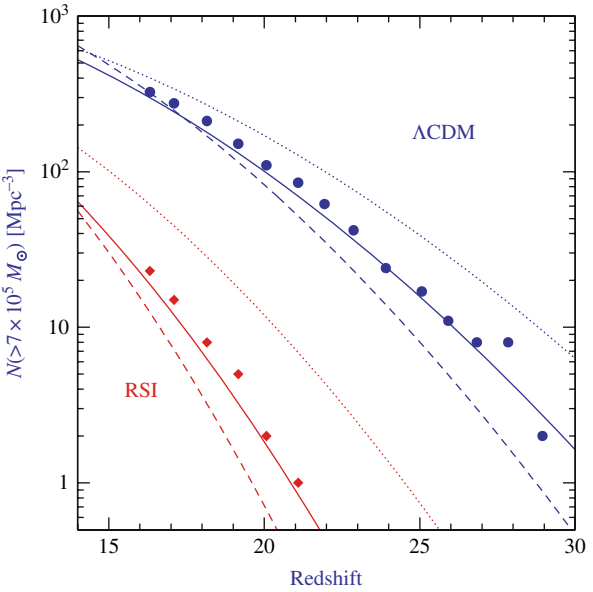

Fig. 67. Halo mass function at high redshift in a 1 Mpc box at the cosmic mean density. The prediction (solid lines) of the hybrid model of Barkana & Loeb (2004) [27] is compared with the number of halos above mass 7×10^5 $M_\odot$ measured in the simulations of Yośhida et al. (2003) [401] [data points are taken from their Fig. 5]. The cosmic mean of the halo mass function (*dotted lines*) deviates significantly from the simulated values, since the periodic boundary conditions within the finite simulation box artificially set the amplitude of large-scale modes to zero. The hybrid model starts with the Sheth-Tormen mass function and applies a correction based on the extended Press-Schechter model; in doing so, it provides a better fit to numerical simulations than the pure extended Press-Schechter model (*dashed lines*) used in the previous figures. We consider two sets of cosmological parameters, the scale-invariant ΛCDM model of Yośhida et al. (2003) [401] (*upper curves*), and their running scalar index (RSI) model (*lower curves*)

box, and predict a first star at $z = 17.8$, a close match given the large Poisson fluctuations introduced by considering a single galaxy within the box.

The artificial bias in periodic simulation boxes can also be seen in the results of extensive numerical convergence tests carried out by [348]. They presented a large array of numerical simulations of galaxy formation run in periodic boxes over a wide range of box size, mass resolution, and redshift. In particular, we can identify several pairs of simulations where the simulations in each pair have the same mass resolution but different box sizes; this allows us to separate the effect of large-scale numerical bias from the effect of having poorly-resolved individual halos.

10.3 Implications

The Nature of Reionization

A variety of papers in the literature [13, 138, 329, 169, 152, 23, 223] maintain that reionization ended with a fast, simultaneous, overlap stage throughout the Universe. This view has been based on simple arguments and has been supported by numerical simulations with small box sizes. The underlying idea was that the ionized hydrogen (H II) regions of individual sources began to overlap when the typical size of each H II bubble became comparable to the distance between nearby sources. Since these two length scales were comparable at the critical moment, there is only a single timescale in the problem – given by the growth rate of each bubble – and it determines the transition time between the initial overlap of two or three nearby bubbles, to the final stage where dozens or hundreds of individual sources overlap and produce large ionized regions. Whenever two ionized bubbles were joined, each point inside their common boundary became exposed to ionizing photons from both sources, reducing the neutral hydrogen fraction and allowing ionizing photons to travel farther before being absorbed. Thus, the ionizing intensity inside H II regions rose rapidly, allowing those regions to expand into high-density gas that had previously recombined fast enough to remain neutral when the ionizing intensity had been low. Since each bubble coalescence accelerates the process, it has been thought that the overlap phase has the character of a phase transition and occurs rapidly. Indeed, the simulations of reionization [152] found that the average mean free path of ionizing photons in the simulated volume rises by an order of magnitude over a redshift interval $\Delta z = 0.05$ at $z = 7$.

These results imply that overlap is still expected to occur rapidly, but only in localized high-density regions, where the ionizing intensity and the mean free path rise rapidly even while other distant regions are still mostly neutral. In other words, the size of the bubble of an individual source is about the same in different regions (since most halos have masses just above $M_{\rm min}$), but the typical distance between nearby sources varies widely across the Universe. The strong clustering of ionizing sources on length scales as large as 30–100 Mpc introduces long timescales into the reionization phase transition. The sharpness of overlap is determined not by the growth rate of bubbles around individual sources, but by the ability of large groups of sources within overdense regions to deliver ionizing photons into large underdense regions.

Note that the recombination rate is higher in overdense regions because of their higher gas density. These regions still reionize first, though, despite the need to overcome the higher recombination rate, since the number of ionizing sources in these regions is increased even more strongly as a result of the dramatic amplification of large-scale modes discussed earlier.

Limitations of Current Simulations

The shortcomings of current simulations do not amount simply to a shift of $\sim$10% in redshift and the elimination of scatter. The effect mentioned above can be expressed in terms of a shift in redshift only within the context of the extended Press-Schechter model, and only if the total mass fraction in galaxies is considered and not its distribution as a function of galaxy mass. The halo mass distribution should still have the wrong shape, resulting from the fact that Δz depends on $M_{\rm min}$. A self-contained numerical simulation must directly evolve a very large volume.

Another reason that current simulations are limited is that at high redshift, when galaxies are still rare, the abundance of galaxies grows rapidly towards lower redshift. Therefore, a $\sim$ 10% relative error in redshift implies that at any given redshift around $z \sim$ 10–20, the simulation predicts a halo mass function that can be off by an order of magnitude for halos that host galaxies (see Fig. 67). This large underestimate suggests that the first generation of galaxies formed significantly earlier than indicated by recent simulations. Another element missed by simulations is the large cosmic scatter. 'This scatter can fundamentally change the character of any observable process or feedback mechanism that depends on a radiation background. Simulations in periodic boxes eliminate any large-scale scatter by assuming that the simulated volume is surrounded by identical periodic copies of itself. In the case of reionization, for instance, current simulations neglect the collective effects described above, whereby groups of sources in overdense regions may influence large surrounding underdense regions. In the case of the formation of the first stars due to molecular hydrogen cooling, the effect of the soft ultraviolet radiation from these stars, which tends to dissociate the molecular hydrogen around them [168, 302, 271], must be reassessed with cosmic scatter included.

Observational Consequences

The spatial fluctuations that we have calculated also affect current and future observations that probe reionization or the galaxy population at high redshift. For example, there are a large number of programs searching for galaxies at the highest accessible redshifts (6.5 and beyond) using their strong Lyα emission [182, 300, 243, 200]. These programs have previously been justified as a search for the reionization redshift, since the intrinsic emission should be absorbed more strongly by the surrounding IGM if this medium is neutral. For any particular source, it will be hard to clearly recognize this enhanced absorption because of uncertainties regarding the properties of the source and its radiative and gravitational effects on its surroundings [24, 26, 311]. However, if the luminosity function of galaxies that emit Lyα can be observed, then faint sources, which do not significantly affect their environment, should be very strongly absorbed in the era before reionization. Reionization can then be detected statistically through the sudden jump in the number of faint

Sources [245]. The above results alter the expectation for such observations. Indeed, no sharp "reionization redshift" is expected. Instead, a Lyα luminosity function assembled from a large area of the sky will average over the cosmic scatter of $\Delta z \sim 1$–2 between different regions, resulting in a smooth evolution of the luminosity function over this redshift range. In addition, such a survey may be biased to give a relatively high redshift, since only the most massive galaxies can be detected, and as we have shown, these galaxies will be concentrated in overdense regions that will also get reionized relatively early.

The distribution of ionized patches during reionization will likely be probed by future observations, including small-scale anisotropies of the cosmic microwave background photons that are rescattered by the ionized patches [8, 162, 312], and observations of 21 cm emission by the spin-flip transition of the hydrogen in neutral regions [365, 75, 142]. Previous analytical and numerical estimates of these signals have not included the collective effects discussed above, in which rare groups of massive galaxies may reionize large surrounding areas. The transfer of photons across large scales will likely smooth out the signal even on scales significantly larger than the typical size of an H II bubble due to an individual galaxy. Therefore, even the characteristic angular scales that are expected to show correlations in such observations must be reassessed.

The cosmic scatter also affects observations in the present-day Universe that depend on the history of reionization. For instance, photoionization heating suppresses the formation of dwarf galaxies after reionization, suggesting that the smallest galaxies seen today may have formed prior to reionization [73, 343, 37]. Under the popular view that assumed a sharp end to reionization, it was expected that denser regions would have formed more galaxies by the time of reionization, possibly explaining the larger relative abundance of dwarf galaxies observed in galaxy clusters compared to lower-density regions such as the Local Group of galaxies [368, 38]. The above results undercut the basic assumption of this argument and suggest a different explanation. Reionization occurs roughly when the number of ionizing photons produced starts to exceed the number of hydrogen atoms in the surrounding IGM. If the processes of star formation and the production of ionizing photons are equally efficient within galaxies that lie in different regions, then reionization in each region will occur when the collapse fraction reaches the same critical value, even though this will occur at different times in different regions. Since the galaxies responsible for reionization have the same masses as present-day dwarf galaxies, this estimate argues for a roughly equal abundance of dwarf galaxies in all environments today. This simple picture is, however, modified by several additional effects. First, the recombination rate is higher in overdense regions at any given time, as discussed above. Furthermore, reionization in such regions is accomplished at an earlier time when the recombination rate was higher even at the mean cosmic density; therefore, more ionizing photons must be produced in order to compensate for the

enhanced recombination rate. These two effects combine to make overdense regions reionize at a higher value of $F_{\rm col}$ than underdense regions. In addition, the overdense regions, which reionize first, subsequently send their extra ionizing photons into the surrounding underdense regions, causing the latter to reionize at an even lower $F_{\rm col}$. Thus, a higher abundance of dwarf galaxies today is indeed expected in the overdense regions.

The same basic effect may be even more critical for understanding the properties of large-scale voids, 10–30 Mpc regions in the present-day Universe with an average mass density that is well below the cosmic mean. In order to predict their properties, the first step is to consider the abundance of dark matter halos within them. Numerical simulations show that voids contain a lower relative abundance of rare halos [248, 82, 39], as expected from the raising of the collapse threshold for halos within a void. On the other hand, simulations show that voids actually place a larger fraction of their dark matter content in dwarf halos of mass below 10^{10} $M_\odot$ [157]. This can be understood within the extended Press-Schechter model. At the present time, a typical region in the Universe fills halos of mass 10^{12} $M_\odot$ and higher with most of the dark matter, and very little is left over for isolated dwarf halos. Although a large number of dwarf halos may have formed at early times in such a region, the vast majority later merged with other halos, and by the present time they survive only as substructure inside much larger halos. In a void, on the other hand, large halos are rare even today, implying that most of the dwarf halos that formed early within a void can remain as isolated dwarf halos till the present. Thus, most isolated dwarf dark matter halos in the present Universe should be found within large-scale voids [25].

However, voids are observed to be rather deficient in dwarf galaxies as well as in larger galaxies on the scale of the Milky Way mass of $\sim 10^{12}$ $M_\odot$ [198, 120, 284]. A deficit of large galaxies is naturally expected, since the total mass density in the void is unusually low, and the fraction of this already low density that assembles in large halos is further reduced relative to higher-density regions. The absence of dwarf galaxies is harder to understand, given the higher relative abundance expected for their host dark matter halos. The standard model for galaxy formation may be consistent with the observations if some of the dwarf halos are dark and do not host stars. Large numbers of dark dwarf halos may be produced by the effect of reionization in suppressing the infall of gas into these halos. Indeed, exactly the same factors considered above, in the discussion of dwarf galaxies in clusters compared to those in small groups, apply also to voids. Thus, the voids should reionize last, but since they are most strongly affected by ionizing photons from their surroundings (which have a higher density than the voids themselves), the voids should reionize when the abundance of galaxies within them is relatively low.

References

1. Abel, T., Haehnelt M.G.: ApJ. **520**, 13 (1999)
2. Abel, T., Norman, M.L., Madau P.: ApJ. **523**, 66 (1999)
3. Abel, T., Mo H.J.: ApJ. **494**, L151 (1998)
4. Abel, T., Bryan, G., Norman M.: ApJ. **540**, 39 (2000)
5. Abel, T., Bryan, G.L., Norman M.L.: Science **295**, 93 (2002)
6. Adelberger, K.L., Steidel C.C.: ApJ. **544**, 218 (2000)
7. Adelberger, K.A., et al.: ApJ. **584**, 45 (2003)
8. Aghanim, N., Desert, F.X., Puget, J.L., Gispert R.: A&A **311**, 1 (1996)
9. Aguirre, A., Hernquist, L., Weinberg, D., Katz, N., Gardner J.: ApJ. **560**, 599 (2001)
10. Aguirre, A., Hernquist, L., Katz, N., Gardner, J., Weinberg D.: ApJ. **556**, L11 (2001)
11. Allison, A.C., Dalgarno A.: ApJ. **158**, 423 (1969)
12. Alvarez, M.A., Bromm, V., Shapiro P.R.: ApJ. **639**, 612 (2006), (astro-ph/0507684)
13. Arons, J., Wingert D.W.: ApJ. **177**, 1 (1972)
14. Babich, D., Loeb A.: ApJ. in press (2006), (astro-ph/0509784)
15. Babul, A., White S.D.M.: MNRAS **253**, P31 (1991)
16. Baltz E.A.,: arXiv:(astro-ph/0412170) (2004)
17. Baltz, E.A., Gnedin, N.Y., Silk J.: ApJl. **493**, L1 (1998)
18. Baraffe, I., Heger, A., Woosley S.E.: ApJ. **550**, 890 (2001)
19. Bardeen, J.M. Bond, J.R., Kaiser, N., Szalay A.S.: ApJ. **304**, 15 (1986)
20. Barger, A.J., et al.: ApJ. **584**, L61 (2003)
21. Barkana, R., Haiman, Z., Ostriker J.P.: ApJ. **558**, 482 (2001)
22. Barkana, R., Loeb A.: ApJ. **523**, 54 (1999)
23. Barkana, R., Loeb A.: Phys. Rep. **349**, 125 (2001)
24. Barkana, R., Loeb A.: Nature **421**, 341 (2003)
25. Barkana R.: MNRAS **347**, 57 (2004)
26. Barkana, R., Loeb A.: ApJ. **601**, 64 (2004)
27. Barkana, R., Loeb A.: ApJ. **609**, 474 (2004)
28. Barkana, R., Loeb A.: ApJ. **624**, L65 (2005)
29. Barkana, R., Loeb A.: MNRAS **363**, L36 (2005)
30. Barkana, R., Loeb A.: ApJ. **626**, 1 (2005)
31. Bate, M.R., Bonnell, I.A., Bromm V.: MNRAS **332**, L65 (2002)
32. Bate, M.R., Bonnell, I.A., Bromm V.: MNRAS **339**, 577 (2003)
33. Battaglia, M., De Roeck, A., Ellis, J.R., Gianotti, F., Olive K.A., Pape L.: Eur. Phys. J. C **33**, 273 (2004) [arXiv:hep-ph/0306219].
34. Begelman, M.C., Blandford R.D., Rees M.J.: Rev. Mod. Phys. **56**, 255 (1984)
35. Begelman, M., Ruszkowski M.: Phil.Trans. Roy. Soc. Lond. **A363**, 655 (2005)
36. Bennett, C.L., et al.: ApJ. **464**, L1 (1996)
37. Benson, A.J., Frenk, C.S., Lacey, C.G., Baugh, C.M., Cole S.: MNRAS **333**, 177 (2002)
38. Benson, A.J., Frenk, C.S., Baugh, C.M., Cole, S., Lacey C.G.: MNRAS **343**, 679 (2003)
39. Benson, A.J., Hoyle, F., Torres, F., Vogeley M.S.: MNRAS **340**, 160 (2003)
40. Bertschingerl E.: ApJS. **58**, 39 (1985)
41. Bharadwaj, S., Ali S.S.: MNRAS **352**, 142 (2004); **356**, 1519 (2005)

42. Bianchi, S., Ferrara, A., Davies, J.I., Alton P.B.: MNRAS **311**, 601 (2000)
43. Binney, J., Tremaine S.: Galactic Dynamics (Princeton: Princeton University Press) (1987)
44. Birkhoff G.D.: Relativity and Modern Physics. Cambridge: Harvard University Press (1923)
45. Blain, A.W., Natarajan P.: MNRAS **312**, L35 (2000)
46. Bland-Hawthorn, J., Maloney P.R.: ApJ. **510**, L33 (1999)
47. Bloom, J.S., Kulkarni, S.R., Djorgovski S.G.: AJ. **123**, 1111 (2002)
48. Bode, P., Ostriker J.P., Turok N.: ApJ. **556**, 93 (2001); Barkana, R., Haiman, Z., Ostriker J.P.: ApJ. **558**, 482 (2001)
49. Boisse P.: A&A **228**, 483 (1990)
50. Bond, J.R., Arnett, W.D., Carr B.J.: ApJ. **280**, 825 (1984)
51. Bond, J.R., Szalay, A.S., Silk J.: ApJ. **324**, 627 (1988)
52. Bond, J.R., Cole, S., Efstathiou, G., Kaiser N.: ApJ. **379**, 440 (1991)
53. Bonnor W.B.: MNRAS **116**, 351 (1956)
54. Borgani, S., et al.: MNRAS **336**, 409 (2002)
55. Bower R.J.: MNRAS **248**, 332 (1991)
56. Boyle, B.J., et al.: ApJ. **317**, 1014 (2000)
57. Bromm, V., Coppi, P.S., Larson R.B.: ApJ. **527**, L5 (1999)
58. Bromm, V.: PhD thesis, Yale University (2000)
59. Bromm, V., Kudritzki, R.P., Loeb A.: ApJ. **552**, 464 (2001)
60. Bromm, V., Ferrara, A., Coppi, P.S., Larson R.B.: MNRAS **328**, 969 (2001)
61. Bromm, V., Loeb A.: ApJ. **575**, 111 (2002)
62. Bromm, V., Coppi, P.S., Larson R.B.: ApJ. **564**, 23 (2002)
63. Bromm, V., Loeb A.: ApJ. **596**, 34 (2003)
64. Bromm, V., Loeb A.: Nature **425**, 812 (2003)
65. Bromm, V., Yoshida, N., Hernquist L.: ApJ. **596**, L135 (2003)
66. Bromm, V., Larson R.B.: Ann. Rev. Astr. & Astrophys. **42**, 79 (2004)
67. Bromm, V., Loeb A.: NewA. 9, 353 (2004)
68. Bromm, V., Loeb A.: ApJ. **624**, 382 (2006a)
69. Bromm, V., Loeb A.: To appear in Proc. of "Gamma Ray Bursts in the Swift Era"; preprint (2006b), (astro-ph/0601216)
70. Bryan, G.L., Machacek, M., Anninos, P., Norman M.L.: ApJ. **517**, 12 (1999)
71. Bryan, G.L., Norman M.: ApJ. **495**, 80 (1998)
72. Bullock, J.S., Kolatt, T.S., Sigad, Y., Somerville, R.S., Kravtsov, A.V., Klypin, A.A., Primack, J.R., Dekel A.: MNRAS **321**, 559 (2001)
73. Bullock, J.S., Kravtsov, A.V., Weinberg D.H.: ApJ. **548**, 33 (2001)
74. Carr, B.J., Ikeuchi S.: MNRAS **213**, 497 (1985)
75. Carilli, C.L., Gnedin, N.Y., Owen F.: ApJ. **577**, 22 (2002)
76. Carilli, C.L., Gnedin, N.Y., Furlanetto, S., Owen F.: N. Astro. Rev., **48**, 1053 (2004)
77. Casas-Miranda, R., Mo, H.J., Sheth, R.K., Börner G.: MNRAS **333**, 730 (2002)
78. Castellani, V., Chieffi, A., Tornambe A.: ApJ. **272**, 249 (1983)
79. Cen, R., Ostriker J.P.: ApJ. **417**, 404 (1993)
80. Cen, Y., Ostriker J.P.: ApJ. **519**, L109 (1999)
81. Cen, R., Haiman Z.: ApJl. **542**, L75 (2000)
82. Cen, R., Ostriker J.P.: ApJ. **538**, 83 (2000)
83. Cen R.: ApJ. **591**, 12 (2003)
84. Cen, R., Miralda-Escudé, J., Ostriker, J.P., Rauch M.: ApJ. **437**, L9 (1994)

85. Cen, R., Haiman, Z., Mesinger A.: ApJ. **621**, 89 (2005)
86. Chatterjee, P., Hernquist, L., Loeb A.: ApJ. **572**, 371 (2002); Phys. Rev. Lett. **88**, 121103 (2002)
87. Chen, X., Kamionkowski, M., Zhang X.: Phys. Rev. D **64**, 021302 (2001)
88. Chen, X., Miralda-Escudé J.: ApJ. **602**, 1 (2004)
89. Chiu, W.A. Ostriker J.P.: ApJ. **534**, 507 (2000)
90. Christlieb, N., et al.: Nature **419**, 904 (2002)
91. Churazov, E., Sunyaev, R., Forman, W., Boehringer H.: MNRAS **332**, 729 (2002)
92. Ciardi, B., Ferrara, A., Abel T.: ApJ. **533**, 594 (2000)
93. Ciardi, B., Ferrara, A., White S.D.M.: MNRAS **344**, L7 (2003)
94. Ciardi, B., Loeb A.: ApJ. **540**, 687 (2000)
95. Ciardi, B., Stoehr, F., White S.D.M.: MNRAS **343**, 1101 (2003)
96. Ciardi, B., Madau P.: ApJ. **596**, 1 (2003)
97. Cole, S., Kaiser N.: MNRAS **237**, 1127 (1989)
98. Couchman, H.M.P., Rees M.J.: MNRAS **221**, 53 (1986)
99. Cowie, L.L., Songaila A.: Nature, **394**, 44 (1998)
100. Croft, R., et al.: ApJ. **580**, 634 (2002)
101. Croom, S.M., et al.: MNRAS **322**, 29 (2001); **325**, 483 (2001); **335**, 459 (2002)
102. Davé, R., Hernquist, L., Katz, N., Weinberg D.: ApJ. **511**, 521 (1999)
103. Daly, R.A., Loeb A.: ApJ. **364**, 451 (1990)
104. Dekel, A., Silk J.: ApJ. **303**, 39 (1986)
105. Diemand, J., Moore, B., Stadel J.: Nature **433**, 389 (2005)
106. Dietrich, M., et al.: ApJ. **589**, 722 (2003)
107. Dijkstra, M., Haiman, Z., Loeb A.: ApJ. **613**, 646 (2004)
108. Di Matteo, T., Springel, V., Hernquist L.: Nature **433**, 604 (2005)
109. Dodelson S.: Modern Cosmology, (Acacdemic Press: San Diego, CA), (2003)
110. Dove, J.B., Shull, J.M., Ferrara A.: (1999) (astro-ph/9903331)
111. Dove, J.B., Shull J.M.: ApJ. **430**, 222 (1994)
112. Draine, B.T., Hao L.: ApJ. **569**, 780 (2002)
113. Dwek, E., Arendt, R.G., Krennrich F.: ApJ. **635**, 784 (2005)
114. Ebert R.: Z. Astrophysik, **37**, 217 (1955)
115. El Eid, M.F., Fricke, K.J., Ober W.W.: A&A **119**, 54 (1983)
116. Efstathiou, G., Rees M.J.: MNRAS **230**, P5 (1988)
117. Efstathiou G.: MNRAS **256**, 43 (1992)
118. Eisenstein, D.J., Hu W.: ApJ. **511**, 5 (1999)
119. Eisenstein, D.J., et al.: ApJ. **633**, 560 (2005)
120. El-Ad, H., Piran T.: MNRAS **313**, 553 (2000)
121. Ellison, S., Songaila, A., Schaye, J., Petinni M.: AJ. **120**, 1175 (2001)
122. Elvis, M., et al.: ApJS. **95**, 1 (1994)
123. Fabian A.: Phil. Trans. Roy. Soc. Lond. **A363**, 725 (2005)
124. Fan, X., et al.: AJ. **120**, 1167 (2000)
125. Fan, X., et al.: AJ. **121**, 54 (2001); AJ. **122**, 2833 (2001); AJ. **125**, 1649 (2003)
126. Fan, X., et al.: AJ. **123**, 1247 (2002)
127. Fan, X., et al.: AJ. **128**, 515 (2004)
128. Fan, X., et al.: preprint (2005), (astro-ph/0512082)
129. Ferrara, A., Pettini, M., Shchekinov Y.: MNRAS **319**, 539 (2000)
130. Ferrarese L.: ApJ. **578**, 90 (2002)
131. Field G.B.: Proc. IRE **46**, 240 (1958); Field G.B.: Astrophys. J. **129**, 536 (1959); Field G.B.: ApJ. **129**, 551 (1959)

132. Fillmore, J.A., Goldreich P.: ApJ. **281**, 1 (1984)
133. Fisher, K.B., Scharf, C.A., Lahav O.: MNRAS **266**, 219 (1994)
134. Floyd D.J.E.: preprint (2003), (astro-ph/0303037)
135. Ford, H.C., et al.: ApJ. **435**, L27 (1994)
136. Fruchter, A., Krolik, J.H., Rhoads J.E.: ApJ. **563**, 597 (2001)
137. Fryer, C.L., Woosley, S.E., Heger A.: ApJ. **550**, 372 (2000)
138. Fukugita, M., Kawasaki M.: MNRAS **269**, 563 (1994)
139. Furlanetto, S.R., Loeb A.: ApJ. **556**, 619 (2001)
140. Furlanetto, S.R., Loeb A.: ApJ. **579**, 1 (2002)
141. Furlanetto, S.R., Loeb A.: ApJ. **588**, 18 (2003)
142. Furlanetto, S., Sokasian, A., Hernquist L.: MNRAS **347**, 187 (2004)
143. Furlanetto, S.R., Zaldarriaga, M., Hernquist L.: ApJ. **613**, 1 (2004)
144. Furlanetto S.R., Loeb A.: ApJ. **634**, 1 (2005)
145. Furlanetto, S.R., Zaldarriaga, M., Hernquist L.: MNRAS **365**, 1012 (2006)
146. Gao, L., White, S.D.M., Jenkins, A., Stoehr, F., Springel V.: MNRAS **355**, 819 (2004)
147. Gehrels, N., et al.: ApJ. **611**, 1005 (2004)
148. Gnedin, N.Y., Ostriker J.P.: ApJ. **486**, 581 (1997)
149. Gnedin N.Y.: MNRAS **294**, 407 (1998)
150. Gnedin, N.Y., Hui L.: MNRAS **296**, 44 (1998)
151. Gnedin, N.Y., Ferrara, A., Zweibel E.G. : ApJ. **539**, 505 (2000)
152. Gnedin N.Y.: ApJ. **535**, 530 (2000a)
153. Gnedin N.Y.: ApJ. **542**, 535 (2000b)
154. Gnedin N.Y., Shaver P.A.: ApJ. **608**, 611 (2004)
155. Goodman J.: Phys. Rev. **D52**, 1821 (1995)
156. Gott J.R.: ApJ. **201**, 296 (1975)
157. Gottlöber, S., Łokas, E.L., Klypin, A., Hoffman Y.: MNRAS **344**, 715 (2003)
158. Green, A.M., Hofmann, S., Schwarz D.J.: MNRAS **353**, L23 (2004)
159. Green, A.M., Hofmann, S., Schwarz D.J.: preprint (2005), (astro-ph/0503387)
160. Green, P.J., et al.: ApJ. **558**, 109 (2001); Chartas, G., et al.: ApJ. **579**, 169 (2002); Gallagher, S.C., et al.: ApJ. **567**, 37 (2002); King, A.R., Pounds K.A.: MNRAS **345**, 657 (2003); Pounds, K.A., et al.: MNRAS **346**, 1025 (2003)
161. Gould A.: ApJl. **386**, 5 (1992); Stanek, K.Z., Paczynski B., Goodman J.: ApJl. **413**, 7 (1993); Ulmer A., Goodman J.: ApJ. **442**, 67 (1995)
162. Gruzinov, A., Hu W.: ApJ. **508**, 435 (1998)
163. Gunn, J.E., Peterson B.A.: ApJ. **142**, 1633–1641 (1965)
164. Gunn, J.E., Gott J.R.: ApJ. **176**, 1 (1972)
165. Gunn J.E.: ApJ. **218**, 592 (1977)
166. Haardt, F., Madau P.: ApJ. **461**, 20 (1996)
167. Haiman, Z., Thoul, A., Loeb A.: ApJ. **464**, 523 (1996a)
168. Haiman, Z., Rees, M.J., Loeb A.: ApJ. **476**, 458 (1997); erratum, **484**, 985
169. Haiman, Z., Loeb A.: ApJ. **483**, 21 (1997); erratum – ApJ. **499**, 520 (1998)
170. Haiman, Z., Loeb A.: ApJ. **503**, 505 (1998)
171. Haiman, Z., Spaans M.: ApJ. **518**, 138 (1999)
172. Haiman, Z., Abel, T., Madau P.: ApJ. **551**, 599 (2000)
173. Haiman, Z., Loeb A.: ApJ. **552**, 459 (2001)
174. Haiman, Z., Holder G.P.: ApJ. **595**, 1 (2003)
175. Haiman, Z., Cen R.: ApJ. **623**, 627 (2005)
176. Hamilton A.: ApJ. **385**, L5, (1992)

177. Haislip, J., et al.: Nature, **440**, 181, (2006), (astro-ph/0509660)
178. Hernquist, L., Katz, N., Weinberg, D.H., Miralda-Escudé J.: ApJ. **457**, L51 (1996)
179. Hoffman, Y., Shaham J.: 3. ApJ. **297**, 16 (1985)
180. Hofmann, S., Schwarz, D.J., Stöker H.: Phys. Rev. D **64**, 083507 (2001)
181. Hoopes, C.G., Walterbos, R.A.M., Rand R.J.: ApJ. **522**, 669 (1999)
182. Hu, E.M., et al.: ApJ. **568**, 75 (2002) erratum **576**, 99
183. Hurwitz, M., Jelinsky, P., Dixon W.: ApJ. **481**, L31 (1997)
184. Iliev, I.T., Mellema, G., Pen, U.L., Merz, H., Shapiro, P.R., Alvarez M.A.: preprint (2005), astro-ph/0512187
185. Jeans J.H.: Astronomy and Cosmogony (Cambridge: Cambridge University Press) (1928)
186. Jenkins, A., et al.: MNRAS **321**, 372 (2001)
187. Jungman, G., Kamionkowski, M., Griest K.: Phys. Rep. **267**, 195 (1996); Bergstrom, L.: Rep. Prog. Phys. **63**, 793 (2000); Gaitskell, R.J., Ann. Rev. of Nuclear and Particle Science **54**, 315 (2004)
188. Kaiser N.: ApJ. **284**, L9, (1984)
189. Kaiser N.: MNRAS **227**, 1, (1987)
190. Kamionkowski, M., Liddle A.R.: Phys. Rev. Lett. **84**, 4525 (2000)
191. Kashlinsky, A., Arendt, R.G., Mather, J., Moseley S.H.: Nature **438**, 45 (2005)
192. Kassim N.E., Weiler K.W.: Low Frequency Astrophysics from Space, Springer-Verlag: New-York, (1990); Stone R.G. et al.: Radio Astronomy at Long Wavelengths American Geophysical Union: Washington DC, (2000); see also http://rsd-www.nrl.navy.mil/7213/weiler/lfraspce.html
193. Kauffmann, G., White S.D.M.: MNRAS **261**, 921 (1993)
194. Kauffmann, G., White, S.D.M., Guiderdoni B.: MNRAS **264**, 201 (1993)
195. Kauffmann, G., Haehnelt M.: MNRAS **311**, 576 (2000)
196. Kearns E.T.: Frascati Phys. Ser. **28** 413 (2002); [hep-ex/0210019] Bahcall, J.N., Pena-Garay C.: JHEP **0311**, 004 (2003)
197. King A.: ApJ. **596**, L27 (2003)
198. Kirshner, R.P., Oemler, A., Schechter, P.L., Shectman S.A.: ApJL **248**, L57 (1981)
199. Kitayama, T., Ikeuchi S.: ApJ. **529**, 615 (2000)
200. Kodaira, K., et al.: PASJ **55**(2) L17 (2003)
201. Kogut, A., et al.: ApJS **148**, 161 (2003)
202. Kohler, K., Gnedin, N.Y., Hamilton A.J.S.: ArXiv Astrophysics e-prints, (2005), arXiv:(astro-ph/0511627)
203. Kolb, E.W., Turner M.S.: The Early Universe (Redwood City, CA: Addison-Wesley) (1990)
204. Kormendy J.: preprint (2003), (astro-ph/0306353)
205. Kronberg, P.P., et al.: ApJ. **560**, 178 (2001)
206. Kroupa P.: Science **295**, 82 (2002)
207. Kudritzki, R.P., et al.: ApJ. **536**, 19 (2000)
208. Kudritzki R.P.: ApJ. **577**, 389 (2002)
209. Kulkarni, S.R., et al.: Proc. SPIE, **4005**, 9 (2000)
210. Lacey, C.G., Cole S.: MNRAS **262**, 627 (1993)
211. Lacey, C.G., Cole S.: MNRAS **271**, 676 (1994)
212. Lamb, D.Q., Reichart D.E.: ApJ. **536**, 1 (2000)

213. Larson R.: In Proc. of the 33rd ESLAB Symposium, Star Formation from the Small to the Large Scale, Noordwijk, The Netherlands, November 2–5, 1999, ESA Special Publications Series (SP-445), edited by Favata, F., Kaas, A.A., and Wilson A. (1999), (astro-ph/9912539)
214. Larson R.B.: MNRAS **332**, 155 (2002)
215. Leitherer, C., Ferguson, H.C., Heckman, T.M., Lowenthal J.D.: ApJ. **452**, 549 (1995)
216. Liddle, A.R., Lyth D.H.: *Cosmological Inflation and Large-Scale Structure*, Cambridge University Press: Cambridge, (2000)
217. Lilje, P.B. Efstathiou G.: MNRAS **236**, 851 (1989)
218. Loeb, A., To appear in "Physica Plus", (2007) astro-ph/0702298
219. Loeb, A., Rasio F.: ApJ. **432**, 52 (1994)
220. Loeb A.: In E. Smith, A. Koratkar (Eds.), ASP Conf. Series Vol. 133, Science With The Next Generation Space Telescope, ASP, San Francisco, p 73 (1998) (astro-ph/9704290)
221. Loeb, A., Haiman Z.: ApJ. **490**, 571 (1997)
222. Loeb, A., Rybicki G.B.: ApJ. **524**, 527 (1999)
223. Loeb, A., Barkana R.: Ann. Rev. Astron. & Ap. **39**, 19 (2001)
224. Loeb, A., Peebles P.J.E.: ApJ. **589**, 29 (2003)
225. Loeb, A., Zaldarriaga M.: Phys. Rev. Lett., **92**, 211301 (2004)
226. Loeb, A., Barkana, R., Hernquist L.: ApJ. **620**, 553 (2005)
227. Loeb, A., Zaldarriaga M.: Phys. Rev. **D71**, 103520 (2005)
228. Lu, L., Sargent, W., Barlow, T.A., Rauch M.: astroph/9802189, (1998)
229. Ma, C., Bertschinger E.: ApJ. **455**, 7 (1995)
230. MacFadyen, A.I. Woosley, S.E., Heger A.: ApJ. **550**, 410 (2001)
231. Machacek, M.E., Bryan, G.L., Abel T.: ApJ. **548**, 509 (2001)
232. Machacek, M.E. Bryan, G.L., Abel T.: MNRAS **338**, 27 (2003),
233. Mackey, J., Bromm, V., Hernquist L.: ApJ. **586**, 1 (2003)
234. Mac Low, M.-M., Ferrara A.: ApJ. **513**, 142 (1999)
235. Madau, P., Shull J.M.: ApJ. **457**, 551 (1996)
236. Madau, P., Meiksin, A., Rees M.J.: ApJ. **475**, 429 (1997)
237. Madau P.: In the proceedings of the 9th Annual October Astrophysics Conference in Maryland, After the Dark Ages: When Galaxies were Young, edited by Holt S.S., and Smith, E.P. (1999), astro-ph/9901237
238. Madau, P., Haardt, F., Rees M.J.: ApJ. **514**, 648 (1999)
239. Madau, P., Rees M.J.: ApJl. **542**, L69 (2000)
240. Madau, P., Ferrara, A., Rees M.J.: ApJ. **555**, 92 (2001)
241. Madau, P., Silk J.: MNRAS **359**, L37 (2005)
242. Magorrian, J., et al.: AJ. **115**, 2285 (1998)
243. Maier, C., et al.: A&A **402**, 79 (2003)
244. Maldacena J.M.: JHEP **0305**, 013 (2003)
245. Malhotra, S., Rhoads J.: ApJL. **647**, 95 (2006), (astro-ph/0511196)
246. Martin C.L.: ApJ. **513**, 156 (1999)
247. Martini P.: preprint (2003), (astro-ph/0304009)
248. Mathis, H., White S.D.M.: MNRAS **337**, 1193 (2002)
249. McQuinn, M. Zahn, O., Zaldarriaga, M., Hernquist, L., Furlanetto S.R.: ApJ. **653**, 815 (2006), preprint (astro-ph/0512263)
250. Mesinger, A., Haiman Z.: ApJ. **611**, L69 (2004)
251. Meyer, D.M., York D.G.: ApJ. **315**, L5 (1987)

252. Milgrom, M.: New Astron. Rev. **46**, 741 (2002); Bekenstein, J.D.: Phys. Rev. D **70**, 083509 (2004)
253. Miralda-Escudé J.: ApJ. **501**, 15 (1998)
254. Miralda-Escudé, J., Rees M.J.: ApJ. **497**, 21 (1998)
255. Miralda-Escudé, J., Haehnelt, M., Rees M.J.: ApJ. **530**, 1 (2000)
256. Miralda-Escudé J.: ApJ. **597**, 66–73, (2003)
257. Mo, H.J., White S.D.M.: MNRAS **282**, 347 (1996)
258. Morales, M.F., Hewitt J.: ApJ. **615**, 7 (2004)
259. Morales M.F.: ApJ. **619**, 678 (2005)
260. Mori, M., Ferrara, A., Madau P.: ApJ. **571**, 40 (2002)
261. Murray, N., Quataert, E., Thompson T.A.: ApJ. **618**, 569 (2004)
262. Natarajan, P., Sigurdsson, S., Silk J.: MNRAS **298**, 577 (1998)
263. Natarajan, P., Albanna, B., Hjorth, J., Ramirez-Ruiz, E., Tanvir, N., Wijers R.A.M.J.: MNRAS **364**, L8 (2005)
264. Nath, B.B. Trentham N.: MNRAS **291**, 505 (1997)
265. Navarro, J.F., Frenk, C.S., White S.D.M.: ApJ. **490**, 493 (NFW) (1997)
266. Navarro, J.F., Steinmetz M.: ApJ. **478**, 13 (1997)
267. Navarro, J.F., et al.: MNRAS **349**, 1039 (2004)
268. Neufeld D.A.: ApJl. **370**, L85 (1991)
269. Oh P.S.: ApJ. **527**, 16 (1999)
270. Oh, P.S., Haiman, Z., Rees M.: ApJ. **553**, 73 (2001)
271. Oh, S.P., Haiman Z.: MNRAS **346**, 456 (2003)
272. Oh, S.P., Scannapieco E.: ApJ. **608**, 62 (2004)
273. Omukai, K., Nishi R.: ApJ. **508**, 141 (1998)
274. Omukai K.: ApJ. **534**, 809 (2000)
275. Omukai, K., Palla F.: ApJ. **561**, L55 (2001)
276. Omukai K., Inutsuka S.: MNRAS **332**, 59 (2002)
277. Ostriker, J., Cowie L.: ApJL. **243**, 1270 (1981)
278. Ostriker, J.P., McKee C.F.: Rev. Mod. Phys. **60**, 1 (1988)
279. Palla, F., Salpeter, E.E., Stahler S.W.: ApJ. **271**, 632 (1983)
280. Peebles, P.J.E., Dicke R.H.: ApJ. **154**, 891 (1968)
281. Peebles, P.J.E., Yu J.T.: ApJ. **162**, 815 (1970)
282. Peebles P.J.E.: The Large-Scale Structure of the Universe (Princeton: PUP) (1980)
283. Peebles P.J.E.: Principles of Physical Cosmology, pp. 176, 177 (Princeton: PUP) (1993)
284. Peebles P.J.E.: ApJ. **557**, 495 (2001)
285. Pei Y.C.: ApJ. **438**, 623 (1995)
286. Pen U.L.: New Astron. **9**, 417 (2004)
287. Pentericci, L., et al.: AJ. **123**, 2151–2158 (2002)
288. Perna, R., Loeb A.: ApJ. **501**, 467 (1998)
289. Pettini, M., Kellogg, M., Steidel, C.C., Dickinson, M., Adelberger, K.L., Giavalisco M.: ApJ. **508**, 539 (1998)
290. Press, W.H., Schechter P.: ApJ. **193**, 437 (1974)
291. Press, W.H., Schechter P.: ApJ. **187**, 425 (1974)
292. Pudritz R.E.: Science **295** 68 (2002)
293. Quinn, T., Katz, N., Efstathiou G.: MNRAS **278**, L49 (1996)
294. Rand R.J.: ApJ. **462**, 712 (1996)
295. Razoumov, A.O., Scott D.: MNRAS **309**, 287 (1999)

296. Rees, M.J., Sciama D.W.: Nature **217**, 511 (1968)
297. Rees M.J.: MNRAS **176**, 483 (1976)
298. Rees M.J.: MNRAS **222**, 27 (1986)
299. Reynolds, R.J., Tufte, S.L., Kung, D.T., McCullough, P.R., Heiles C.R.: ApJ. **448**, 715 (1995)
300. Rhoads, J.E., et al.: AJ. **125**, 1006 (2003)
301. Ricotti, M., Shull J.M.: ApJ. **542**, 548 (2000)
302. Ricotti, M., Gnedin, N.Y., Shull M.J.: ApJ. **575**, 49 (2002)
303. Ripamonti, E., Haardt, F., Ferrara, A., Colpi M.: MNRAS **334**, 401 (2002)
304. Rix, H.W., et al.: preprint (1999), (astro-ph/9910190)
305. Rybicki, G.B., Lightman A.P.: Radiative Processes in Astrophysics, pp. 29–32. Wiley, New York (1979)
306. Rybicki, G.B., Loeb A.: ApJl. **520**, L79 (1999)
307. Sachs, R.K., Wolfe A.M.: ApJ. **147**, 73 (1967)
308. Saijo, M., Baumgarte, T.W., Shapiro, S.L., Shibata M.: ApJ. **569**, 349 (2002); Saijo M.: ApJ. **615**, 866 (2004)
309. Salvaterra, R., Ferrara A.: MNRAS **339**, 973 (2003)
310. Santos, M.R., Bromm, V., Kamionkowski M.: MNRAS **336**, 1082 (2002)
311. Santos M.R.: MNRAS **349**, 1137 (2003)
312. Santos, M.G., Cooray, A., Haiman, Z., Know, L., Ma C.P.: ApJ. **598**, 756 (2003)
313. Santos, M.G., Cooray, A., Knox L.: ApJ. **625**, 575 (2004)
314. Scalo J.: ASP conference series Vol 142, The Stellar Initial Mass Function, eds. Gilmore G., Howell, D., p. 201 (San Francisco: ASP) (1998)
315. Scalo J., Wheeler J.C.: ApJ. **566**, 723 (2002)
316. Scannapieco, E., Broadhurst T.: ApJ. **549**, 28 (2001)
317. Scannapieco, E., Ferrara, A., Broadhurst T.: ApJ. **536**, 11 (2000)
318. Scannapieco, E., Madau, P., Woosley, S., Heger, A., Ferrara A.: ApJ. **633**, 1031 (2005)
319. Schaye, J., Aguirre, A., Kim, T.-S., Theuns, T., Rauch, M., Sargent W.L.W.: ApJ. **596**, 768 (2003)
320. Schmid, C., Schwarz, D.J., Widerin P.: Phys. Rev. D. **59**, 043517 (1999)
321. Schneider, R., Ferrara, A., Natarajan, P., Omukai K.: ApJ. **571**, 30 (2002)
322. Scott, D., Rees M.J.: MNRAS **247**, 510 (1990)
323. Sedov L.I.: Similarity and Dimensional Methods in Mechanics (New York: Academic) (1959)
324. Sedov L.I.: Similarity and Dimensional Methods in Mechanics (10th ed.; Boca Raton: CRC) (1993)
325. Seljak, U., Zaldarriaga M.: Phys. Rev. Lett. **82**, 2636, (1999); Hu W.: ApJ. **556**, 93 (2001); Hirata C.M., Seljak U.: Phys. Rev. D **67**, 043001(2003)
326. Sesana, F., Haardt, F., Madau, P., Volonteri M.: ApJ. **611**, 623 (2004)
327. Shandarin S.: Astrofizika **16**, 769 (1980)
328. Shapiro, P.R., Giroux M.L.: ApJ. **321**, L107 (1987)
329. Shapiro, P.R., Giroux, M.L., Babul A.: ApJ. **427**, 25 (1994)
330. Shapiro, P.R., Raga A.C.: preprint of the contribution to The Seventh Texas-Mexico Conference on Astrophysics: Flows, Blows, and Glows, eds. W. Lee and S. Torres-Peimbert (astro-ph/0006367) (2000)
331. Sheth R.K.: MNRAS **300** 1057 (1998)
332. Sheth, R.K., Tormen G.: MNRAS **308**, 119 (1999)
333. Sheth, R.K., Mo, H.J., Tormen, G.: MNRAS **323**, 1 (2001)

334. Sheth, R.K., Tormen G.: MNRAS **329**, 61 (2002)
335. Sheth, R.K., et al.: ApJ. **594**, 225 (2003)
336. Shu, F.H., Lizano, S., Galli, D., Cantó, J., Laughlin G.: ApJ. **580**, 969 (2002)
337. Silk J.: ApJ. **151**, 459 (1968)
338. Silk, J., Rees M.: A&A Lett. **331**, L1 (1998)
339. Simcoe, R.A., Sargent, W.L.W., Rauch M.: ApJ. **606**, 92 (2004)
340. Sobolev V.V.: Moving Envelopes of Stars, Cambridge: Harvard University Press (1960)
341. Sokasian, A., Abel, T., Hernquist, L., Springel V.: MNRAS **344**, 607 (2003)
342. Sokasian, A., Yoshida, N., Abel, T., Hernquist, L., Springel V.: MNRAS **350**, 47 (2004)
343. Somerville R.S.: ApJ. **572**, L23 (2002)
344. Somerville, R.S., Livio M.: ApJ. **593**, 611 (2003)
345. Songaila A.: ApJ. **490**, L1 (1997)
346. Songaila, A., Cowie L.L.: AJ. **112** 335 (1996)
347. Spergel, D.N., et al.: ApJ. Suppl. **148**, 175 (2003)
348. Springel, V., Hernquist L.: MNRAS **339**, 312 (2003)
349. Springel, V., Di Matteo, T., Hernquist L.: ApJl **620**, L79 (2005)
350. Stanek, K.Z., et al.: ApJ. **591**, L17 (2003)
351. Steidel, C.C., Pettini, M., Adelberger K.L.: ApJ. **546**, 665 (2000)
352. Stoehr, F., White, S.D.M., Springel, V., Tormen, G., Yoshida N.: MNRAS **345**, 1313 (2003)
353. Storey, P.J., Hummer D.G.: MNRAS **272**, 41 (1995)
354. Sunyaev, R.A., Zeldovich Y.B.: APSS **7**, 3 (1970)
355. Tacconi, L.J., et al.: ApJ. **580**, 73 (2002)
356. Tan, J.C., McKee C.F.: ApJ. **603**, 383 (2004)
357. Tegmark, M., Silk, J., Evrard A.: ApJ. **417**, 54 (1993)
358. Tegmark, M., Silk, J., Rees, M.J., Blanchard, A., Abel, T., Palla F.: ApJ. **474**, 1 (1997)
359. Tegmark, M., et al.: Phys. Rev **D69**, 103501 (2004)
360. Thacker, R.J., Scannapieco, E., Davis M.: ApJ. **581**, 836 (2002)
361. Theuns, T., Mo, H.J., Schaye J.: MNRAS **321**, 450 (2001)
362. Thoul, A.A., Weinberg D.H.: ApJ. **465**, 608 (1996)
363. Todini, P., Ferrara A.: MNRAS **325**, 726 (2001)
364. Totani T.: ApJ. **486**, L71 (1997)
365. Tozzi, P., Madau, P., Meiksin, A., Rees M.J.: ApJ. **528**, 597 (2000)
366. Tozzi, P., Madau, P., Meiksin, A., Rees M.J.: ApJ. **528**, 597 (2000); Gnedin, N.Y., Shaver P.A.: ApJ. **608**, 611 (2003)
367. Tremaine, S., et al.: ApJ. **574**, 740 (2002)
368. Tully, R.B., Somerville, R.S., Trentham, N., Verheijen M.A.W.: ApJ. **569**, 573 (2002)
369. Tumlinson, J., Shull J.M.: ApJ. **528**, L65 (2000)
370. Tumlinson, J., Venkatesan, A., Shull J.M.: ApJ. **612**, 602 (2004)
371. Tytler, D., et al.: In Meylan, G. (ed.) QSO Absorption Lines, p. 289. ESO Astrophysics Symposia; Springer Heidelberg, (1995)
372. Valageas, P., Silk J.: A&A **347**, 1 (1999)
373. Voit G.M.: ApJ. **465**, 548 (1996)
374. Waxman, E., Draine B.T.: ApJ. **537**, 796 (2000)
375. Weinberg S.: Gravitation and Cosmology. Wiley, New York (1972)

376. Wechsler, R.H., Bullock, J.S., Primack, J.R., Kravtsov, A.V., Dekel A.: ApJ, **568**, 52 (2002)
377. Weinberg, D.H., Hernquist, L., Katz N.: ApJ. **477**, 8 (1997)
378. White S.D.M., Rees M.J.: MNRAS **183**, 341 (1978)
379. White, S.D.M., Springel V.: preprint of the contribution to the 1999 MPA/ESO workshop, The First Stars, eds. Weiss, A., Abel, T., Hill V. (astro-ph/9911378) (1999)
380. White, R.L., Becker, R.H., Fan, X., Strauss M.A.: Probing the ionization state of the universe at $z > 6$, AJ **126**, 1–14, (2003)
381. White S.D.M.: Les Houches Lectures, MPA preprint 831 (1994)
382. Wijers, R.A.M.J., Bloom, J.S., Bagla, J.S., Natarajan P.: MNRAS **294**, L13 (1998)
383. Willott, C.J., McLure, R.J., Jarvis M.J.: ApJl **587**, L15 (2003)
384. Witt, A.N., Gordon K.G.: ApJ. **463**, 681 (1996)
385. Wolfire, M.G., Cassinelli J.P.: ApJ. **319**, 850 (1987)
386. Wood, K., Loeb A.: ApJ. **545**, 86 (2000)
387. Wouthuysen S.A.: AJ. **57**, 31 (1952)
388. Wu, K.K.S., Lahav, O., Rees M.J.: Nature **397**, 225 (1999)
389. Wyithe, J.S.B., Loeb, A., Barnes D.G.: ApJ. **634**, 715 (2005)
390. Wyithe, J.S.B., Loeb A.: ApJ. **581**, 886 (2002)
391. Wyithe, J.S.B., Loeb A.: ApJ. **586**, 693 (2003)
392. Wyithe, J.S.B., Loeb A.: ApJ. **590**, 691 (2003)
393. Wyithe, J.S.B., Loeb A.: ApJ. **595**, 614 (2003)
394. Wyithe, J.S.B., Loeb A.: ApJ. **612**, 597 (2004)
395. Wyithe, J.S.B., Loeb A.: Nature **427**, 815 (2004)
396. Wyithe, J.S.B., Loeb A.: ApJ. **610**, 117 (2004)
397. Wyithe J.S.B., Loeb A.: ApJ. **621**, 95 (2004)
398. Wyithe, J.S.B., Loeb A.: Nature **432**, 194 (2004)
399. Wyithe, J.S.B., Loeb A.: ApJ. **625**, 1 (2005)
400. Wyithe, J.S.B., Loeb, A., Carilli C.: ApJ. **628**, 575 (2005)
401. Yoshida, N., Sokasian, A., Hernquist, L., Springel V.: ApJ. **598**, 73 (2003)
402. Yoshida, N., Bromm, V., Hernquist L.: ApJ. **605**, 579 (2004)
403. Zaldarriaga, M., Furlanetto S.R., Hernquist L.: ApJ. **608**, 622 (2004)
404. Zhang, Y., Anninos, P., Norman M.L.: ApJ. **453**, L57, (1995)
405. Zygelman B.: ApJ. **622**, 1356 (2005)

Cosmological Feedbacks from the First Stars

A. Ferrara

Abstract. It is very natural for us in a clear night to raise the head and look at the starry sky. We are so used to such emotion that we hardly imagine a dark sky, where no shimmering light attracts our constantly changing attention. However, for how peculiar this might seem to us, there was a time when no stars were born yet and the universe resembled a vast, quiet (and somewhat boring) sea of inert hydrogen and helium gas. Darkness was everywhere, as the light produced by the Big Bang rapidly shifted out of the visible range and into the infrared. Such situation lasted for many million years (now we speculate about 200 Myr).

The study of these remote times is a relatively new area in cosmology. These epochs can still be compared to the Old Wild West territories which only brave pioneers have dared to explore. The gold mines containing precious information about the dawn of the universe, when the first luminous sources brightened up and their light unveiled the already ongoing formation of large numbers of pregalactic systems, are still not at reach of the yet most powerful experimental devices currently available to us. This situation is going to change soon, but at the moment theorists' predictions remain in the realm of sophisticated (and intellectually exciting) speculations. For how uncertain all this might be considered, there is a great deal of knowledge that can be gathered even under these unfavorable conditions, as witnessed by the impressive momentum gained by the field in the very recent years. Thus, as observations are slowly, but steadily filling the gap with theory, this school very suitably serves the scope of setting a common framework among theory and observation. The present Lectures aim at discussing the current understanding of the properties of first cosmic stars. These stars had a dramatic impact on the surrounding environment regulated by a complex network of physical processes to which we – somewhat ambiguously – usually refer to as *feedback*.

Before the start, a short caveat on the suggested literature. It would be impossible to give credit to all the results and people whose efforts are constantly re-shaping the field in this brief essay. In fact we do not aim at giving a comprehensive view of all the aspects concerning the subject; for this we defer the reader to the most recent and technical reviews available: [4, 11, 15]. Instead, we hope that we will be able to outline the physics of the most relevant processes regulating the evolution of the first luminous sources and their effects.

A. Ferrara: *Cosmological Feedbacks from the First Stars*. In: D. Schaerer et al.: First Light in the Universe, Saas-Fee Advanced Courses, pp. 161–258 (2008)
DOI 10.1007/978-3-540-74163-3_2

1 Star Formation in Primordial Gas

Present-day gas contains a mass fraction in heavy elements that is about 2%. However, atoms, molecules (such as C^+, O, CO) and dust grains are very effective at radiating thermal energy and therefore this small fraction of "metals" (as heavy elements are usually called in the astrophysical jargon) can control the gas thermodynamics. Indeed the time scale for thermal equilibrium is much shorter than the typical dynamical time (essentially, the free fall time scale). As a result the equilibrium temperature in the typical molecular cloud hardly deviates from the standard one of approximately 10 K; stated differently, the equation of state of the gas can be considered as isothermal.

On the other hand, a gas of primordial composition essentially does not contain heavy elements. Because H and He atoms are very poor radiators for temperatures below $\approx 10^4$ K, if the gas were to remain purely atomic, the cloud would follow an almost adiabatic evolution.

Before we proceed to the collapse of primordial gas clouds, it is instructive to briefly recall the main cooling processes in primordial gas clouds; the corresponding rates can be found easily in the literature.

- *Radiative recombinations*: The thermal energy loss associated with the recombination of protons with electrons is caused by the photon emitted during the process. The recombined atom, generally in an excited state, eventually decays to the lowest energy level by emitting photons. Accordingly, the energy loss per recombination is the difference between the energy of the electrons in a bound state of the hydrogen atom and the kinetic energy of the free electron.
- *Collisional ionizations*: The collisional ionization of the hydrogen atom is also a cooling process. The energy loss in this process is related to the ionization potential energy: the thermal energy of electrons is converted into the ionization energy by this process.
- *Bound-bound transitions*: This is the most important cooling process around 10,000 K. Collisionally excited atoms emit radiation of energy equal to the energy difference between two levels when electrons decay. The level population must be determined by solving the detailed excitation/de-excitation balance equation rates for each level.
- *Bremsstrahlung emission*: Radiation due to the acceleration of a charge in Coulomb field of another charge is called as bremsstrahlung or free-free emission.

Knowing all (i.e. for both H and He) rates, and further assuming thermal ionization-recombination equilibrium (although this assumption is not adequate for some applications; full time-dependent equations must be solved to determine the ionization level in that case), we can estimate the cooling rate, $\Lambda(T)/n_{\rm H}^2$, where T and $n_{\rm H}$ are the gas temperature and hydrogen density, respectively. The cooling function is dominated by bremsstrahlung for $T > 10^{5.5}$ K and by collisional excitation of H and He in the range

$10^4 < (T/\mathrm{K}) < 10^{5.5}$. However, below 10^4 K the cooling rate drops off very sharply. This fact has very important consequences as we will see.

To make further progress we have to introduce three fundamental timescales. The first one is the cooling time, which is defined as

$$t_\mathrm{c} = \frac{3kT}{2n\Lambda(T)} \tag{1}$$

where n is the total gas number density. The second important time scale is the free-fall time:

$$t_\mathrm{ff} = \left(\frac{3\pi}{32G\rho}\right)^{-1/2}, \tag{2}$$

where G is the gravitational constant and ρ the total density of matter (i.e. including dark matter). Finally, the Hubble timescale must be considered:

$$t_\mathrm{H} = H(z)^{-1} = H_0^{-1}[\Omega_\Lambda + \Omega_\mathrm{m}(1+z)^3]^{-1/2}, \tag{3}$$

Ω_Λ and Ω_m being the nondimensional vacuum energy and matter density, respectively. The interplay among these three times governs both star and galaxy formation as we will see in the following. Let's examine some examples in more detail.

If the cooling time, t_c, is shorter than the typical dynamical time, t_ff, the thermal energy of a gas cloud can efficiently be carried away by radiation. Thus, the cloud undergoes an almost free fall collapse, since the pressure gradient force is negligible. In the opposite limit, if $t_\mathrm{ff} \ll t_\mathrm{c}$, the cloud collapses almost adiabatically; the pressure gradient force overwhelms the gravitational force. This descends from the fact that the effective "adiabatic index" of gravity (equal to 4/3 in a 3-dimensional collapse), is smaller than that of an ideal gas and the pressure gradient force can halt the gravitational collapse. As a result the cloud is virialized. After virialization, the cloud settles down in a quasi-equilibrium state and evolves within the cooling time scale. Finally, if the cooling time scale is longer than the Hubble time scale, t_H, at a given redshift, there has not been sufficient time for the cloud to contract.

From the above discussion we conclude that for primordial gas clouds with temperature below 10^4 K, the drop of the cooling function causes the cooling time to increase dramatically, thus preventing their collapse. The collapse condition required to form the first stars can then only be achieved thanks to another available coolant in the early universe: molecular hydrogen.

1.1 Cooling by Molecular Hydrogen

The process of H_2 radiative cooling is the key ingredient to allow primordial gas to cool below 10,000 K. The hydrogen molecules have energy levels corresponding to vibrational and rotational transitions. Vibrational transitions are

more important at high temperatures $10^3 < (T/\mathrm{K}) < 10^4$; rotational ones are more significant for $T < 10^3\,\mathrm{K}$. It is worth noting that there is a fundamental difference between H_2 and HI line cooling. In fact, Einstein's A-coefficients are much smaller for H_2, because hydrogen molecules have no dipole moment, resulting in correspondingly lower absorption coefficients. At the same time, the critical column density above which the cloud is optically thick becomes much larger. The small A-coefficients also affect the level population of the excited states. The collisional de-excitation process is effective for densities $> 10^4\,\mathrm{cm}^{-3}$, whereas the critical density for HI is higher than $10^{15}\mathrm{cm}^{-3}$.

Fitting formulae to the H_2 cooling function in the optically thin regime have been derived by several authors. This is the regime that is appropriate to the initial stages of the fragmentation/collapse process, when the H_2 fraction and the gas density are low. At later evolutionary stages (proto-stellar core phase, see Sect. 2) a proper treatment of the line radiative transfer becomes necessary. A useful approximation is given by the formula suggested by [33], according to which the cooling function is expressed as

$$\Lambda_{\mathrm{H}_2} = n_{\mathrm{H}_2}(n_H L^H_{\mathrm{vr}} + n_{\mathrm{H}_2} L^{\mathrm{H}_2}_{\mathrm{vr}}), \tag{4}$$

where the first term represents the contribution from H-H_2 collisional excitation and the second is due to H_2–H_2 collisions. The detailed expressions for the functions L^{H,H_2}_{vr} are rather cumbersome and can be found in the cited paper (see also [65]). The level populations are determined by the balance between the de-excitation rates from higher levels to lower levels and the excitation rates. The excitation rates are always dominated by the collisional excitation processes, whereas the de-excitation rates are determined by both collisional processes and radiative decay. The collisional de-excitation rates are proportional to n^2; radiative decay rates are proportional to n. It follows that the de-excitation rates are dominated by the collisional process at high density, and the radiative decay dominates the de-excitation process for low density. The density at which the collisional excitation rate equals the radiative decay one is called the *critical density*, n_{cr}; its value is $\approx 10^4\,\mathrm{cm}^{-3}$ for the lowest rotational transition. Thus, the number density of excited molecules is $\propto n^2$ for $n \ll n_{\mathrm{cr}}$ at low density and $\propto n$ for $n \gg n_{\mathrm{cr}}$. Finally, the emissivity ($n^2 L$) is the product of the number density of the excited molecules and the radiative transition rates. Other fitting formulae have been proposed by [27, 40] which are compared in Fig. 1. Note that while these approximations agree well in the high density regime they markedly differ at low densities.

In addition to collisional processes, additional cooling and heating are produced when H_2 molecules are either dissociated or formed. Hydrogen molecules have lower potential energy than the state of two separated neutral hydrogen atoms. Thus, the hydrogen molecules absorb the thermal energy of the colliding particles during the dissociation processes. The resulting cooling rate is

$$\Lambda_{\mathrm{diss}} = 7.16 \times 10^{-12} \left(\frac{\mathrm{d}n_{\mathrm{H}_2}}{\mathrm{d}t}\right)_{-} \ \mathrm{ergs}^{-1}\mathrm{cm}^{-3} \tag{5}$$

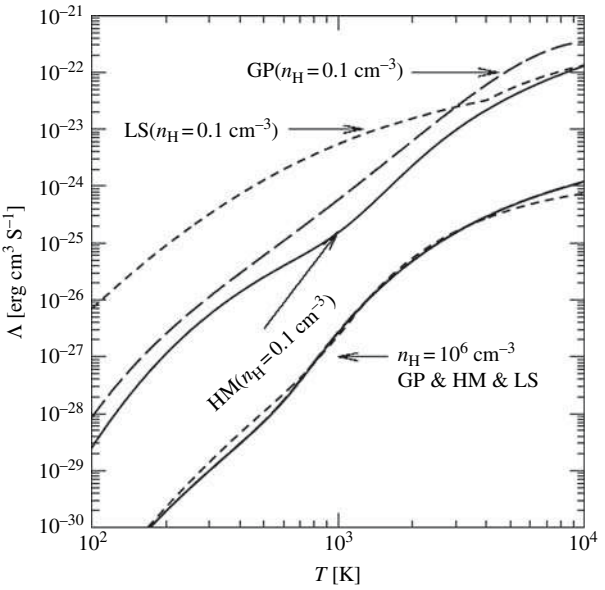

Fig. 1. H_2 cooling functions due to HI –H_2 collisions are plotted for two different densities. Solid/Long dashed/Short dashed curves refer to HM/GP/LP formulae, respectively (see text)

In the above equation, the term in parenthesis is the dissociation rate of H_2. As we will see next, collisional dissociation of hydrogen molecules occurs via three main channels: (i) H_2–H^+ collisions, H_2–HI collisions and (iii) H_2–H_2 collisions. The first channel is efficient when the gas is highly ionized; the other two dominate in low ionization environments.

The H_2 formation process is essentially the inverse of the dissociation one. The hydrogen molecules are in an excited state when they form. Hence, if the subsequent collisional de-excitation process is more efficient than the spontaneous decay process, the excitation energy goes into thermal energy of the surrounding particles via collisions. If the spontaneous decay rate is larger than the collisional de-excitation rate, the energy is instead radiated away by H_2 line emission, thus not contributing to the thermal budget. To account for this a density correction is usually applied as follows:

$$\Gamma_{\rm form} = 7.16 \times 10^{-12} \left(\frac{{\rm d}n_{\rm H_2}}{{\rm d}t}\right)_+ \frac{1}{1+(n_{\rm cr}/n_{\rm H})}\ {\rm ergs}^{-1}{\rm cm}^{-3}. \qquad (6)$$

1.2 Molecular Hydrogen Formation/Dissociation

Molecular hydrogen is a very fragile molecule and its abundance is sensitive to a large number of processes and environmental conditions. This is even more

true in the early universe, where dust grains (whose surface provides a favorable reaction site for HI atoms to combine and form H_2) are not present. Thus H_2 formation has to proceed using the much less efficient available gas phase reactions. The main reactions one needs to consider to obtain an accurate estimate of the H_2 abundance are listed in the following.

- H^- *channel*

$$H + e \rightarrow H^- + \gamma \tag{7}$$
$$H^- + H \rightarrow H_2 + e \tag{8}$$

- H^+ *channel*

$$H + H^+ \rightarrow H_2^+ + \gamma \tag{9}$$
$$H_2^+ + H \rightarrow H_2 + H^+ \tag{10}$$

- 3-*body channel*

$$3H \rightarrow H_2 + \gamma \tag{11}$$
$$2H + H_2 \rightarrow 2H_2 \tag{12}$$

- H-*collision channel*

$$H(n = 1) + H(n = 2) \rightarrow H_2 + \gamma \tag{13}$$

The first two processes are the most important ones simply because a dipole moment (provided by the ions) is necessary to form H_2 in two body reactions. We also notice that these two reactions require electrons or protons as catalysts, implying that the degree of ionization of the gas play a crucial role for the formation of H_2. In other words, a mechanism to provide a suitable ionization level is necessary for the formation of H_2 in the early universe. Three-body reactions become important at densities $n_H > 10^8\,cm^{-3}$, typical of the collapse phase of proto-stellar cores. Finally, the fourth process involves the collision between an excited HI atom and one in the ground state. The cross section of this reaction becomes much larger than that of the direct collision of hydrogen atoms in the ground state, due to the dipole moment produced by the difference of the electron energy levels. The significance of this channel is limited as it becomes important at high redshift ($z > 1000$), as CMB photons destroy H_2^+ and H^- at $z > 250$ and the main channels are inhibited. In addition, the higher CMB temperature provides the excitation source of HI . We now briefly turn to examine the dissociation channels:

- H-H_2 *impact*

$$H_2 + H \rightarrow 3H \quad \text{or} \tag{14}$$
$$2H_2 \rightarrow H_2 + 2H \tag{15}$$

These processes are important for $T > 2000\,K$ as at lower T the collisions are not sufficiently energetic to dissociate H_2.

- H^+-H_2 *impact*

$$H_2 + H^+ \rightarrow H_2^+ + H \tag{16}$$
$$H_2^+ + e \rightarrow 2H \tag{17}$$

  This channel is most important when the gas cools from a hot phase.
- e-H_2 *impact*

$$H_2 + e \rightarrow 2H + e \tag{18}$$

  Also this process is mainly active at high T, although it is sub-dominant with respect to the previous one.
- *Photodissociation*

$$H_2 + \gamma \rightarrow H_2^* \tag{19}$$
$$H_2^* \rightarrow 2H + \gamma \tag{20}$$

  This is the so-called Solomon process, which requires a two-step reaction. It is particularly important as it is responsible for one of the strongest feedbacks affecting early galaxy and star formation, the *radiative feedback*. We will devote considerable analysis to this physical process in Sect. 7.

We have stressed above that the presence of free electrons is fundamental to produce the necessary H_2 cooling allowing for the collapse of primordial structure. During the recombination epoch, the recombination time for hydrogen became increasingly long until it exceeded the Hubble time. At that point the electron fraction "freezed-out", reaching an almost constant value $y_e \approx 3 \times 10^{-4}$. Provided these free electrons, hydrogen molecules formed through the channels above; its abundance finally freezed-out at $z \approx 70$. The best estimates today give the following relic abundance, y_{H_2}, for this species (only very slightly dependent on cosmological parameters):

$$y_{H_2} = 1.1 \times 10^{-6}, z < 100, \tag{21}$$
$$= 10^{-7}, 100 < z < 250, \tag{22}$$
$$= 10^{-7}(1 + z/250)^{-14}, z > 250 \tag{23}$$

Although this amount of relic H_2 might have some interesting consequences per se, it is too tiny to be able to affect substantially the cooling of primordial objects. That is, the virial temperature of the objects that can cool is not noticeably decreased below 10^4 K, as if there were no H_2 at all.

1.3 Struggling for more Molecular Hydrogen

The way to increase the gas abundance of H_2 has to do with the collapse of structures. As dark matter perturbations turn-around and start to collapse under the gravitational pull, baryons follow into their potential wells. The

density increase due to the collapse and virialization boosts the formation rate of H_2 to levels at which the cooling and fragmentation of the proto-galaxy become possible. During this phase y_{H_2} increases from its relic abundance up to a typical value of $\approx 10^{-3}$. This can be understood as follows.

It is well known that the density evolution of a "top hat" perturbation evolves with redshift/time as

$$\frac{\rho}{\bar{\rho}} = \frac{9(\alpha - \sin\alpha)^2}{2(1-\cos\alpha)^3} \tag{24}$$

where $\bar{\rho}$ is the mean cosmic matter density at a given redshift and α is the "development angle", which is related to redshift z by

$$\frac{1+z_{\rm vir}}{1+z} = \frac{(\alpha - \sin\alpha)^{2/3}}{2\pi}; \tag{25}$$

$z_{\rm vir}$ is defined as the virialization epoch of the perturbation, i.e. the time at which the halo density stabilizes to the virial value $\rho_{\rm vir} = 18\pi^2\bar{\rho}$.

The above density evolution must be coupled with the energy and chemical equations for the gas:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{3kT}{2\mu m_{\rm p}} = \frac{p}{\rho^2}\frac{\mathrm{d}\rho}{\mathrm{d}t} - \Lambda, \tag{26}$$

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = \sum_j k_j y_j + n_{\rm H}\sum_{k,l} k_{k,l} y_k y_l + n_{\rm H}^2 \sum_{m,n,s} k_{m,n,s} y_m y_n y_s, \tag{27}$$

where $y_{\rm i} \equiv n_{\rm i}/n_H$ is the fraction of i-th species, and k is the rate coefficient of the corresponding reactions described above. Typically, one needs to take into account six species, namely e, H, H^+, H^-, H_2 and H_2^+ . Note, that for simplicity, in the previous equation we have neglected the chemical energy term due to variation of the number of particles, which is usually negligible but could become important under some conditions.

1.4 Enough for Collapse?

By solving the previous equations we can derive the chemical and thermal evolution of the gas residing in a given halo at a given redshift and assess if this is able to cool, fragment and form stars. The only calculation required to answer this question is to compare the cooling time (1) and the free-fall time (2). The results of this procedure are shown in Fig. 2. By equating $t_{\rm c}$ to $t_{\rm ff}$ we obtain the boundary in the redshift-virial temperature plane of the region in which objects are allowed to cool ($t_{\rm c} \ll t_{\rm ff}$). These systems will undergo a highly dynamical collapse, as their thermal pressure becomes negligible. Note that the cooling region expands into $T_{\rm vir} < 10^4\,{\rm K}$, i.e. differently from the classical Rees-Ostriker cooling diagram. It is also useful to compare the cooling time scale with the Hubble timescale given in (3). Halos allowed to cool within

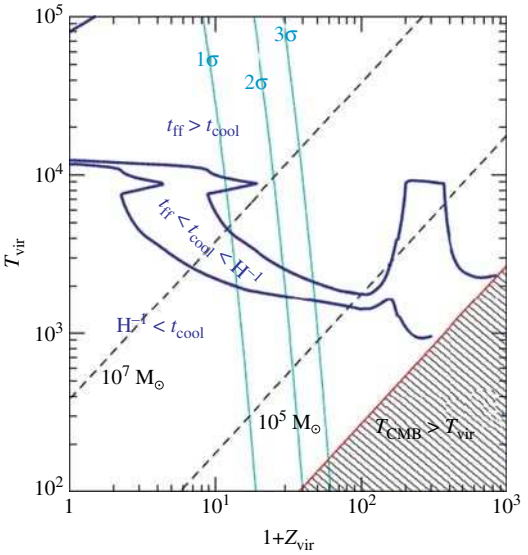

Fig. 2. Cooling diagram. The two thick *solid lines* divide the plane into three regions. The lower right region is forbidden from cooling, due to CMB Compton heating. *Dashed lines* denote loci of constant given mass. Also shown are 1σ, 2σ, 3σ fluctuations of the primordial density field for a concordance ΛCDM cosmological model

the Hubble time have slightly lower virial temperatures that those considered above; however, their collapse proceeds in a slow, quasi-statical manner. The bottom line is that H_2 cooling allows 2–3σ fluctuations at $z \approx 30$ to collapse. These correspond to halos with virial temperatures of about 3000 K, hence allowing a population of so-called minihalos to possibly form the first cosmic stars (unless radiative feedback is strong, see Sect. 7).

1.5 Additional Physics

The above picture contains the most important ingredient to investigate primordial star formation. However, many authors have suggested recently that at least is another important ingredient must be added. This is constituted by the deuterated hydrogen molecule (HD). HD is the second most abundant primordial molecule. Although HD is less abundant than H_2, its cooling efficiency is larger than H_2 because of its finite dipole moment $\mu_{HD} = 0.83$ debye and accordingly higher transition probabilities, which can compensate for its lower fractional abundance. The energy difference for the lowest transition is $\Delta E/k = 510\,\mathrm{K}$ and $\Delta E/k = 128\,\mathrm{K}$ for H_2 and HD, respectively. This lower transition energy enables HD to reduce the gas temperature to $T \lesssim 100\,\mathrm{K}$. Accordingly, the Jeans mass ($\propto T^{3/2}$) could be significantly reduced.

As for the H_2, HD is very suitably produced in situations in which the gas is rapidly cooling out of equilibrium, giving rise to a electron-rich, low temperature out-of-equilibrium physical conditions. These conditions are most easily created in shocked gas. Recent studies [34] have shown that at redshifts $z > 10$ HD line cooling allows strongly-shocked primordial gas to cool to the temperature of the cosmic microwave background (CMB), provided that the HD abundance exceeds 10^{-8}. This low temperature would decrease the characteristic mass of the second generation of stars (the first being necessary anyway to produce the shock-wave via their winds and supernova explosions) to a typical mass of about 10 $M_\odot$.

2 The Initial Mass Function

How massive where the first stars? How did their mass distribution look like? These are important questions that we are only starting to glimpse. However, already a considerable amount of studies has been carried out, which has clarified at least some of the many intricacies involved in the answers to the question above. We have to keep in mind, though, that very scarce experimental hints concerning the nature of the first stars are available (they are discussed in Sect. 4. Therefore, our progresses have to mainly be guided by theoretical studies. In the past 30 years, it has been argued that the first cosmological objects formed globular clusters, supermassive black holes, or even low-mass stars. This disagreement of theoretical studies might at first seem surprising. However, the first objects formed via the gravitational collapse of a thermally unstable reactive, optically thick medium, as we will see in this Lecture, which inhibits conclusive calculations. The problem is particularly acute because the evolution of all other cosmological objects (and in particular the larger galaxies that follow) depends on the evolution of the first stars.

The two above questions have been more recently tackled through numerical techniques. Results from these studies of the collapse and fragmentation of primordial gas clouds suggest that the first stars were predominantly very massive, with typical masses $M > 100$ $M_\odot$. Regardless of the detailed initial conditions, the primordial gas attains characteristic values of temperature and density corresponding to a characteristic fragmentation scale, given approximately by the Jeans mass, and are explained by the microphysics of H_2 cooling (see Sect. 1). Having constrained the characteristic mass scale, still leaves undetermined the overall range of stellar masses and the power-law slope which is likely to be a function of mass. In addition, it is presently not known whether binaries or, more generally, clusters of zero-metallicity stars, can form. This question has important implications for the star formation process. If Population III star formation typically resulted in a binary or multiple system, much of the progenitor cloud's angular momentum could go into the orbital motion of the stars in such a system. On the other hand, if an isolated formation mode were to predominate, the classical angular momentum barrier could be much

harder to breach. Answers to these questions share the pre-requisite that the physics of the proto-stellar collapse is deeply investigated and understood. In the following we will briefly describe the basic physics and current status of the research in this area.

2.1 Protostellar Collapse

The fundamental concepts of the proto-stellar collapse were layed down already almost 40 years ago by [37]. Initially there are no pressure gradients in the cloud, so the entire gas content starts to collapse in free fall; the gas is optically thin, and the compression energy generated by the collapse is freely radiated away producing an isothermal evolution. As a result, the density at the cloud boundary drops, while the interior one increases, thus establishing a pressure gradient. In turn, this causes the collapse in this region to be significantly retarded from a free-fall. The density therefore rises more rapidly in the center that the outer parts of the cloud and the density distribution becomes peaked at the center. The collapse of the central part of the cloud continues approximately as a free-fall (even if pressure gradients are not negligible), and since the free-fall time depends on $\rho^{-1/2}$, the collapse proceeds more rapidly at the center. The density distribution becomes more and more peaked at the center as the collapse proceeds.

The fact that density and velocity distributions approach constant limiting forms allow to describe the collapse via "similarity" solutions. To see how this can be accomplished, let's write the hydrodynamic equations in Eulerian form:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial r} + \frac{Gm}{r^2} + RT\frac{\mathrm{d}\ln\rho}{\mathrm{d}r} = 0 \tag{28}$$

$$\frac{\partial m}{\partial t} + 4\pi r^2 \rho u = 0 \tag{29}$$

$$\frac{\partial m}{\partial r} - 4\pi r^2 \rho = 0 \tag{30}$$

where the temperature T is kept constant, and m is the mass within a sphere of radius r. Searching for self-similar solutions of the type

$$u(r,t) = b(t)u_1(s) \tag{31}$$

$$\rho(r,t) = c(t)\rho_1(s) \tag{32}$$

$$m(r,t) = d(t)m_1(s) \tag{33}$$

where $s = r/a(t)$, by substitution in the equations above, one finds that acceptable solutions must satisfy

$$b(t) = 1, \quad c(t) = a(t)^{-2}, \quad d(t) = a(t); \tag{34}$$

with such prescription, the hydrodynamic equations reduce to a set of ODEs for the three functions $u_1(s), \rho_1(s), m_1(s)$. By further eliminating m_1, one can write the system

$$\left(\frac{s}{\tau}+u_1\right)\frac{\mathrm{d}u_1}{\mathrm{d}s}+4\pi G\rho_1(s+u_1\tau)+RT\frac{\mathrm{d}\ln\rho_1}{\mathrm{d}s}=0 \tag{35}$$

$$\frac{\mathrm{d}u_1}{\mathrm{d}s}+\left(\frac{s}{\tau}+u_1\right)\left(\frac{\mathrm{d}\ln\rho_1}{\mathrm{d}s}+\frac{2}{s}\right)=0 \tag{36}$$

where $\tau=(\mathrm{d}a/\mathrm{d}t)^{-1}$ is a constant. A final transformation is required to cast these two equations in nondimensional form:

$$x=\frac{s}{\tau\sqrt{RT}},\quad \xi=\frac{-u_1}{\sqrt{RT}},\quad \eta=4\pi G\rho_1\tau^2. \tag{37}$$

With these relations, (35) and (36) become

$$\frac{\mathrm{d}\xi}{\mathrm{d}x}=\frac{x-\xi}{x}\frac{\eta x(x-\xi)-2}{(x-\xi)^2-1}, \tag{38}$$

$$\frac{\mathrm{d}\ln\eta}{\mathrm{d}x}=\frac{x-\xi}{x}\frac{\eta x-2(x-\xi)}{(x-\xi)^2-1}. \tag{39}$$

Once complemented with the boundary conditions $\xi(x=0)=0$ and $\eta x=2$ at $x-\xi=1$, the above set gives the exact solution for the isothermal collapse evolution. Before discussing their numerical solution we note that in the limiting case $x\gg 1$ (when $x-\xi\to\infty$ and $\eta\to 0$), one finds that $\mathrm{d}\ln\xi/\mathrm{d}\ln x=0$ and $\mathrm{d}\ln\eta/\mathrm{d}\ln x=-2$. This implies that $u(r)$ becomes constant and the density profile approaches the asymptotic form $\rho\propto r^{-2}$. This feature is confirmed by the numerical solution of the equations shown in Fig. 3. The collapse proceeds essentially as a free-fall; however, pressure forces are not negligible and in fact shape the final form of the solutions. It is easy to show that the pressure gradient to gravitational force is about 60% in the region in which $x\ll 1$ and decreases at larger distances. It is also interesting to note that the solutions only depend on the assumed temperature and not on other properties of the collapsing cloud. Also clearly seen is the strong density enhancement (i.e. the core) in the central regions in which $\log x<0$.

2.2 Core Evolution

To investigate the subsequent evolutionary phases of the collapsing core it is necessary to add a substantial amount of physical processes as the energy equation, formation/destruction of molecules and their effect on the cooling function, radiative transfer of the cooling radiation. The first attempt to model such complications has been successfully carried out by [50] and [55]. From those studies, we can summarize the evolution of the central core as follows.

At the beginning, since a sufficient amount of H_2 has not formed yet, the temperature rises adiabatically by compression. At low densities H_2 is formed mainly via the (for details see Sect. 1) H^- channel. However, as the temperature and density boost the H_2 formation rate, a sufficient amount of H_2 to cool

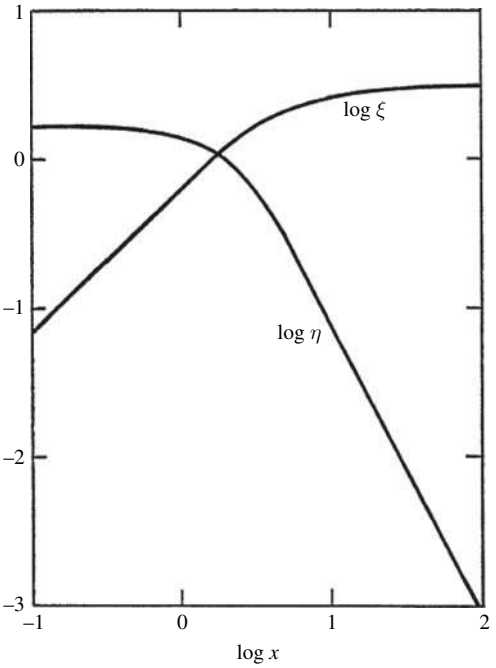

Fig. 3. Numerical solution for isothermal collapse described by (35) and (36)

within a free-fall time is formed, causing the temperature to drop to about 300 K. When the density reaches $\approx 10^8\,\mathrm{cm}^{-3}$, three-body reactions become efficient and hydrogen turns fully molecular at $n = 10^{10-11}\,\mathrm{cm}^{-3}$. At the same time, the cloud becomes optically thick to a few H_2 lines. However, the cooling remains efficient enough to induce dynamical collapse (i.e., the ratio of specific heat $\gamma \equiv \mathrm{d}\log p/\mathrm{d}\log\rho < 4/3$) due to the presence of a sufficient number of optically thin lines. When the central number density approaches $10^{14}\mathrm{cm}^{-3}$, H_2 collision-induced emission begins to dominate the cooling. Eventually, the cloud becomes optically thick to this continuum at $N \approx 10^{16}\,\mathrm{cm}^{-3}$, and the radiative cooling rate drops rapidly beyond that point. Simultaneously, H_2 dissociation begins, due to the high temperature; such dissociation prevents the temperature an additional temperature increase until $n > 10^{20}\,\mathrm{cm}^{-3}$. At that point the collapse becomes virtually adiabatic. After a minor further contraction, a small hydrostatic core of mass $\approx 10^{-3}\,\mathrm{M}_\odot$, or protostar, forms. Its density and temperature are $n \approx 10^{22}\,\mathrm{cm}^{-3}$, $T = 3 \times 10^4\,\mathrm{K}$.

2.3 Fragmentation

Although we have now understood how a protostellar core forms, the final mass of the star depends on the accretion history of the remaining gas onto the core itself. Because under primordial conditions standard ways to stop

such accretion seem to fail due to the lack of magnetic fields, dust and heavy element opacity, and the uncertain presence of an accretion disk-like structure, the final mass of the star will depend largely on the mass of the clump from which it is born. It is therefore crucial to understand the process of fragmentation of primordial clouds.

It is obvious that much hinges on the physics of cooling, primarily the number of channels available for the gas to cool and the efficiency of the process. As already discussed, cooling is important when $t_c \ll t_{ff}$. This condition implies that the energy deposited by gravitational contraction cannot balance the radiative losses; as a consequence, temperature decreases with increasing density. Under such circumstances, the cloud cools and then fragments. At any given time, fragments form on a scale that is small enough to ensure pressure equilibrium at the corresponding temperature, i.e., the Jeans length scale,

$$R_F \approx \lambda_J \propto c_s t_{ff} \propto n^{\gamma/2-1} \tag{40}$$

where the sound speed $c_s = \sqrt{RT/\mu}$, $T \propto n^{\gamma-1}$ and γ is the adiabatic index. Since c_s varies on the cooling timescale, the corresponding R_F becomes smaller as T decreases. Similarly, the corresponding fragment mass is the Jeans mass,

$$M_F \propto nR_F^{\eta} \propto n^{\eta\gamma/2+(1-\eta)} \tag{41}$$

with $\eta = 2$ for filaments, and $\eta = 3$ for spherical fragments. This hierarchical fragmentation process comes to an end when cooling becomes inefficient because (1) the critical density for LTE is reached or (2) the gas becomes optically thick to cooling radiation; in both cases, at that juncture $t_c \geq t_{ff}$. At this stage, the temperature cannot decrease any further, and it either remains constant (if energy deposition by gravitational contraction is exactly balanced by radiative losses) or increases. The necessary condition to stop fragmentation and start gravitational contraction within each fragment is that the Jeans mass does not decrease any further, thus favoring fragmentation into subclumps. From (41), this implies the condition

$$\gamma \geq 2\frac{\eta - 1}{\eta} \tag{42}$$

which translates into $\gamma \geq 4/3$ for a spherical fragment, and $\gamma \geq 1$ for a filament. Thus, a filament is marginally stable and contracts quasi-statically when $t_c \approx t_{ff}$ and the gas becomes isothermal. Finally, when $t_c \gg t_{ff}$ the fragments become optically thick to cooling radiation, and the temperature increases as the contraction proceeds adiabatically.

2.4 Fragmentation of Metal-Free Clouds

Reference [62] investigated the above fragmentation process numerically in great detail. The gas within a dark matter halo is given an initial temperature of 100 K, and the subsequent thermal and chemical evolution of

the gravitationally collapsing cloud is followed numerically until a central protostellar core forms. The gas within the dark matter halo gets shock-heated to the virial temperature $T_{\rm vir} \gg 100\,{\rm K}$. However, after a short transient phase, the evolutionary track in the (n, T) plane shown in Fig. 4 (top curve, top panel) provides a good description of the thermal evolution of the gas. The metal-free gas is able to cool down to temperatures of a few hundred kelvin regardless of the initial virial temperature. This is the minimum temperature at which molecular line cooling becomes effective. In any case, independent of the virial temperature, the thermal evolution of the gas rapidly converges to the (n, T) track corresponding to $Z = 0$ (the zero-metallicity track). The temperature of the gas decreases with increasing density, thus favoring fragmentation into subclumps.

As the number density increases, it reaches the critical value $n_{cr} = 10^3\,{\rm cm}^{-3}$; the corresponding Jeans mass is 10^4 $M_\odot$. The cooling time at this critical point becomes comparable to the free-fall time as a consequence of the H_2 levels being populated according to LTE. The temperature then starts to rise slowly. At this stage, the stability of the fragments toward further increase in the density needs to be investigated according to (41) above. The

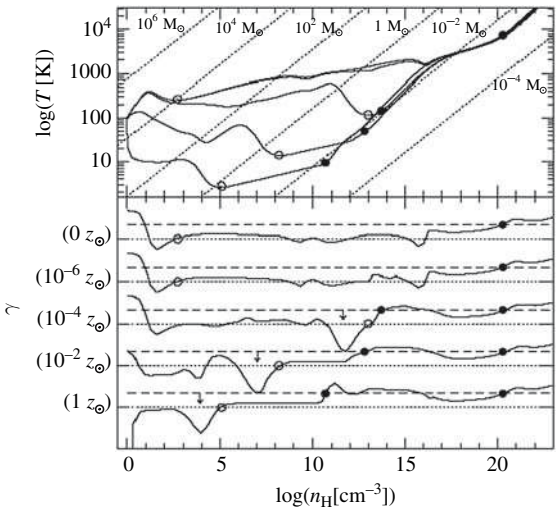

Fig. 4. Top: Evolution of the temperature as a function of the hydrogen number density of protostellar clouds with the same initial gas temperature but varying metallicities $Z = (0, 10^{-6}, 10^{-4}, 10^{-2}, 1)$ $Z_\odot$ (Z increasing from top to bottom curves). The *dashed lines* correspond to the constant Jeans mass for spherical clumps; *open circles* indicate the points where fragmentation stops; filled circles mark the formation of hydrostatic cores. **Bottom**: The adiabatic index γ as a function of the hydrogen number density for the curves shown in the top panel. Dotted (*dashed*) lines correspond to $\gamma = 1 (\gamma = 4/3)$; *open* and *filled circles* as above. From Schneider et al. (2002) [62]

bottom panel of Fig. 4 shows the density dependence of γ for each metallicity track. For a metal-free gas, γ lies in the range $0 < \gamma - 1 < 1/3$, implying that further fragmentation is unlikely to occur unless the fragments are spherical. Although the gravitational evolution will probably favor a tendency toward spherical symmetry, this is not likely to occur until the central density has reached high values as seen from simulations. It is important to stress that, even if the fragments are nearly spherical, fragmentation will be modest and is likely to result only in a low-multiplicity stellar system. As the density increases, quasi-static contraction takes place until the fragments become optically thick to H_2 lines, and adiabatically collapse to increasingly higher central densities and temperatures. At this stage, $\gamma > 4/3$ and a central hydrostatic core (filled circles in Fig. 4) is formed. Each fragment is characterized by a central core of $\approx 10^{-3}$ $M_\odot$ surrounded by a large envelope of gravitationally unstable gas. The core grows in mass because of gas accretion from the envelope. The mass of the formed stars depends on the accretion rate as well as on the fragment mass.

2.5 Fragmentation of Metal-enriched Clouds: Critical Metallicity

We now consider the effects of the presence of heavy elements on the fragmentation process. Figure 4 shows the effects of metal enrichment on the (n, T) tracks for the same initial conditions and different values of the mean metallicity. In general, clouds with lower metallicity tend to be warmer because of their lower radiative cooling ability. As long as the clouds are transparent, cooling and fragmentation occur. Clouds with a mean metallicity $Z < 10^{-6}$ $Z_\odot$ follow the same evolution as that of the gas with primordial composition. However, at $Z > 10^{-4}$ $Z_\odot$, H_2 formation on grain surfaces enhances cooling at low density. Dust grains are now known to condense out the ejecta of SNe (see Sect. 8). When the LTE-NLTE transition occurs for H_2, the cloud can still cool (although less efficiently) because of OI line cooling. At densities $> 10^6$ cm^{-3}, heating due to H_2 formation becomes larger than compressional work, and the temperature starts to increase until thermal emission from grains due to energy transfer between gas and dust dominates the cooling. This occurs at a density $n \approx 10^{10}$ cm^{-3}, where the temperature drops and a new fragmentation phase occurs. The minimum fragment mass is reached at the point indicated by the open circle, when the Jeans mass is 10^{-2} $M_\odot$. Finally, as the density increases, the gas becomes opaque to dust thermal emission, fragmentation stops, and compressional heating causes the fragments to contract adiabatically. Therefore, a *critical metallicity* of $Z_{\rm cr} \approx 10^{-4}$ $Z_\odot$ can be identified, which marks the transition point between metal-free and metal-rich gas evolution. For higher metallicities, cooling is driven by OI, CI, and CO line emission. When the NLTE-LTE transition for the level populations of CO occurs, fragmentation stops and the temperature increases because of H_2 formation. The larger concentration of dust grains (assumed here to be proportional to the mean metallicity) leads to a significant thermal emission that is responsible

for cooling the gas and starting a new phase of fragmentation. This stops when $T_{\rm grain} \approx T$, and thereafter the fragments contract quasi-statically until they become optically opaque to dust emission and adiabatic contraction occurs. Because of the enhanced ability to cool, fragmentation stops at lower temperatures and densities for higher metallicity clouds. However, the Jeans mass corresponding to the minimum fragmentation scale is always orders of magnitude smaller than for a cloud with no metals. In conclusion, the presence of metals not only enables fragmentation down to smaller mass scales 10^{-2} $\mathrm{M}_{\odot}$, but also breaks the one-to-one correspondence between the mass of the formed star and that of the parent fragment by halting the accretion through radiation force onto the dust.

2.6 Accretion Phases

As described in Sect. 2.4, the end product of the collapse is a small protostar surrounded by a large reservoir of gas. After its birth, the protostar grows by some orders in mass by accreting the envelope matter. The mass accretion rate on to the protostar is determined by the radial density distribution at the time of the protostar formation, which is related to the prestellar temperature or equivalently the sound speed $c_{\rm s}$ and roughly given by

$$\dot{M} = \frac{c_{\rm s}^3}{G}. \tag{43}$$

The mass of the forming stars is set when the protostellar accretion stops. Thus it is important to understand how accretion proceeds. At least three different phases can be identified [51], which are discussed below.

Adiabatic Accretion

Early in the evolution, the opacity in the protostar, dominated by free-free absorption is very high due to the low temperatures, resulting in very low interior luminosity. During this phase the timescale for the cooling of the protostellar interior (given by the Kelvin-Helmholtz time) is much longer than the evolutionary timescale of the protostar (i.e. the accretion time) due to the low luminosity,

$$t_{\rm KH} = L_{\star}^{-1} \frac{GM_{\star}^2}{R_{\star}} \gg \frac{M_{\star}}{\dot{M}_{\star}} = t_{\rm acc}. \tag{44}$$

After passing through the accretion shock and settling onto the protostar, the accreted material piles up without further cooling; the protostar swells gradually as the mass increases. The temperature increase accompanied by the growth in the protostellar mass makes the opacity fall rapidly. The lower opacity eventually allows the heat contained inside the protostar to propagate outward as a luminosity wave. Consequently, the interior luminosity increases

suddenly (for a "typical" accretion rate of 4×10^{-3} $M_\odot\,yr^{-1}$ this occurs at $M_\star \approx 8$ $M_\odot$. The arrival of the luminosity wave to the surface causes the sudden swelling of the protostar, marking the end of the adiabatic accretion phase.

Kelvin-Helmholtz Contraction

Once the opacity falls and $t_{KH} \approx t_{acc}$, the protostar cools and contracts. This phase is called the Kelvin-Helmholtz (KH) contraction phase. After the propagation of the luminosity wave, the major opacity source in the protostar is the electron scattering. If the radius increases, then $t_{KH} < t_{acc}$, and the core shrinks due to fast cooling; on the other hand, if the protostar shrinks, $t_{KH} > t_{acc}$, and the core swells adiabatically. Thus, the a self-regulated mechanism to keep $t_{KH} \approx t_{acc}$ exists. This leads to the relation $R_\star \propto \dot{M}_\star/M_\star^2$. During the KH contraction, $L_\star/M_\star$ increases rapidly and the luminosity-mass relation becomes approximately a unique relation independent of the mass accretion rate, as expected in the ideal case of constant opacity. Also, the photospheric luminosity, which is the sum of the interior $L_\star$ and accretion luminosity L_{acc} accretion luminosity, is independent of $\dot{M}$, since $L_{acc} \propto \dot{M}/R$ and $R \propto \dot{M}^{-1}$. Immediately after the onset of the KH contraction, the maximum temperature exceeds 10^6 K and the deuterium starts burning. For typical accretion rates though, deuterium burning contributes at most about 1/4 of the interior luminosity. Similarly, the heating due to the *p*-*p* chain is not sufficient to stop the contraction because the energy generation rate saturates at high temperatures. The rest of the luminosity comes from the gravitational contraction. Hence, despite nuclear heating, the protostar continues contraction.

ZAMS Settling

When the accretion rate is below the above typical value the protostar settles smoothly to a main-sequence star at the end of the KH contraction phase. In the course of the Kelvin-Helmholtz contraction, when the central temperature becomes high enough to synthesize carbon, which catalyze the CN cycle, the H burning becomes significant. At this moment, the gravitational contraction is halted and the interior luminosity to mass ratio stops increasing. Soon, the protostar relaxes to the ZAMS structure, which is characterized by a unique luminosity-mass-radius relation. With the H burning via the CN cycle working as a thermostat, the central temperature remains almost constant at about 10^8 K. After that, the interior luminosity is essentially supplied by the nuclear burning. Since a part of the nuclear generated energy is consumed to heat the stellar interior, the specific entropy in the convective core increases as the stellar mass grows. This results in the decrease of the central density with the temperature kept constant. The mass fraction of the convective core increases.

If instead the accretion rate exceeds the typical one, the protostar starts to violently swell when the luminosity becomes close to the Eddington limit,

without converging to the ordinary ZAMS structure. This expansion is caused by the strong radiation force exerted onto both the stellar surface and the radiative precursor. During the Kelvin-Helmholtz contraction, the luminosity-to-mass ratio increases until the central contraction is halted by the nuclear burning. As the accretion rate increases, the onset of the H ignition is delayed. For sufficiently high accretion rates, the total luminosity reaches the Eddington limit before the H ignition. The ram pressure onto the protostellar surface decreases suddenly, forcing the protostar to expand and reduce the total luminosity not to exceed the Eddington limit. With the expansion, the temperature drops, and the opacity and the radiation force increase at the surface. As a result, the expansion is accelerated. This violent expansion possibly results in stripping of the surface layer, as well as of the material in the accreting envelope. In conclusion, high accretion rates prevent the protostars to relax on the ZAMS structure.

2.7 Final Masses and Feedbacks

The effect of realistic time-dependent accretion rate, initially as high as $0.01\ \mathrm{M_\odot\,yr^{-1}}$ and rapidly decreasing once the stellar mass exceeds $\approx 90\ \mathrm{M_\odot}$ has been studied, guided by the outcome of numerical simulations. The basic result is that accretion could continue at a moderate rate for relatively long times unimpeded, lacking the main physical mechanisms to stop it (e.g. radiation pressure, magnetic fields driving bipolar outflows). Thus, stars with masses $\gg \mathrm{M_\odot}$ can in principle form, provided there is sufficient matter in the parent clumps.

As the lifetimes of primordial massive stars converge to about 2 Myr, the above accretion rates imply an upper limit to stellar masses of $\approx 2000\ \mathrm{M_\odot}$ However, other feedback processes are likely to intervene before this. Once the flux of ionizing photons from the protostar is greater than that of neutral H to its surface, an H II region forms. Accretion may be suppressed if the H II region expands to distances greater than R_{g}, where the escape speed equals the ionized gas sound speed, $\approx 10\ \mathrm{km\ s^{-1}}$. For the fiducial model of [66] this occurs at the poles when $M_\star = 90\ \mathrm{M_\odot}$, and at the equator when $M_\star = 140\ \mathrm{M_\odot}$. These conclusions are based on a simplified free-fall density distribution; in reality the ionizing radiation force decelerates and deflects the flow.

Another ingredient which has so far received little attention but which could be quite important to determine the final mass of the first stars, is rotation of the infalling envelope. For collapse with angular momentum, most streamlines do not come too close to the star, so the H II region quenching due to enhanced densities is greatly weakened. In fact, deflection is more important in reducing the concentration of inflowing gas near the star, so that the H II region becomes larger. Finally, another feedback effect is radiation pressure from Lyα photons created in the H II region and FUV photons emitted by the star. These photons are trapped by the Lyman series damping

wings of the neutral gas infalling towards the HII region. The energy density builds up until the escape rate, set by diffusion in frequency as well as in space, equals the input rate. The resulting pressure acts against the infall ram pressure. Radiation pressure becomes greater than twice the ram pressure at $M_\star \approx 20\ \mathrm{M}_\odot$, depending on the rotation speed. The enhancement of radiation pressure above the optically thin limit is by a factor 1000. Infall is first reversed at the poles, which would allow photons to leak out and reduce the pressure acting in other directions.

In summary, although the typical mass of the first stars is suggested by the above arguments and studies to be much larger than the present one, thus maybe leading to a top-heavy primordial IMF, studies in this area are still in their infancy and definite conclusions have to await for more detailed calculations.

3 First Stars

The primordial star formation process and its final products are presently quite unknown. This largely depends on our persisting ignorance of the fragmentation process and on its relationship with the thermodynamical conditions of the gas. Despite these uncertainties, it is presently accepted that the first stars, being formed out of a gas of primordial composition, are metal-free (Pop III stars). In addition to this theoretical justification, the existence of Pop III stars have been historically invoked for many different reasons (for a review, see [15]). The very first generation of stars must have formed out of probably unmagnetized, pure H/He gas, since heavy elements can only be produced in the interior of stars. These characteristics render the primordial star formation problem very different from the present-day case, and lead to a significant simplification of the relevant physics. Nevertheless, the complexity and interactions of the hydrodynamical, chemical and radiative processes, have forced many early studies to use the steady state shock assumption, a spherical or highly idealized collapse model. Nevertheless, the arguments presented in the previous sections, suggest that Pop III stars could have been much more massive than stars formed today, with a tentative mass range $100 < M/\mathrm{M}_\odot < 1000$. For such a massive star with the concomitant high interior temperatures, the only effective source of opacity is electron scattering. In the absence of metals and, in particular, of the catalysts necessary for the operation of the CNO cycle, nuclear burning proceeds in a nonstandard way. At first, hydrogen burning can only occur via the inefficient p-p chain. To provide the necessary luminosity, the star has to reach very high central temperatures ($T > 10^8\,\mathrm{K}$). These temperatures are high enough for the simultaneous occurrence of helium burning via the triple-α process. After a brief initial period of triple-α burning, a trace amount of heavy elements has been formed. Subsequently, the star follows the CNO cycle. The resulting structure consist of a convective core, containing about 90% of the

mass, and a thin radiative envelope. As a result of the high mass and temperature, the stars are dominated by radiation pressure and have luminosities close to the Eddington limit $L_{\rm edd} = 10^{38}(M/{\rm M}_\odot)\,{\rm erg\,s^{-1}}$. A major uncertainty in the evolution of these stars is mass loss. Although radiative mass loss is probably negligible for stars of such low metallicity, they still might lose an appreciable fraction of their mass because of nuclear-driven pulsations.

The mechanical structure of a massive, radiation pressure-dominated star is approximately given by a polytrope of index $n = 3$ [10]. One can then use the solution of the Lane-Emden equation to express the density and pressure, and, with the polytropic value for the central pressure, one can estimate the temperature at the center of the star as

$$T_{\rm c} = 8.4 \times 10^7\,{\rm K}(M_\star/100\ {\rm M}_\odot)^{1/2}(R/10\ {\rm R}_\odot)^{-1} \tag{45}$$

To constrain $T_{\rm c}$ further, we consider the global energy balance of the star. By balancing the average nuclear energy generation rate due to CNO with the Eddington luminosity, which indicates the energy lost by radiation, one finds:

$$T_{\rm c} = 1.9 \times 10^8\,{\rm K}(M_\star/100\ {\rm M}_\odot)^{-1/8}(R/10\ {\rm R}_\odot)^{3/8} \tag{46}$$

Combining the two expressions above for $T_{\rm c}$ we derive the required mass-radius relation for very massive stars:

$$M = 370\ {\rm M}_\odot(R/10\ {\rm R}_\odot)^{2.2}. \tag{47}$$

Finally, the effective temperature can be estimated from $L \approx L_{\rm edd} = 4\pi R^2\sigma T_{\rm eff}^4$; Using the previous equation to eliminate the radius we get:

$$T_{\rm eff} = 1.1 \times 10^5\,{\rm K}(M_\star/100\ {\rm M}_\odot)^{0.0025}. \tag{48}$$

This analytical estimate for $T_{\rm eff}$ and the very weak dependence on mass are in good agreement with more sophisticated numerical analyses. The relevant point here is that massive stars tend to be very hot. We will see in the following a number of important implications of this fact.

3.1 Emission Spectrum

The high effective temperature of massive stars produces interesting emission features. Before discussing those, however, it is necessary to understand how they evolve in the $L - T_{\rm eff}$ plane, i.e. to follow their stellar tracks. These depend strongly on the assumptions made concerning the mass loss. We will compare and discuss here, following [59], the two extreme cases of (i) no mass loss, and (ii) strong mass loss. The main difference between the "strong" and "no mass loss" sets is the rapid blueward evolution of the stars in the former

case, due to strong increase of the He abundance on the surface of these stars leading to a hot Wolf-Rayet-like phase. While the use of different input physics (e.g. nuclear reaction rates) could lead to somewhat different results if recomputed with more modern codes, the predicted tracks depend essentially only on the adopted mass loss and remain thus completely valid. Due to the lack of CNO elements the ZAMS of massive Pop III stars is much hotter that their solar or low metallicity counterparts. In particular this implies that at $Z = 0$ stars with $M > 5$ $M_\odot$ have unusually high temperatures and in turn non-negligible ionizing fluxes corresponding to normal O-type stars ($T_{\rm eff} > 30,000$ K).

These unique physical characteristics enhance the ionizing photon production of Pop III stars, particularly in the He II continuum, in which they produce up to 10^5 times more photons than Pop II. If, in addition, the first stars were also massive, they would be even hotter and have harder spectra. Metal-free stars with mass above 300 $M_\odot$ resemble a blackbody with an effective temperature of $\approx 10^5$ K, with a production rate of ionizing radiation per stellar mass larger by 1 order of magnitude for H and He I and by 2 orders of magnitude for He II than the emission from Pop II stars. In the less extreme case of metal-free stars with masses < 100 $M_\odot$, the H-ionizing photon production takes twice as long as that of Pop II to decline to 1/10 of its peak value. Nevertheless, due to the red-ward stellar evolution and short lifetimes of the most massive stars, the hardness of the ionizing spectrum decreases rapidly, leading to the disappearance of the characteristic He II recombination lines after about 3 Myr in instantaneous burst models. Emission spectra from primordial stars have been calculated by different authors for stars with masses up to ≈ 1000 $M_\odot$. Some calculations have carried out evolutionary calculations over this entire range of stellar masses, covering the H and He-burning phases and allowing for a moderate overshooting from convective cores, but neglecting rotation. Additionally, for very massive stars ($M > 120$ $M_\odot$), recent mass-loss rate prescriptions for the radiation-driven winds at very low metallicities have been applied and the amplification of the loss rate caused by stellar rotation calculated. Nevertheless, the above studies are based on some strong simplifying assumptions; e.g. all the stars are assumed to be on the Zero Age Main Sequence (ZAMS) (i.e. stellar evolution is neglected) and nebular continuum emission is not included. In particular, this last process cannot be neglected for metal-poor stars with strong ionizing fluxes, as it increases significantly the total continuum flux at wavelengths red-ward of Lyα and leads in turn to reduced emission line equivalent widths. Nebular emission has been included in a more complete and extended studies by [59, 60], who presents realistic models for massive Pop III stars and stellar populations based on non-LTE model atmospheres, recent stellar evolution tracks and up-to-date evolutionary synthesis models, including also different IMFs [15]. Table 1 gives a summary of the emission properties of Pop III stars. The numbers have been derived by integrating the ionizing photon rate in the absence of stellar winds over three different IMFs, i.e. Salpeter, Larson and Gaussian.

Table 1. The different progenitor models for SNII. The masses are in $M_\odot$ and $\mathcal{N}^{II}$ is in units of $M_\odot^{-1}$

SN Model	M_{SNII}	M_{BH}	$Z/Z_\odot$	$\mathcal{N}^{II}$
SNII-A	12	40	10^{-2}	0.00343
SNII-B	12	40	1	0.00343
SNII-C	10	50	10^{-2}	0.00484
SNII-D	10	50	1	0.00484

3.2 Final Fate

As the *p*-*p* chain is never sufficiently efficient to power massive stars, stars of initial zero metallicity contract until central temperatures $> 10^8$ K are reached and CNO seed isotopes are produced by the triple-α process. As a consequence of their peculiar behavior in hydrogen burning, the post-main sequence entropy structure of stars of zero initial metallicity is different from that of stars having finite metallicity. As already stated, although radiative mass loss is probably negligible for stars of such low metallicity, they still might lose an appreciable fraction of their mass because of nuclear-driven pulsations. Recent theoretical analysis on the evolution of metal-free stars predicts that their fate can be classified as follows:

- Stars with masses $10 < M/M_\odot < 40$ proceed through the entire series of nuclear burnings accompanied by strong neutrino cooling: hydrogen to helium, helium to carbon and oxygen, then carbon, neon, oxygen and silicon burning, until finally iron is produced. When the star has built up a large enough iron core, exceeding its Chandrasekhar mass, it collapses, followed by a supernova explosion. In particular, stars with $M > 30$ $M_\odot$ would eventually collapse into a Black Hole (BH)
- For stars of mass $40 < M/M_\odot < 100$ the neutrino-driven explosion is probably too weak to form an outgoing shock. A BH forms and either swallows the whole star or, if there is adequate angular momentum, produces a jet which could result in a GRB.
- Stars with $M > 100$ $M_\odot$ form large He cores that reach carbon ignition with masses in excess of about 45 $M_\odot$. It is known that after helium burning, cores of this mass will encounter the electron-positron pair instability, collapse and ignite oxygen and silicon burning explosively. If explosive oxygen burning provides enough energy, it can reverse the collapse in a giant nuclear-powered explosion (the so-called Pair Instability Supernova, PISN) by which the star would be partly or completely (if $M > 140$ $M_\odot$) disrupted. For even more massive stars ($M > 260$ $M_\odot$) a new phenomenon occurs as a sufficiently large fraction of the center of the star becomes so hot that the photodisintegration instability is encountered before explosive burning reverses the

implosion. This uses up all the energy released by previous burning stages and, instead of producing an explosion, accelerates the collapse leading to the prompt formation of a massive BH, and, again, either a complete collapse or a jet-powered explosion (see later on for more details).

- At even higher masses ($M > 10^5$ M$_\odot$) the evolution depends on the metallicity: if $Z < 0.005$ the star collapses to a BH as a result of post-Newtonian instabilities without ignition of the hydrogen burning; for higher metallicities it explodes, as it could generate nuclear energy more rapidly from β-limited cycle.

3.3 Metal Yields and Explosion Energies

We have seen that very likely the massive star formation mode continued until the cosmic gas reached the critical metallicity value, $Z_{\rm cr} \approx 10^{-5\pm1}$ Z$_\odot$, which enabled the formation of lower mass clumps and allowed some of the above mechanisms to stop accretion. This situation leads to the so-called *star formation conundrum*. If Pop III stars were indeed very massive, stellar evolution models predict that most of these objects would have evolved into black holes, which lock in their nucleosynthetic products. This, in turn, would prevent the metal enrichment of the surrounding gas and force the top-biased star formation mode to continue indefinitely. It is only in the relatively narrow mass window characterizing the PISN in fact, that ejection of heavy elements can occur. Thus, PISN may be essential to increase the intergalactic medium metallicity above $Z_{\rm cr}$ and initiate the transition from top-biased star formation to the formation of "normal" (Population II/Population I) stars with a more typical initial mass function.

In order to quantify the properties of Pop III and Pop II/I objects, we consider the metal yields and explosion energies of pair-creation (SN$_{\gamma\gamma}$) and Type II SNe (SNII), respectively. In the Pop II/I case we adopt the results of Woosley & Weaver (1995) who compute these quantities for SNII as a function of progenitor mass in the range 12 M$_\odot$ $< M_\star <$ 40 M$_\odot$ and initial metallicities $Z = (0, 10^{-4}, 10^{-2}, 10^{-1}, 1)$ Z$_\odot$. The corresponding quantities for PISN are taken from [31] assuming progenitors in the mass range $\Delta_{\gamma\gamma} \equiv 140$ M$_\odot$ $\leq M_\star \leq 260$ M$_\odot$.

Pop II/I Objects

The total heavy-element mass and energy produced by a given object depend on the assumed IMF. We define Pop II/I objects as those that host predominantly Pop II/Pop I stars, i.e. objects with initial metallicity $Z \geq Z_{\rm cr} = 10^{-4}$ Z$_\odot$. For these, we assume a canonical Salpeter IMF, $\Phi(M) \propto M^{-(1+x)}$, with $x = 1.35$, and lower (upper) mass limit $M_{\rm l} = 0.1$ M$_\odot$ ($M_{\rm u} = 100$ M$_\odot$), normalized so that,

$$\int_{M_{\rm l}}^{M_{\rm u}} {\rm d}MM\Phi(M) = 1. \tag{49}$$

The IMF-averaged SNII metal yield of a given heavy element i (in solar masses) is then

$$Y_{\rm i}^{II} \equiv \frac{\int_{M_{\rm SNII}}^{M_{\rm BH}} {\rm d}M\Phi(M)M_{\rm i}}{\int_{M_{\rm SNII}}^{M_{\rm BH}} {\rm d}M\Phi(M)} \tag{50}$$

where $M_{\rm i}$ is the total mass of element i ejected by a progenitor with mass M. The mass range of SNII progenitors is usually assumed to be (8–100) $\rm M_\odot$. However, above $M_{\rm BH} = 50 \pm 10$ $\rm M_\odot$ stars form black holes without ejecting heavy elements into the surrounding medium (Tsujimoto et al. 1995), and from (8–11) $\rm M_\odot$ the pre-supernova evolution of stars is uncertain, resulting in tabulated yields only between 12 $\rm M_\odot$ and 40 $\rm M_\odot$ (Woosley & Weaver 1995). To be consistent with previous investigations (Gibson, Loewenstein, & Mushotzky 1997; Tsujimoto et al. 1995), in our reference model (SNII-C in Table 1) we consider SNII progenitor masses $M_{\rm SNII} = 10$ $\rm M_\odot \leqslant M_\star \leqslant M_{BH} = 50$ $\rm M_\odot$ range, linearly extrapolating from the lowest and highest mass grid values. Finally, SNII yields depend on the initial metallicity of the progenitor star. Here we assume for simplicity that SNII progenitors form out of a gas with $Z \geq 10^{-2}$ $\rm Z_\odot$. Moreover, the predicted SNII yields with initial metallicity in the range 10^{-4} $\rm Z_\odot \leq Z \leq 10^{-2}$ $\rm Z_\odot$ are largely independent of the initial metallicity of the star (Woosley & Weaver 1995). Different combinations of progenitor models relevant for the present analysis are illustrated in Table 1. The last entry in each row is the number of SNII progenitors per unit stellar mass formed,

$$\mathcal{N}^{II} \equiv \frac{\int_{M_{\rm SNII}}^{M_{\rm BH}} {\rm d}M\Phi(M)}{\int_{M_{\rm l}}^{M_{\rm u}} {\rm d}MM\Phi(M)}. \tag{51}$$

Finally, we assume that each SNII releases $\mathcal{E}_{\rm kin} = 1.2 \times 10^{51}$ erg, independent of the progenitor mass (Woosley & Weaver 1995).

Table 3 shows the IMF-averaged yields for some relevant elements, as well as the total mass of metals released in different SNII models. In the same Table, we also show the corresponding yields for Type Ia SNe (SNIa), whose values are mass-independent.

Pop III Objects

The detailed shape of the Pop III IMF is highly uncertain, as these stars have not been directly observed. In this investigation we adopt a model that is based on numerical and semi-analytical studies of primordial gas cooling and

fragmentation ([1]), which show that these processes mainly depend on molecular hydrogen physics. In particular, the end products of fragmentation are determined by the temperature and density at which molecular hydrogen levels start to be populated according to LTE, significantly decreasing the cooling rate. The minimum fragment mass is then comparable to the Jeans mass at these conditions, which is $\sim 10^3$ M$_\odot$, although minor fragmentation may also occur during gravitational contraction, due to the occasional enhancement of the cooling rate by molecular hydrogen three-body formation.

Ultimately, in these conditions the stellar mass is determined by the efficiency of clump mass accretion onto the central protostellar core. For a gas of primordial composition, accretion is expected to be very efficient due to the low opacity and high temperature of the gas. Given the uncertainties is these models, it is interesting to explore different values for the characteristic stellar mass $M_{\rm C}$ and for $\sigma_{\rm C}$, the dispersion around this characteristic mass. In the following we assume that Pop III stars are formed according to a mass distribution given by,

$$\Phi(M)M{\rm d}M = \frac{1}{\sqrt{2\pi}\sigma_C}{\rm e}^{-(M-M_{\rm C})^2/2\sigma_{\rm C}^2}{\rm d}M, \tag{52}$$

where 100 M$_\odot \leqslant M_{\rm C} \leqslant$ 1000 M$_\odot$ and $\sigma_{\rm C}^{\rm min} \leqslant \sigma_{\rm C} \leqslant \sigma_{\rm C}^{\rm max}$. A similar shape for the initial Pop III IMF has been also suggested by Nakamura & Umemura (2001), with a second Gaussian peak centered around a much smaller characteristic mass of 1 M$_\odot$, leading to a bimodal IMF.

The minimum dispersion is taken to be $\sigma_{\rm C}^{\rm min} = 0.1\,M_{\rm C}$ and the maximum value is set so that less than 1% of stars form below a threshold mass of $M_{\rm BH} \sim 50$ M$_\odot$, i.e. $\sigma_{\rm C}^{\rm max} = (M_{\rm C} - 50$ M$_\odot)/3$, such that they will not leave behind small stars or metal-free compact objects observable today. With this choice, only progenitors with mass in the range $\Delta_{\gamma\gamma}$ are effective in metal enrichment. The IMF-averaged yields, $Y_{\rm i}^{\gamma\gamma}$, and the number of progenitors per unit stellar mass formed, $\mathcal{N}^{\gamma\gamma}$ can be obtained from expressions equivalent to (50) and (51) with the assumed IMF and metal yields from [31]. For each pair of $M_{\rm C}$ and $\sigma_{\rm C}$ values, we also compute the parameter

$$f_\gamma\gamma \equiv \frac{\int_{\Delta_{\gamma\gamma}} {\rm d}MM\Phi(M)}{\int_0^\infty {\rm d}MM\Phi(M)}, \tag{53}$$

i.e. the fraction of Pop III stars ending in PISN. Similarly, the IMF-averaged specific kinetic energy is

$$\mathcal{E}_{\rm kin}^{III} \equiv \frac{\int_{\Delta_{\gamma\gamma}} {\rm d}M\Phi(M)E_{\rm kin}}{\int_{\Delta_{\gamma\gamma}} {\rm d}M\Phi(M)}. \tag{54}$$

The various models are described in Table 2 and the corresponding yields are given in Table 3.

Table 2. PISN progenitor models. The first three entries are in $M_\odot$, $\mathcal{N}^{\gamma\gamma}$ is in units $M_\odot^{-1}$ and $\mathcal{E}_{\rm kin}$ is in units of 10^{51} erg. The last entry represents the kinetic energy per unit gas mass, $\mathcal{E}_{\rm g}^{\rm III} = f_*^{\rm III} f_{\rm w} \mathcal{N}^{\gamma\gamma} \mathcal{E}_{\rm kin}^{\rm III}$ in units of 10^{51} erg $M_\odot^{-1}$ computed assuming a Pop III star formation efficiency $f_*^{\rm III} = 0.1$, and wind efficiency $f_{\rm w} = 0.1$ (see text)

SN Model	$M_{\rm C}$	$\sigma_{\rm C}^{\rm min}$	$\sigma_{\rm C}^{\rm max}$	$\mathcal{N}^{\gamma\gamma}$	$f_{\gamma\gamma}$	$\mathcal{E}_{\rm kin}^{\rm III}$	$\mathcal{E}_{\rm g}^{\rm III}$
PISN-A	100	10	17	$< 6 \times 10^{-5}$	$[0.03 - 8]10^{-3}$	$6 - 7.5$	$< 4 \times 10^{-6}$
PISN-B	200	20	50	$[5 - 4]10^{-3}$	$0.99 - 0.8$	$38.4 - 37.5$	$[2 - 1.5]10^{-3}$
PISN-C	260	26	70	$[2.1 - 2.2]10^{-3}$	$0.5 - 0.4$	$67.1 - 46.4$	$[1.4 - 1]10^{-3}$
PISN-D	300	30	83	$[0.4 - 1.4]10^{-3}$	$0.09 - 0.3$	$70.5 - 48.2$	$[2.6 - 6.6]10^{-4}$
PISN-E	500	50	149	$< 2.2 \times 10^{-4}$	< 0.05	$70.6 - 47.8$	$< 10^{-4}$
PISN-F	1000	100	314	$< 3 \times 10^{-5}$	$< 6.3 \times 10^{-3}$	$69.9 - 43.6$	1.3×10^{-5}

Table 3. The IMF-averaged metal yields (in $M_\odot$) for SNII, SNIa, and PISN

SN Model	Y_O	Y_{Si}	Y_S	Y_{Fe}	Y_{met}
SNII-A	1.43	0.133	0.064	0.136	2.03
SNII-B	1.56	0.165	0.078	0.115	2.23
SNII-C	1.11	0.1	0.047	0.126	1.6
SNII-D	1.19	0.124	0.061	0.11	1.76
SNIa	0.148	0.158	0.086	0.744	1.23
PISN-A	48.4–47.5	2.07–4.37	0.559–13.2	$[0.29–13.3]10^{-2}$	59.3–61.3
PISN-B	44.2–43.5	20.9–18.9	8.77–7.89	5.81–7.41	89–86.8
PISN-C	37.4–41.8	24.3–20.8	11.2–9.01	22.1–11.6	104–92.4
PISN-D	35–41.4	23.6–21.1	11.0–9.19	24.9–12.5	104–93.4
PISN-E	32.8–41.3	22.5–20.8	10.5–9.07	25.7–12.4	100–92.9
PISN-F	34.3–42.1	23.1–19.8	10.8–8.51	24.8–10.6	102–90.2

3.4 Effects of Rotation: Hypernovae

One of the most interesting recent developments in the study of supernovae from very massive stars is the discovery of some very energetic explosions, whose kinetic energy (KE) exceeds 10^{52} erg, about 10 times the KE of normal core-collapse SNe ($E_{51} = E/10^{51}$ erg). The most luminous and powerful of these objects ($E_{51} = 30$), the Type Ic supernova (SN Ic) 1998bw, was probably linked to the gamma-ray burst GRB 980425 thus establishing for the first time a connection between gamma-ray bursts (GRBs) and the well-studied phenomenon of core-collapse SNe. In addition, SN 1998bw was exceptional for a SN Ic: it was as luminous at peak as a SN Ia. Because of its large KE, SN 1998bw was called a "Hypernova (HN)". These objects span a wide range of properties, although they all appear to be highly energetic compared to normal core-collapse SNe. What is the reason for such large and anomalous explosion energies? The most widely accepted explanation has to do with the explosion of a *rotating* massive star ([47]).

This has become evident as researchers have started to plot the explosion energy E as a function of the main-sequence mass $M_\star$ of the progenitor star as derived from fitting the optical light curves and spectra of various hypernovae and normal supernovae (Fig. 5) it appears that E increases with $M_\star$, forming a "Hypernova Branch", reaching values much larger than the canonical $E_{51} = 1$. However other SNe (1997D, 1999br among the others) are located well below that branch, forming a "Faint SN Branch". This trend might be interpreted as follows. Stars with $M_\star < 20$–25 $M_\odot$ form a neutron star (SN 1987A may be a borderline case between the neutron star and black hole formation). Stars with $M_\star > 20$–25 $M_\odot$ form a black hole; whether they become hypernovae or faint SNe may depend on the angular momentum in the collapsing core, which

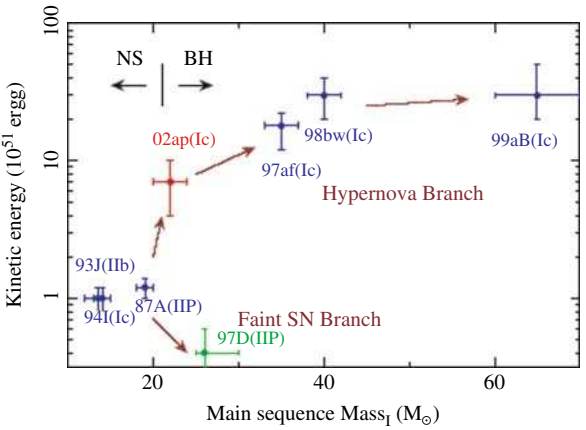

Fig. 5. Explosion energies vs. the main sequence mass of the progenitors for several bright supernovae/hypernovae

in turn depends on the stellar winds, metallicity, magnetic fields, and binarity. Hypernovae might have rapidly rotating cores owing possibly to the spiraling-in of a companion star in a binary system. The core of faint SNII might not have a large angular momentum, because the progenitor had a massive H-rich envelope so that the angular momentum of the core might have been transported to the envelope possibly via a magnetic torques. Between these two branches, there may be a variety of SNe.

The nucleosynthetic products of hypernovae have very clear signatures, which markedly differ from normal core/collapse Sne. In core-collapse supernovae/hypernovae, stellar material undergoes shock heating and subsequent explosive nucleosynthesis. Iron-peak elements are produced in two distinct regions, which are characterized by the peak temperature, $T_{\rm p}$, of the shocked material. For $T_{\rm p} > 5 \times 10^9$ K, material undergoes complete Si burning whose products include Co, Zn, V, and some Cr after radioactive decays. For $4 \times 10^9 < T_{\rm p} < 5 \times 10^9$ K, incomplete Si burning takes place and its after decay products include Cr and Mn. Three main differences between a HN and a normal SN of the same mass can be individuated: (1) Both complete and incomplete Si-burning regions in HN shift outward in mass compared with normal supernovae, so that the mass ratio between the complete and incomplete Si-burning regions becomes larger. As a result, higher energy explosions tend to produce larger [(Zn, Co, V)/Fe] and smaller [(Mn, Cr)/Fe]. (2) In the complete Si-burning region of hypernovae, elements produced by α-rich freeze-out are enhanced. Hence, elements synthesized through capturing of α-particles, such as ^{44}Ti, ^{48}Cr, and ^{64}Ge (decaying into ^{44}Ca, ^{48}Ti, and ^{64}Zn, respectively) are more abundant. (3) Oxygen burning takes place in more extended regions for the larger KE. Then more O, C, Al are burned to produce a

larger amount of burning products such as Si, S, and Ar. Therefore, hypernova nucleosynthesis is characterized by large abundance ratios of [Si,S/O].

3.5 First Stars and Gamma-Ray Bursts

As discussed in Sect. 3.4, metal-free (and also moderately metal-poor) stars more massive than approximately 260 $M_\odot$ collapse completely to BHs. Similar arguments apply to stars in a lower mass window (30 $M_\odot$–140 $M_\odot$), which are also expected to end their evolution as BHs. If this is the case, a numerous population of Intermediate Mass Black Holes (IMBHs) – with masses $M_\bullet \approx 10^2 - 10^3$ $M_\odot$, between those of stellar and SuperMassive Black Holes (SMBHs) – may be the end-product of those episodes of early star formation. As these pre-galactic BHs become incorporated through a series of mergers into larger and larger halos, they sink to the center because of dynamical friction, accrete a fraction of the gas in the merger remnant to become supermassive, form a binary system, and eventually coalesce. If only one IMBH with $M_\bullet > 150$ $M_\odot$ formed in each of the minihalos collapsing at $z \approx 20$ from 3-σ fluctuations, then the mass density of Pop III IMBHs would be comparable to that of the supermassive variety observed in the nuclei of galaxies. Studies of this scenario have included a number of physical ingredients such as gas accretion, hardening of the IMBH binary and triple interactions. These results show that the hierarchical growth can reproduce the observed luminosity function of optically selected quasars in the redshift range $1 < z < 5$. In addition, a prediction of the model is that a population of "wandering" black holes in galactic halos and in the IGM should exist, contributing around 10% of the present day total black hole mass density, 4×10^5 $M_\odot$ Mpc^{-3}. IMBHs that have not yet ended up in SMBHs, can be either (i) en route toward thereby accounting for the X-ray-bright off-center sources detected locally by ROSAT, or (ii) constituting the dark matter candidates composing the entire baryonic halos of galaxies.

Once a proto-BH has formed into the stellar core, accretion continues through a disk. It is widely accepted, though not confirmed, that magnetic fields drive an energetic jet which produces a burst of TeV neutrinos by photon-meson interaction, and eventually breaks out of the stellar envelope appearing as a Gamma-Ray Burst (GRB). Based on recent numerical simulations and neutrino emission models, the expected neutrino diffuse flux from these Pop III GRBs could be within the capabilities of present and planned detectors as AMANDA and IceCube High-energy neutrinos from Pop III GRBs could dominate the overall flux in two energy bands, 10^4–10^5 GeV and 10^5–10^6 GeV, of neutrino telescopes. The enhanced sensitivities of forthcoming detectors in the high-energy band (AMANDA-II), will provide a fundamental insight into the characteristic explosion energies of Pop III GRBs, and will constitute a unique probe of the IMF of the first stars. Based on such results, Pop III GRBs could be associated with a new class of events detected by BEPPO-SAX,

the Fast X-ray Transients (FXTs), bright X-ray sources with peak energies in the 2–10 keV band and durations between 10 and 200 s.

4 Observational Signatures of First Stars

Despite the efforts made in the past decades, the search for zero-metallicity stars has been proven unfruitful as no Pop III star has been detected yet. The first very low metallicity star, with a logarithmic abundance relative to solar of $[\mathrm{Fe/H}] = -4.5$, was identified more than 15 years ago. Since then, more stars with $-4.5 < [\mathrm{Fe/H}] < -2$ have been detected. Only recently new records have been established: [13] have measured an abundance of $[\mathrm{Fe/H}] = -5.3$ in HE0107-5240, a star with a mass of $\approx 0.8\,\mathrm{M}_\odot$; [25] found $[\mathrm{Fe/H}] = -5.4$ for HE1327-2326. Does the existence of this stars suggest that Pop III also contained low-mass and long-lived objects? Is there a strong lower bound to the metallicity of the halo stars implying that the first stars were too massive to have survived until today? What is the age-metallicity relation of metal-poor stars? How did metal enrichment proceeded within galaxies and in the intergalactic medium? What are the best strategies to search for a truly primordial stellar population? This set of related questions represent the core of the material presented in this Lecture. Before addressing them in detail, though, it is useful to review the background observations and introduce some definitions and data.

4.1 Stellar Relics

The distribution of stellar metallicities in the halo of the Galaxy has been intensively studied since the early 1980. The main difficulty of this experiment is that metal-poor (MP) stars are extremely rare in the solar neighborhood. As a direct consequence, knowledge of the form of the Metallicity Distribution Function (MDF), in particular the shape of its low-metallicity tail, has been limited by small-number statistics. The most widespread technique applied to identify a "fair" sample of halo stars exploits their motions (for a review see [5]). Spectroscopic follow-up of stars selected from proper-motion surveys has allowed the determination of at least a reasonably accurate picture of the global shape of the halo MDF. Early studies were able to demonstrate that the MDF of halo stars peaks at a metallicity $[\mathrm{Fe/H}] = -1.6$, including tails extending up to the solar metallicity on the high side, and to metallicities at least down to $[\mathrm{Fe/H}] = -3.0$, or slightly below, on the low side. Both suggested MDFs are consistent with one another, and with the predictions of the so-called "Simple Model". The actual shape of the MDF at the lowest metallicities, and its precise cutoff, is limited by the small numbers of stars in these samples. Even taken together, these surveys only include some 250 very metal-poor stars with $[\mathrm{Fe/H}] < -2.0$, and but a handful with $[\mathrm{Fe/H}] < -3.0$. Recent kinematically unbiased surveys for metal-poor stars have revealed the

presence of at least two hyper metal-poor (HMP) stars, with [Fe/H] < -5.0 ([13, 25]). The ongoing effort to search for MP stars has a long history which coincides essentially with the surveys that have been carried on during the last decades. In brief, the techniques used and the results obtained can be summarized as follows (for a detailed description see [5]).

*The HK Survey**

The HK survey was initiated over 25 years ago. During the course of the survey, a total of some 300 objective-prism plates covering 2800 deg^2 in the northern hemisphere and 4100 deg^2 in the southern hemisphere were obtained. The selection of candidate metal-poor stars was accomplished based on visual inspection. Medium-resolution spectroscopic and broadband photometric follow-up of candidate MP from the HK survey has been underway for two decades. The survey has allowed to obtain about 14,500 spectra of candidates of which 75% are unique.

*The Hamburg/ESO Survey**

The HES objective-prism survey provides the opportunity to greatly increase the number of very metal-poor stars identified by the HK survey. It reaches about two magnitudes deeper than the HK survey (B = 17.5 vs. B = 15.5) and also covers regions of the southern sky not sampled by the HK survey; a total of about 8225 deg^2 of the sky above $|b| = 30^{\circ}$ is presently available from the HES. The selection of metal-poor candidates from the HES database of digital objective-prism spectra is performed using quantitative criteria. Medium-resolution spectroscopic follow-up of candidate metal-poor stars from the HES has been underway for the past 5 years, primarily with 2.5–4 m class telescopes. The number of HES targets (including stars other than metal-poor candidates) with available medium-resolution spectroscopy obtained to date is about 7500.

*The Sloan Digital Sky Survey**

The SDSS includes a large number of Milky Way stars (about 70,000) at medium-resolution (2.5 Å) spectroscopy, with *ugriz* photometry. Most stars contained in the SDSS database were not targeted specifically to be metal-poor halo objects, and indeed they represent a complex assembly of objects selected for calibration and reddening determinations, directed studies of various classes of stars (e.g. horizontal-branch stars, carbon stars, white dwarfs, late-type K and M dwarfs, etc), as well as objects originally targeted as quasars that turned out to be stars. In spite of its rather inhomogeneous assembly, the SDSS stellar database does provide a useful means for sampling the tail of the halo MDF.

Figure 6 shows the distribution of [Fe/H] for about 6000 stars from the HK survey that have measured or inferred colors. Note that the bi-modal character of this distribution is simply the result of the imperfect selection of candidate low-metallicity targets; the large number of stars with [Fe/H]

> -1.5 are the "mistakes". The 1200 stars with metallicities [Fe/H] < -2.0 are the sought-after objects. In the right panel a similar plot for the HES stars is shown. Although the total number of targets is only about half of the HK survey stars, the far higher efficiency of the HES in the identification of very metal-poor stars has greatly reduced the number of mistakes with [Fe/H] > -1.5. As a result, the total number of HES stars with [Fe/H] < -2.0 exceeds that of the HK survey by some 300 stars. There are almost twice as many [Fe/H] < -3.0 stars found in the HES as have been identified in the HK survey. The combined samples of stars from the HK and HES includes some 2700 stars with [Fe/H] < -2.0 and almost 400 stars with [Fe/H] < -3.0. Note that HK-survey stars that were rediscovered by the HES have not been eliminated yet. Hence, these numbers will be reduced somewhat in the final tally. When using these data we need to keep in mind two important caveats: (i) owing to the manner in which the surveys were constructed (to find the lowest metallicity stars), a bias is introduced as metallicity rises above [Fe/H] $= -2.5$. This can be corrected by comparing the shapes of the MDFs for stars in the range $-2.5 <$ [Fe/H] < -2.0 with those of the kinematically-based surveys discussed earlier, and derive a "bias correction factor" that can by applied to the numbers of stars in this interval to account for the fraction lost by the selection of candidates; (ii) a careful check on the abundance estimates for the many Carbon-rich MP stars that are found at low metallicity in both surveys must be performed: a possible underestimate of the metallicity of the cooler, or more carbon-enhanced stars, can take place. To correct for this one would have to evaluate such stars on a spectrum-by-spectrum basis, and either remove them or correct their abundances.

In any case, two features are evident from the resulting MDF shown in the Figure. The low-metallicity cutoff appears at [Fe/H]$= -4.0$, although the small numbers of stars with [Fe/H] < -3.5 still make the actual location of the cutoff somewhat uncertain. The other feature is the possibility of a small "bump" in the MDF at around [Fe/H] $= -2.5$, which needs to be confirmed by higher quality spectra.

r-Process Element Abundances

If the first stars are characterized by large masses, the first cosmic metals must come from PISN; thus, nucleosynthetic tracers provide at least a lower limit on early star formation. In zero-metallicity stars, the efficiency of nuclear-powered radial pulsations and radiatively driven stellar winds is greatly reduced, so mass loss is likely to be small. Although one of the most abundant elements produced by PISN, iron provides an uncertain measure of very massive star formation. In fact, the amount of silicon burned into iron in the final stellar explosion increases with the kinetic energy of the initial collapse, and thus the iron yield varies strongly with the core mass. Yields of other nuclides depend less on $M_\star$. While the mass yield of oxygen is quite high, the so-called α

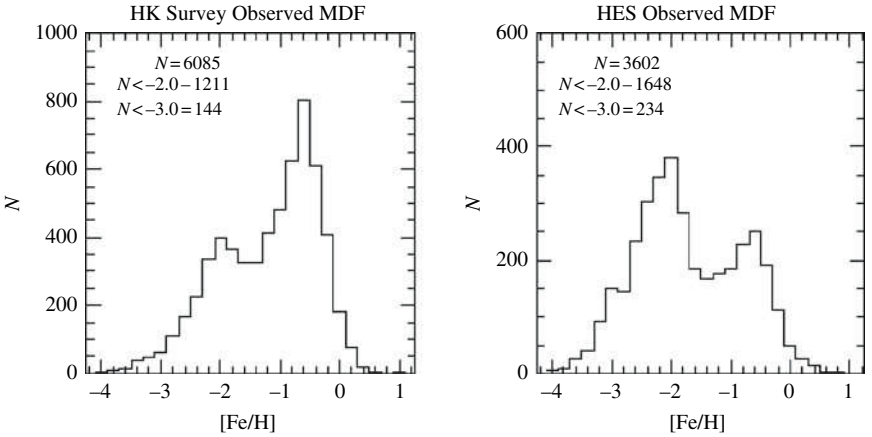

Fig. 6. Left: Metallicity Distribution Functions for the combined HK survey HES candidate metal-poor stars, for stars with [Fe/H] < -2.0. **Right**: The same distribution, but with the Carbon-rich extremely MP stars with [C/Fe] > 1.0 removed. From Beers et al. (2005) [5]

elements like Si, S, and Ca have enrichment factors relative to solar that are far above O. These trends are specific to very low/zero-metallicity massive stars; these elements are therefore excellent tracers of Population III star formation. According to the most recent estimates ([31]) no elements above the iron group are produced for the pair instability explosion. If these models are correct, the main prediction is that very massive stars should produce copious α-elements with very little associated r-element production.

The above prediction can now be confronted with observations of heavy r-process abundances in metal-poor halo stars as a function of [Fe/H]; so far this study has provided the following results (discussed in detail, for example, by [49]). For [Fe/H]< -3 the Ba abundance is roughly constant; however, over a small range in Fe abundance, $-3.1 <$ [Fe/H] < -2.5, there is a 2 dex spread in the Ba abundance. This fact can be explained as follows. PISN start to pollute the surrounding medium with very small amounts of r-process elements relative to their iron production rate. In this case, the initial iron enrichment up to [Fe/H]< -3 is almost exclusively due to these stars. The finite abundance at can be associated with very minor co-production of Ba in PISN. According to the critical metallicity concept (see Sect. 2), the IMF of the first stars should turn to "normal" once the metallicity of the gas rises above a certain value. At that point the main contributors to heavy elements production are core-collapse SNe. As they are not associated with significant co-production of iron, they will produce the wide scatter observed in Ba abundance over a small range in [Fe/H]. From such arguments, we would expect the trend seen in Fe to be most strongly exhibited in the α-elements. Indeed, Si and Ca show the same trend of a low Ba abundance at

very low metallicity followed by a wide scatter over a small range in metallicity. On the other hand, we would expect these trends to be absent in elements less abundantly produced by very massive stars, which do not experience significant initial enrichment before normal star formation takes over. This is indeed seen: carbon shows merely a linear correlation between C and Ba production with a great deal of scatter. This provides a first (albeit admittedly indirect) supporting evidence for a early cosmic phase of very massive star formation. Thus, Si and Ca (more suitably than Fe which shows a strong dependence on stellar mass) can be used to quantify the amount of early very massive star formation. The main result obtained from applying this technique to the available MP star catalogs indicates that the amount of ionizing photons corresponding to the abundances of these elements is sufficient to reionize the universe. In addition it shows the power of analyses that exploit local data for cosmological inferences.

Metallicity Distribution

As an alternative strategy to study the first stars, it appears that number counts of very low-metallicity Milky Way (MW) stars can not only provide evidences for "living fossils", i.e. stars formed immediately after the end of Dark Ages, but also probe the primordial IMF and hierarchical models to cosmic epochs ($z \geq 6$) currently unreachable to most experiments.

Suppose that Pop III stars are born inside the so-called Pop III objects, i.e. systems with total mass $\lesssim 2 \times 10^8$ $M_\odot$, which strongly rely on molecular hydrogen cooling in order to collapse, given their low virial temperatures. These objects are extremely fragile to the energy release by their own supernovae; as a result, their gas is quickly evacuated and they can only witness a single burst of star formation. The remnant is a "naked stellar object" a tiny agglomerate of long lived low-mass stars, essentially retaining the metallicity of the gas from which they formed. These are the stars that we are the prime candidates for the very low-metallicity stars observed in the MW halo. The calculation can be carried on along the following lines.

By using the extended Press-Schechter formalism it is possible to calculate the conditional probability that a virialized halo of mass M_1 at z_1 will become part of a more massive halo of mass M_0 at a later time, z_0. This allows us to calculate the mass functions of the fragments which through mergers will end up as part of a MW sized system today. Then one associates a maximum metallicity to the stars to be formed and contained within these halos through the mean mass weighted metallicity of the universe at each redshift, taken from cosmological simulations of metal enrichment. As a result of the merger process, these halos will probably contain a few stars with yet lower metallicities formed in halos of the preceding hierarchies; in this sense the aforementioned metallicities are to be considered as an upper limit for the stars of these halos.

The lower mass limit, $M_{\rm crit}$, excludes systems having masses and hence virial temperatures too low to allow H_2 cooling to lead to collapse (see Sect. 1. The upper limit, $M_{\rm by}$, corresponds to the upper mass at which the mechanical feedback due to supernova can expel the baryonic component entirely, thus quenching further star formation. Systems with masses larger than $M_{\rm crit}$ cool and the gas forms a disk in centrifugal equilibrium (often seen in the numerical simulations), with an exponential surface density profile fixed by the total baryonic mass and angular momentum. This last is given by the λ parameter, which is chosen at random from the distribution:

$$P(\lambda) = \frac{1}{\sigma_\lambda (2\pi)^{1/2}} \exp\left[\frac{-\ln^2(\lambda/<\lambda>)}{2\sigma_\lambda^2}\right] \frac{{\rm d}\lambda}{\lambda}, \tag{55}$$

with $<\lambda>= 0.05$ and $\sigma_\lambda = 1.0$. The initial dark halo profile is chosen such to have a central constant density core, as seen in present day dark halos from dwarf to cluster scales.

The final temperature of the gas in the disk is taken to be 300 K, as appropriate for gas where the main cooling mechanism is H_2. This temperature determines $c_{\rm s}$, the sound speed within the disk, which together with κ, the epicycle frequency obtained from the detailed final rotation curve, yields the Toomre's stability parameter for the disk:

$$Q = \frac{c_{\rm s}\kappa}{\pi G \Sigma}. \tag{56}$$

In the regions where $Q > 1$ the total tidal shears, together with the velocity dispersion, are sufficient to stabilize the disk locally against its self gravity. These regions are therefore not subject to star formation. Once the total gas mass turned into stars as a function of the metallicity has been determined, the most flexible approach is to use the Larson IMF to turn this into a total number of stars. This function offers a convenient parameterization of the IMF:

$${\rm d}N/{\rm d}\log\, m \propto (1 + m/m_{\rm s})^{-1.35}. \tag{57}$$

In the above equation $m_{\rm s}$ is a characteristic mass scale, of order 0.35 $M_\odot$ for a present day solar neighborhood IMF. This mass scale can be increased to explore the consequences of a top heavy IMF.

Figure 7, taken from [32] summarizes the results of such analysis. It is clear that theoretical predictions fall somewhat above the observational measurements, implying that the assumption of a constant $m_{\rm s} = 0.35$ $M_\odot$, essentially a present day solar neighborhood IMF, is not valid. Reconciling the estimates with the observations requires not only the assumption of a higher $m_{\rm s}$ in the past, but throughout the redshift range studied, an increasing trend for this value. The increase of the CMB temperature with redshift alone, implies an increase in the Jeans mass for the star forming clouds, once one is beyond $z = 2$, where the CMB temperature equals the ≈ 8 K of cold star forming

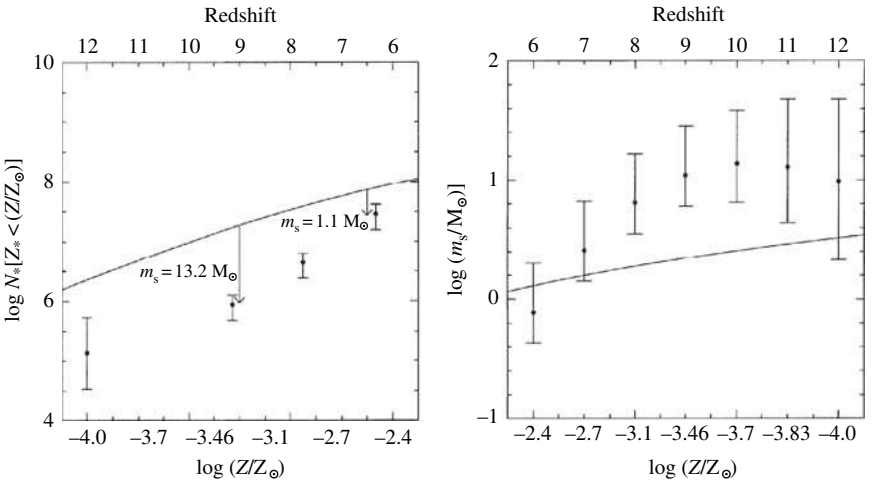

Fig. 7. Left: Average number of Pop III stars with metallicity lower than $[Z/Z_{\odot}]$ expected in MW sized galaxies today, for a solar neighborhood IMF (*solid curve*). The dots with error bars show the data from HK Survey. The *labeled arrows* show the values of $m_{\rm s}$ required to reconcile model and data, at $z = 6.5$ and $z = 9.0$. **Right**: Inferred values of $m_{\rm s}$ as a function of redshift (*points with error bars*). the *solid curve* gives the redshift evolution of the Jeans mass of cold star forming clouds, identified with $m_{\rm s}$, resulting from the temperature evolution of the CMB. From Hernandez & Ferrara (2001) [32]

clouds. It has therefore been suggested that at high redshift the IMF should be increasingly weighted towards larger stellar masses. The labeled arrows in Fig. 7 show the values of $m_{\rm s}$ required at $z = 6.5$ and $z = 9.0$, to match the observed points. The right panel illustrated the evolution of $m_{\rm s}$ inferred for the IMF of Pop III objects, by imposing that model and observations agree. The solid curve indicates the redshift evolution of the Jeans mass in star forming clouds, which we identify with $m_{\rm s}$. Although for $z > 6$ results are consistent with the increase in $m_{\rm s}$ coming solely from that in the Jeans mass due to the CMB, the trend for $6 < z < 9$ is clearly for a rise in $m_{\rm s}$ beyond what this effect can account for. For $z > 9$, the values of the characteristic mass appear to stabilize around the value $m_{\rm s} = 10$–$15\ {\rm M}_{\odot}$, indicative of a very top heavy IMF at these early epochs, although the larger error bars makes determining trends in this redshift range harder.

Although more modern versions of this idea should be explored to include the oncoming HES and SDSS data described above before drawing definite conclusions, it appears that the stellar relics in the Milky Way halo can teach us a lot about stars formation that occurred in the remote past.

4.2 Indirect Probes: The Near Infrared Background

Numerous arguments favor an excess contribution to the extragalactic background light between 1 μm and a few μm when compared to the expectation based on galaxy counts and Milky Way faint star counts.

Recent measurements of the Near Infrared Background (NIRB) shown an intensity excess with respect to observed light from galaxies in deep field surveys. The discrepancy is maximal at 1.4 μm, corresponding to 17–48 $\mathrm{nW\,m^{-2}sr^{-1}}$ (about 2–5 times the known galaxy contribution); significant discrepancies are found also at longer wavelengths. While these measurements are likely to be affected by certain systematics and issues related to the exact contribution from zodiacal light within the Solar System, one explanation is that a contribution to the NIRB originates from high redshift, very massive stars. The redshifted line emission from Lyα emitting galaxies at $z > 9$ would produce an integrated background in the near-infrared wavelengths observed today. In case this interpretation of the NIRB is correct, it would directly constrain the number of ionizing sources at high redshifts and thus would have direct implications on reionization. Until the recent availability of Spitzer and NICMOS/HST Ultra Deep Field data, there were no serious counterarguments to the interpretation of the entire excess being produced by very massive stars, thus providing a convincing, although indirect, probe of their existence. Instead, some support to this explanation came from the fact that the same stars can also account for the observed small ($\approx$ arcsec) angular fluctuations detected in the same bands. However, the picture of this problem has changed recently, as worked out in [56] Salvaterra & Ferrara (2006) and it is briefly summarized in the following.

The mean specific background intensity, $J(\nu_0, z_0)$, at redshift z_0 observed at frequency ν_0, produced by a population of sources (e.g. Pop III stellar clusters) characterized by a comoving emissivity $\epsilon_\nu(z)$, can be written as

$$J(\nu_0, z_0) = \frac{(1+z_0)^3}{4\pi} \int_{z_0}^{\infty} \epsilon_\nu(z) \mathrm{e}^{-\tau_{\mathrm{eff}}(\nu_0, z_0, z)} \frac{\mathrm{d}l}{\mathrm{d}z} \mathrm{d}z, \tag{58}$$

where $\nu = \nu_0(1+z)/(1+z_0)$, $\mathrm{d}l/\mathrm{d}z$ is the proper line element, τ_{eff} is the IGM effective optical depth (see Sect. 9) at ν_0 between redshift z_0 and z. The source term ϵ_ν can be written, in general, as a convolution of the light curve of the stellar cluster $l_\nu(t)$ with the source formation rate (per unit comoving volume). Under the approximation that the source formation rate is constant over the source lifetime τ_{lf}, and considering the average specific luminosity per unit mass, $\bar{l}_\nu$, over such timescale, the emissivity is given by

$$\epsilon_\nu(z) \simeq \bar{l}_\nu \tau_{\mathrm{lf}} f_\star \frac{\Omega_{\mathrm{b}}}{\Omega_{\mathrm{m}}} \frac{\mathrm{d}}{\mathrm{d}t} \int_{M_{\mathrm{min}}(z)}^{\infty} \frac{\mathrm{d}n}{\mathrm{d}M_{\mathrm{h}}}(M_{\mathrm{h}}, z) M_{\mathrm{h}} \mathrm{d}M_{\mathrm{h}}, \tag{59}$$

where $n(M_{\rm h}, z)$ is the comoving number density of halos of mass $M_{\rm h}$ at redshift z. The integral represents the collapsed mass per comoving volume contained in dark matter halos with mass above $M_{\rm min}(z)$. Two different threshold masses must be considered. depending on the relevant cooling agents: (i) atomic hydrogen (i.e. $T_{\rm vir} > 10^4\,{\rm K}$); and (ii) molecular hydrogen (see Sect. 1); $f_\star$ is the Pop III star formation efficiency. Once a (time-evolving) spectrum (see Sect. 3) for the massive Pop III stars has been selected the average flux $F_{\nu_0}(z, M_{\rm h})$ from Pop III stellar cluster of mass $M_{\rm cl}$ at redshift z can be derived:

$$F_{\nu_0} = \frac{M_{\rm cl}}{4\pi \Delta\nu_0\, d_{\rm L}(z)^2} \int\limits_{\nu_{\rm min}}^{\nu_{\rm max}} \bar{l}_\nu {\rm e}^{-\tau_{\rm eff}(\nu_0, z_0, z)} {\rm d}\nu, \tag{60}$$

where $d_{\rm L}(z)$ is the luminosity distance, $\Delta\nu_0$ is the instrumental bandwidth, and $\nu_{\rm min}, \nu_{\rm max}$ are the restframe frequencies corresponding to the observed ones.

The above equations allow to compute the diffuse NIRB background due to Pop III stars and compare it with the number count data. In order to determine the minimum requirements for the sources responsible for the NIR light, it is useful to consider the maximal contribution of both "normal" deep field galaxies and zodiacal light; it is then possible to determine the minimum star formation efficiency, $f_\star^{\rm m}$, necessary to fit the NIRB spectrum. Let us dub as *minimal* NIRB model, for clarity, that in which $f_\star^{\rm m} = 0.8\ (0.4)$ for H-cooling (H_2-cooling) halos. Finally, for the given value of $f_\star^{\rm m}$, we compute the surface density of Pop III stellar clusters required to fit the *minimal* model in the flux range F_{ν_0} and $F_{\nu_0} + {\rm d}F_{\nu_0}$ as

$$\frac{{\rm d}N}{{\rm d}\Omega dF_{\nu_0}}(F_{\nu_0}, z_{\rm end}) = \int\limits_{z_{\rm end}}^{\infty} \left(\frac{{\rm d}V_{\rm c}}{{\rm d}z{\rm d}\Omega}\right) n_{\rm c}(z, F_{\nu_0}){\rm d}z, \tag{61}$$

where ${\rm d}V_{\rm c}/{\rm d}z{\rm d}\Omega$ is the comoving volume element per unit redshift per unit solid angle, $n_{\rm c}(z, F_{\nu_0})$ is the comoving number of objects at redshift z with observed flux in $[F_{\nu_0}, F_{\nu_0} + {\rm d}F_{\nu_0}]$, given by

$$n_{\rm c}(z, F_{\nu_0}) \simeq \frac{{\rm d}M_{\rm h}}{{\rm d}F_{\nu_0}}(z, F_{\nu_0}) \tau_{\rm lf} \frac{{\rm d}^2 n}{{\rm d}M_{\rm h}{\rm d}t}(M_{\rm h}, z). \tag{62}$$

This set of prediction can be confronted with the high redshift galaxy number counts from the Spitzer/IRAC and HST/Nicmos instruments. In particular, [9] looked at the prevalence of galaxies at $z \approx 8 - 12$ by applying the dropout technique to the wide variety of deep F110W- and F160W-band (hereafter J_{110} and H_{160}, respectively) fields that have been imaged with the Near Infrared Camera and Multi-Object Spectrometer (NICMOS). The primary selection criterion for high-z sources is $J_{110} - H_{160} > 1.8$. Using this criterion, [9] found eleven sources. Eight of these are ruled out as credible

$z \simeq 10$ sources, either as a result of detection ($> 2\sigma$) blueward of J_{110} or because of their colors red-ward of the break ($H_{160} - K \approx 1.5$). The nature of the three remaining sources could not be assessed from the data, but this number appears consistent with the expected contamination from low-redshift interlopers. Hence, [9] concluded that the actual number of $z \simeq 10$ sources in the NICMOS parallel fields must be three or fewer.

Adopting the same selection criterion, the theory predicts that thousands of these sources should be detected. More precisely, 1165 (5634) J-dropouts should be observed for H-cooling (H_2 cooling) halos below the conservative magnitude limit $H_{160} = 28$. These numbers are at odd with the mere 3 (tentative) detections reported by [9].

What are the implications of this result? Even if all candidates turn out to be genuine $z = 10$ Pop III galaxies, the contribution to the NIRB is less than $0.05(0.04)\,\mathrm{nW\,m^{-2}\,sr^{-1}}$ in the J (H) band, i.e. $\approx 1/340$ of the minimum observed NIRB excess. Alternatively, if Pop III ionizing photons largely escape from star forming environments, the Lyα line could be produced in the intergalactic medium. The consequent strong surface brightness decrease would prevent detection and association with the parent galaxy; however, these Lyα photons would still contribute to the NIRB. The three NICMOS candidates (corresponding to $f_\star^{\rm m} = 1.5\%$) imply then a small NIRB contribution from Pop III, amounting only to $1/24$ of the minimum observed excess. Again, these sources would be below detectability for *Spitzer* at larger wavelengths.

The conclusion is that if the NIRB excess is made entirely by high redshift Pop III clusters, these should have already been observed in deep galaxy searches in large numbers. The few (possibly none!) detections of such objects offer little room for the Pop III explanation of the NIRB excess, leaving us with a puzzling question concerning the origin of this mysterious background component.

4.3 Direct Probes at Intermediate Redshifts

Although a critical metallicity-induced transition must have occurred at high redshift in the bulk of cosmic star formation sites, massive Pop III stars are likely to have left some imprints that are observable at intermediate ($3 < z < 6$) redshifts. One point that might have important observational consequences is the fact that cosmic metal enrichment has proceeded very inhomogeneously, with regions close to star formation sites rapidly becoming metal-polluted and overshooting $Z_{\rm cr}$, and others remaining essentially metal-free. Thus, the two modes of star formation, Pop III and normal, must have been active at the same time and possibly down to relatively low redshifts, opening up the possibility of detecting Pop III stars. This is particularly important as metal-free stars are yet to be detected. Thus, it might well be that the best approach to detect the first stellar clusters is to develop careful observation strategies of high-redshift Pop III host candidates.

Reference [58] have worked out a comprehensive model to predict the probability of spotting massive stars at intermediate redshifts. As metal-free stars are powerful Lyα line emitters, it is natural to use this indicator as a first step in any search for primordial objects. This is even more promising as surveys aimed at finding young, high redshift systems have already discovered a considerable number of such emitters. In principle, the calculation of the Lyα luminosity, L_α, from a stellar cluster is very simple, being directly related to the corresponding hydrogen ionizing photon rate, $Q(H)$, by

$$L_\alpha = c_{\rm L}(1 - f_{\rm esc})Q(H), \tag{63}$$

where $c_{\rm L} \equiv 1.04 \times 10^{-11}$ erg and $f_{\rm esc}$ is the escape fraction of ionizing photons from the galaxy. As this escape fraction is relatively uncertain, we adopt here an educated guess of $f_{\rm esc} = 0.2$, which is based on a compilation of both theoretical and observational results; $Q(H)$ is computed from evolutionary stellar models.

Having calculated the Lyα emission, we can then derive the probability that a given high-redshift Lyα detection is due to a cluster of Pop III stars (modulo the assumption of a IMF and a star formation efficiency). The resulting probabilities are displayed in Fig. 8. In this figure, the isocontours in the Lyα luminosity-redshift plane indicate the probability to find Pop III objects (i.e. galaxies hosting Pop III stars) in a given sample of Lyα emitters for various feedback strenghts (feedback processes will be discussed in detail in Sect. 6, parameterized by the value of $\varepsilon_{\rm III}$ (for the definition of this parameter see Table 2). These are compared with the available data points and a line corresponding to a typical detection flux threshold of 1.5×10^{-17} ergs cm^{-2} s^{-1}.

In this figure, we see that Pop III objects populate a well-defined region of the L_α-redshift plane, whose extent is governed by the feedback strength. Note that the lower boundary is practically unaffected by changes in $\varepsilon_{\rm III}$, as most Pop III objects are in a limited mass range such that they are large enough to cool efficiently, but small enough that they are not clustered near areas of previous star formation. At lower $\varepsilon_{\rm III}$ values, the non-zero probability area widens considerably: in this case, a smaller volume of the universe is polluted and Pop III star formation continues at lower redshifts and in the higher mass, more luminous objects that form later in hierarchical models. Above the typical flux threshold, Lyα emitters are potentially detectable at all redshifts beyond 5. Furthermore, the fraction of Pop III objects increases with redshift, independent of the assumed feedback strength. For the fiducial case $\varepsilon_{\rm III} = 10^{-3}$, for example, the fraction is only a few percent at $z = 4$ but increases to approximately 15% by $z = 6$. We then conclude that the Lyα emission from already observed high-z sources can indeed be due to Pop III objects, if such stars were biased to high masses. Hence collecting large data samples to increase the statistical leverage may be crucial for detecting the elusive first stars.

Although this type of investigations is still in its infancy, encouraging results are rapidly accumulating. For example, [24] report the serendipitous discovery of a very red, gravitationally lensed, star-forming galaxy (the Lynx arc), with a redshift of 3.357. The analysis of the emission lines demonstrates that the stars powering the H II region are very hot ($T \approx 80{,}000$ K), suggestive of very massive stars according to our discussion above. In addition the metal enrichment pattern analysis shows a substantial overabundance of silicon, which might be an additional nucleosynthetic signature of past PISN a Population III cluster. Although there could be different explanations for the result (e.g. [70]), we might be seen essentially pristine star formation activity in act relatively close to us for the first time. This is only the first step towards

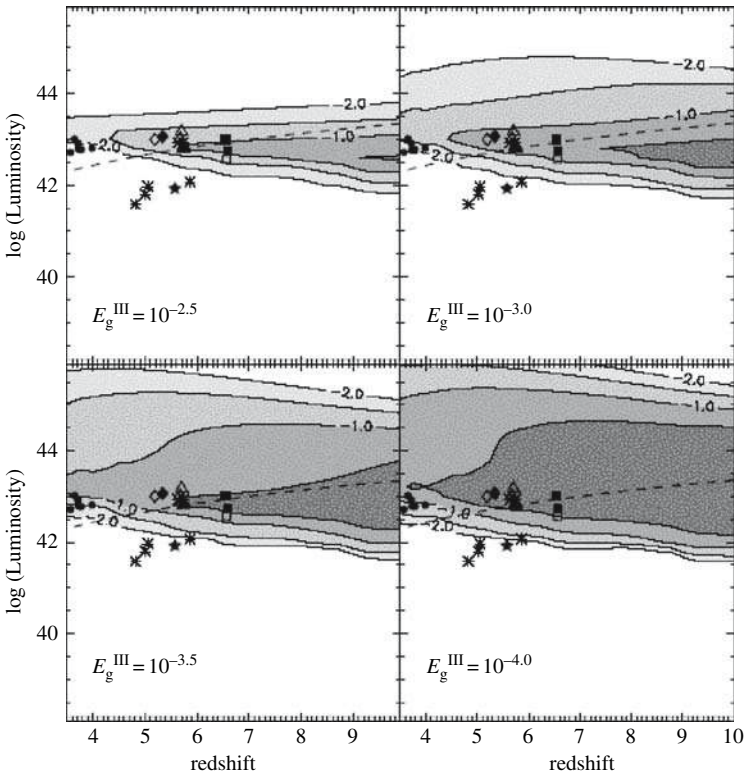

Fig. 8. Fraction of Pop III objects as a function of Lyα luminosity and redshift. Isocontours of fractions $\geqslant 10^{-2}, 10^{-1.5}, 10^{-1}$ and $10^{0.5}$ are shown. Burst-mode star formation with a $f_\star^{\rm II} = f_\star^{\rm III} = 0.1$ is assumed for all objects. In the Pop III case however, a lower cutoff mass of 50 $M_\odot$ is assumed in this figure. Each panel is labeled by the assumed $\varepsilon_{\rm III}$ value. For reference, the dashed line gives the luminosity corresponding to an observed flux of 1.5×10^{-17} ergs cm^{-2} s^{-1}, and the various points correspond to observed galaxies. From Scannapieco et al. (2003) [58]

more stringent studies of this type which might turn out to constitute a new technique to search for long sought-after metal-free stars.

5 Blastwaves and Winds

A shock represents a "hydrodynamic surprise", in the sense that such a wave travels faster than signals in the fluid, which are bound to propagate at the gas sound speed, $c_{\rm s}$. It is common to define the shock speed, $v_{\rm s}$, in terms of $c_{\rm s}$ by introducing the shock Mach number, M:

$$v_{\rm s} = Mc_{\rm s} = M(\gamma P/\rho)^{1/2}, \tag{64}$$

where P and ρ are the gas pressure and density, respectively. The change of the fluid velocity v is governed by the momentum equation

$$\rho\frac{\partial v}{\partial t} + \rho v\cdot\nabla v = -\nabla P - \frac{1}{8\pi}\nabla B^2 + \frac{1}{4\pi}B\cdot\nabla B, \tag{65}$$

where B is the magnetic field. The density is determined by the continuity equation

$$\frac{\partial\rho}{\partial t} + v\cdot\nabla\rho = -\rho\nabla v. \tag{66}$$

5.1 Hydrodynamics of Shock Waves

The large-scale properties of a shock in a perfect gas, with $B = 0$ are fully described by the three ratios v_2/v_1 ρ_2/ρ_1 and P_2/P_1, determined in terms of the conditions ahead of the shock (i.e. v_1, ρ_1 and P_1) by the Rankine-Hugoniot jump conditions resulting from matter, momentum and energy conservation:

$$\rho_1 v_1 = \rho_2 v_2, \tag{67}$$

$$P_1 + \rho_1 v_1^2 = P_2 + \rho_2 v_2^2, \tag{68}$$

$$v_2\left(\frac{1}{2}\rho_2 v_2^2 + U_2\right) - v_1\left(\frac{1}{2}\rho_1 v_1^2 + U_1\right) = v_1 P_1 - v_2 P_2, \tag{69}$$

where U is the internal energy density of the fluid. If the fluid behaves as a perfect gas on each side of the front we have $U = P/(\gamma - 1)$.

In the adiabatic limit it is sufficient to consider the mass and momentum equations only. By solving such system we obtain:

$$\frac{v_2}{v_1} = \frac{\rho_1}{\rho_2} = \frac{\gamma - 1}{\gamma + 1} + \frac{2}{\gamma + 1}\frac{1}{M^2}, \tag{70}$$

where the Mach number is defined as $M^2 = v_1^2/c_1^2$ and c_1 is the sound velocity in the unperturbed medium ahead of the shock. For strong shocks, i.e. $M \gg 1$ and $\gamma = 5/3$, $\rho_2/\rho_1 = 4$ and

$$T_2 = \frac{3\mu}{16k} v_{\rm s}^2, \tag{71}$$

where $v_{\rm s} = v_1 - v_2$

In the isothermal limit, i.e. when the cooling time of the post-shock gas becomes extremely short and the initial temperature is promptly re-established,

$$\frac{\rho_2}{\rho_1} = \frac{v_1^2}{c_{\rm s}^2} = M^2, \tag{72}$$

where $c_{\rm s}$ is the sound velocity on both sides of the shock. While in the adiabatic case the compression factor is limited to four, in an isothermal shock much larger compressions are possible due to the softer equation of state characterizing the fluid.

5.2 Hydromagnetic Shock Waves

Assuming B is parallel to the shock front (so that $B \cdot \nabla B = 0$), (68) can be replaced by

$$P_1 + \rho_1 v_1^2 + \left(\frac{B_1^2}{8\pi}\right) = P_2 + \rho_2 v_2^2 + \left(\frac{B_2^2}{8\pi}\right). \tag{73}$$

Magnetic flux conservation requires $B_1/\rho_1 = B_2/\rho_2$, so that in the adiabatic limit the magnetic pressure $B^2/8\pi$ increases by only a factor 16.

In the isothermal limit the compression depends only linearly on the Alfvénic Mach number $M_{\rm a}$

$$\frac{\rho_2}{\rho_1} = \sqrt{2}\frac{v_1}{v_{\rm a,1}} = M_{\rm a}, \tag{74}$$

where $v_{\rm a} = B/\sqrt{4\pi\rho_1}$ is the Alfvén velocity. Hence, some of the shock kinetic energy is stored in the magnetic field lines rather than being used to compress the gas as in the field-free case; in other words a magnetized gas is less compressible than an unmagnetized one.

Structure of Radiative Shocks

The structure of a shock can be divided into four regions ([45]). The radiative precursor is the region upstream of the shock in which radiation emitted by the shock acts as a precursor of the shock arrival. The shock front is the region in which the relative kinetic energy difference of the shocked and un-shocked gas is dissipated. If the dissipation is due to collisions among the atoms or

molecules of the gas, the shock is collisional. On the other hand, if the density is sufficiently low that collisions are unimportant and the dissipation is due to the collective interactions of the particles with the turbulent electromagnetic fields, the shock is collisionless. Next comes the radiative zone, in which collisional processes cause the gas to radiate. The gas cools and the density increases. Finally, if the shock lasts long enough, a thermalization region is produced, in which radiation from the radiative zone is absorbed and re-radiated.

5.3 Supernova Explosions

In the explosion of a supernova three different stages in the expansion may be distinguished. In the initial phase the interstellar material has little effect because of its low moment of inertia; the velocity of expansion of the supernova envelope will then remain nearly constant with time. This phase terminates when the mass of the swept up gas is about equal to the initial mass $M_{\rm e}$ expelled by the supernova, i.e. $(4\pi/3)r_{\rm s}^3\rho_1 = M_{\rm e}$, where ρ_1 is the density of the gas in front of the shock and $r_{\rm s}$ is the radius of the shock front. For $M_{\rm e} = 0.25\ {\rm M}_\odot$ and $\rho_1 = 2\times10^{-24}\ {\rm g\ cm^{-3}}$, $r_{\rm s} \simeq 1$ pc, which will occurs about 60 years after the explosion. During the second phase (Sedov-Taylor phase) the mass behind the shock is determined primarily by the amount of interstellar gas swept up, but the energy of this gas will remain constant. When radiative cooling becomes important (radiative phase), the temperature of the gas will fall to a relatively low value. The motion of the shock is supported by the momentum of the outward moving gas and the shock may be regarded as isothermal. The velocity of the shell can be computed from the condition of momentum conservation (snowplow model). This phase continues until $v_{\rm s} = \max\{v_{\rm t}, c_{\rm s}\}$, where $v_{\rm t}$ is the turbulent velocity of the gas, when the shell loses its identity due to gas random motions.

Blastwave Evolution

To describe the evolution of a blastwave, the so-called thin shell approximation is frequently used. In this approximation, one assumes that most of the mass of the material swept up during the expansion is collected in a thin shell. We start by applying (we neglect the ambient medium density and shell self-gravity) the virial theorem to the shell:

$$\frac{1}{2}\frac{{\rm d}^2 I}{{\rm d}t^2} = 2E = 2(E_{\rm k} + E_{\rm t}), \tag{75}$$

where I is the moment of inertia, and $E_{\rm k}$ ($E_{\rm t}$) is the kinetic (thermal) energy of the shell. Let us introduce the structure parameter K in the following form:

$$K = \frac{1}{MRv_2} \int_0^R \mathrm{d}m\, rv, \tag{76}$$

where M is the ejected mass, R the shell radius. Substituting in (75) with $v = v_2(r/R)$, we obtain

$$K \frac{\mathrm{d}}{\mathrm{d}t}(Mrv_\mathrm{s}) = 2E(v_\mathrm{s}/v_2). \tag{77}$$

To recover the Sedov-Taylor adiabatic phase we integrate up to R and assume $\rho(R) = \rho_0 R^{-m}$, $E = const.$, $K = (6-2m)/(7-2m)$

$$R(t) = \left[\frac{(5-m)E}{\pi\rho_0 K}\right]^{1/(5-m)} t^{2/(5-m)}. \tag{78}$$

The previous expression can be obtained also via a dimensional approach. In fact, in the adiabatic phase the energy is conserved and hence

$$E \propto Mv_\mathrm{s}^2 \propto \rho_0 R^{5-m} t^{-2}, \tag{79}$$

so that

$$R^{5-m} \propto \frac{E}{\rho_0} t^2; \tag{80}$$

the fraction of the total energy transformed in kinetic form is

$$E_\mathrm{k} = \frac{1}{2} KM \left(\frac{3}{4} v_\mathrm{s}^2\right) = \frac{3E}{2(5-m)}. \tag{81}$$

We note that in the limit of a homogeneous ambient medium ($m = 0$) we recover the standard Sedov-Taylor evolution $R \propto t^{2/5}$, with 30% of the explosion energy in kinetic form.

A similar reasoning can be followed to recover the radiative phase behavior. In this phase $v_2 = v_s$, i.e. the shock wave is almost isothermal and $K = 1$, i.e. the shell is thin. Under these assumptions the energy is lost at a rate

$$\dot{E} = -4\pi R^2 \left(\frac{1}{2}\rho_0 v_\mathrm{s}^2\right). \tag{82}$$

There are two possible solutions of (75) in this case: (a) the pressure-driven snowplow in which

$$R(t) \propto t^{2/(7-m)}, \tag{83}$$

or the momentum-conserving solution

$$R(t) \propto t^{1/(4-m)}. \tag{84}$$

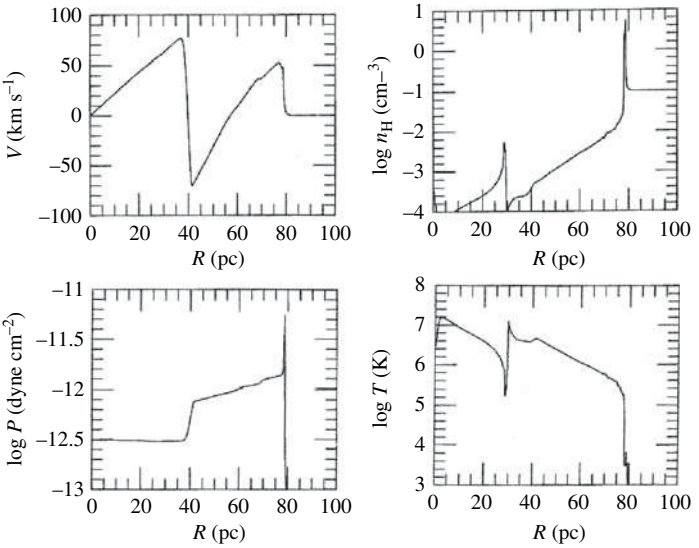

Fig. 9. The structure of spherical supernova remnant calculated using a high-resolution numerical hydrodynamical code. The remnant age is $t = 5 \times 10^5$ yr. From Cioffi et al. (1988) [16]

In both cases, the expansion proceeds at a lower rate with respect to the adiabatic phase.

In order to study the evolution of an explosion in a more realistic way it is necessary to resort to numerical simulations as the one shown in Fig. 9. From there the inner structure of the remnant is clearly visible: indeed most of the mass is concentrated in a thin shell (located at about 80 pc after 0.5 Myr from the explosion); a reverse shock is also seen which is proceeding towards the center and thermalizing the ejecta up to temperatures of the order of 10^7 K. In between the two shocks is the contact discontinuity between the ejecta and the ambient medium, through which the pressure remains approximately constant.

5.4 Cosmological Blastwaves

The application of the previous theory to cosmological explosions has to cope with the fact that the background medium is expanding. The simplest explosion models which describes such situation is a three-phase model ([44, 67]) constituted by (i) a dense, cool spherical shell of outer radius R and thickness $R\delta$, containing a fraction $(1 - f_{\rm m})$ of the total baryonic mass enclosed; (ii) A uniform neutral intergalactic medium of density $\rho_{\rm m} = \rho_{\rm b} + \rho_{\rm d}$, including the contribution of baryonic and dark matter, respectively; (iii) a hot, isothermal plasma of pressure p and temperature T inside the shell. The shell is essentially driven by the thermal pressure of the interior gas, which has to

overcome the inertia of the swept up material and gravity force. If one assumes that the shell sweeps up almost all the IGM gas ahead, then its mass can be written as $m(t) = (4/3)\pi R^3(1 - f_{\rm m})\rho_{\rm b}$, with $f_{\rm m} \ll 1$. One can then write the mass, momentum and energy conservation for the shell motion in an expanding universe, for which $\dot{\rho}/\rho = -3H$:

$$\frac{\dot{m}}{m} = (R^3\rho_b)^{-1}\frac{\rm d}{{\rm d}t}(R^3\rho_b) = 3\left(\frac{\dot{R}}{R} - H\right) \text{ for } \frac{\dot{R}}{R} > H; \text{zero otherwise.} \quad (85)$$

$$\frac{\rm d}{{\rm d}t}\dot{R} = \frac{8\pi pG}{\Omega_b H^2 R} - \frac{3}{R}(\dot{R} - HR)^2 - (\Omega - \frac{1}{2}\Omega_{\rm b})\frac{H^2 R}{2} \quad (86)$$

$$\dot{E} = L - p{\rm d}V/{\rm d}t = L - 4\pi pR^2\dot{R} \quad (87)$$

The physics expressed by these equations can be understood as follows. The mass of the shell increases in time as long as it velocity is larger than the Hubble expansion. The newly added material must be accelerated to the shell velocity, thus resulting in a net braking force. The internal pressure term has therefore to counteract both this force and the gravitational one. The third equation expresses energy conservation: the luminosity L incorporates all sources of heating and cooling of the plasma. Typically these include the supernova energy injection, cooling by Compton drag against the CMB, bremsstrahlung and ionization losses. These equations have in general to be solved numerically. However, some of the main features of the blastwave evolution can be identified by a simple dimensional analysis. Three different regimes can be isolated during the evolution. At first, for bubble ages $t \ll t_{\rm H}$ the gravity and Hubble flow are negligible and one can easily show that $R \propto t^{3/5}$ as we have already noticed in the case of non-cosmological blastwaves previously. When $t \approx t_{\rm H}$, the behavior becomes quite complicated as several effects control the evolution at the same time: SN explosions have ceased, slowing the expansion; cooling and $p{\rm d}V$ work reduces p essentially to zero: the blast enters the momentum conserving phase in which $R \propto t^{1/4}$; gravity becomes important, decelerating the expansion. Finally, when the age becomes larger than the Hubble time $t \gg t_{\rm H}$, the shell gets frozen into the Hubble flow, i.e. $R \propto t^{2/3}$ for $\Omega = 1$. It is very instructive to compute the evolution of the energy content of the expanding bubble as a function of redshift. Figure 10 illustrates how the energy of the supernovae is distributed among kinetic, thermal and gravitational energies; the role of the various cooling processes is also made clear.

5.5 Porosity

For many applications it is useful to compute the cosmic volume occupied by (either blastwaves or ionized) bubbles produced by sources (galaxies, quasars).

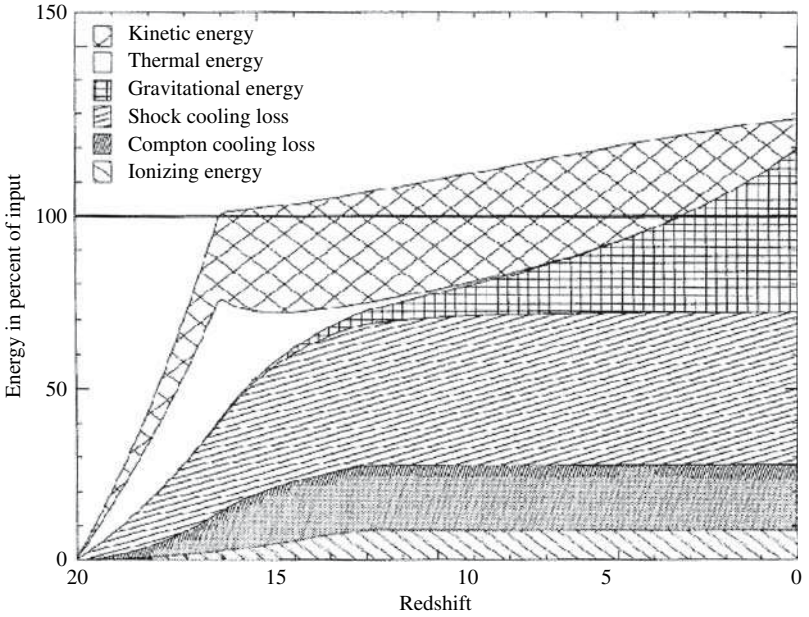

Fig. 10. Illustrative graph of the evolution of the energy content of an expanding bubble as a function of redshift

This requires an assumption about the number density distribution of galaxy halos and therefore a background cosmological model. This is usually accomplished by using the so-called Press-Schechter formalism, according to which the fraction of mass in gravitationally bound objects of total mass larger than M at redshift z is given by

$$f = 1 - \mathrm{erf}\left[\frac{\delta_\mathrm{c}}{\sqrt{2}\sigma(M)}\right] \tag{88}$$

where σ is the linearly extrapolated rms mass fluctuation in a sphere of given radius (usually $8\,h^{-1}$ Mpc). The expression for such quantity is

$$\sigma^2 \equiv \left(\frac{\sigma_0}{1+z}\right)^2 \propto \frac{\sigma_0}{(1+z)^2}\int_0^\infty \mathrm{d}k P(k) W^2(k) \tag{89}$$

where $P(k)$ is the fluctuation power spectrum and W a suitable window function. The value of the critical overdensity is $\delta_\mathrm{c} = 1.69$, the linearly extrapolated value at which a spherically symmetric perturbation virializes. If $R(z, z_\star)$ denotes the radius t redshift z of a bubble created at $z = z_\star$ by a source of baryonic mass $M_\mathrm{b} = \Omega_\mathrm{b} M$, then the filling factor, i.e. the fraction of cosmic volume occupied by the bubbles, is

$$\phi(z) = -\int_z^\infty \mathrm{d}z_\star \frac{4}{3}\pi R(z, z_\star)^3 \frac{\rho_\mathrm{b}}{M_\mathrm{b}} \frac{\mathrm{d}f(z_\star)}{\mathrm{d}z_\star} \tag{90}$$

This value can exceed unity by construction: this situation obviously corresponds to the occurrence on the so-called overlap, during which a point in space is overrun by different bubbles. For this reason it is more convenient to introduce the porosity parameter, P, which is defined as

$$P \equiv 1 - \mathrm{e}^{-\phi}. \tag{91}$$

Note that the above argument assumes that the sources are spatially uncorrelated (i.e. Poisson-distributed). If this is not the case (as indeed happens for galaxy populations) the values given by the previous formulae represent an overestimate of the true ones.

6 Mechanical Feedbacks in Cosmology

The word "feedback" is by far one of the most used ones in modern cosmology where it is applied to a vast range of situations and astrophysical objects. However, for the same reason, its meaning in the context is often unclear or fuzzy. Hence a review on feedback should start from setting the definition of feedback on a solid basis. We have found quite useful to this aim to go back to the Oxford Dictionary from where we take the following definition:

Fee'dback *n. 1. (Electr.) Return of fraction of output signal from one stage of circuit, amplifier, etc. to input of same or preceding stage (***positive, negative***, tending to increase, decrease the amplification, etc). 2. (Biol., Psych., etc) Modification or control of a process or system by its results or effects, esp. by difference between the desired and actual results.*

In spite of the broad description, we find this definition quite appropriate in many ways. First, the definition outlines the fact that the concept of feedback invokes a back reaction of a process on itself or on the causes that have produced it. Secondly, the character of feedback can be either negative or positive. Finally, and most importantly, the idea of feedback is intimately linked to the possibility that a system can become self-regulated. Although some types of feedback processes are disruptive, the most important ones in astrophysics are probably those that are able to drive the systems towards a steady state of some sort. To exemplify, think of a galaxy which is witnessing a burst of star formation. The occurrence of the first supernovae will evacuate/heat the gas thus suppressing the star formation activity. Such feedback is then acting back on the energy source (star formation); it is of a negative type, and it could regulate the star formation activity in such a way that only a sustainable amount of stars is formed (regulation). However, feedback can fail to produce such regulation either in small galaxies where the gas can be ejected by the first SNe or in cases when the star formation timescale is too

short compared to the feedback one. As we will see there are at least three types of feedback, and even the mechanical feedback described in the example above is part of a larger class of feedback phenomena related to the energy deposition of massive stars. We then start by briefly outline the importance of feedback processes in cosmology.

6.1 The Need for Feedback in Cosmology

One of the main aims of physical cosmology is to understand in detail galaxy formation starting from the initial density fluctuation field with a given spectrum (typically a CDM one). Such ab initio computations require a tremendous amount of physical processes to be included before it becomes possible to compare their predictions with experimental data. In particular, it is crucial to model the interstellar medium of galaxies, which is know to be turbulent, have a multi-phase thermal structure, can undergo gravitational instability and form stars. To account for all this complexity, in addition one should treat correctly all relevant cooling processes, radiative and shock heating (let alone magnetic fields!). This has proven to be essentially impossible even with present day best supercomputers. Hence one has to resort to heuristic models where simplistic prescriptions for some of these processes must be adopted. Of course such an approach suffers from the fact that a large number of free parameters remains which cannot be fixed from first principles. These are essentially contained in each of the ingredients used to model galaxy formation, that is, the evolution of dark halos, cooling and star formation, chemical enrichment, stellar populations. Fortunately, there is a large variety of data against which the models can be tested: these data range from the fraction of cooled baryons to cosmic star formation histories, the luminous content of halos, luminosity functions and faint galaxy counts. The feedback processes enter the game as part of such iterative try-and-learn process to which we are bound by our ignorance in dealing with complex systems as the galaxies. Still, we are far from a full understanding of galaxies in the framework of structure formation models. The hope is that feedback can help us to solve some "chronic" problems found in cosmological simulations adopting the CDM paradigm. Their (partial) list includes:

1. **Overcooling**: The predicted cosmic fraction of cooled baryons is larger than observed. Moreover models predict too many faint, low mass galaxies;
2. **Disk Angular Momentum**: The angular momentum loss is too high and galactic disk scale lengths are too small;
3. **Halo Density Profiles**: Profiles are centrally too concentrated;
4. **Dark Satellites**: Too many satellites predicted around our Galaxy.

6.2 The Overcooling Problem

Among the various CDM problems, historically the overcooling has been the most prominent and yet unsolved one. In its original formulation it has been

first spelled out by [71]. Let us assume that, as a halo forms, the gas initially relaxes to an isothermal distribution which exactly parallels that of the dark matter. The virial theorem then relates the gas temperature $T_{\rm vir}$ to the circular velocity of the halo $V_{\rm c}$,

$$kT_{\rm vir} = \frac{1}{2}\mu m_{\rm p} V_{\rm c}^2 \text{ or } T_{\rm vir} = 36 V_{\rm c,km/s}^2 {\rm K}, \tag{92}$$

where $\mu m_{\rm p}$ is the mean molecular weight of the gas. At each radius in this distribution we can then define a cooling time as the ratio of the specific energy content to the cooling rate,

$$t_{\rm cool}(r) = \frac{3\rho_{\rm g}(r)/2\mu m_{\rm p}}{n_{\rm e}^2(r)\Lambda(T)}, \tag{93}$$

where $\rho_{\rm g}(r)$ is the gas density profile and $n_e(r)$ is the electron density. $\Lambda(T)$ is the cooling function. The cooling radius is defined as the point where the cooling time is equal to the age of the universe, i.e. $t_{\rm cool}(r_{\rm cool}) = t_{\rm Hubble} = H(z)^{-1}$.

Considering the virialized part of the halo to be the region encompassing a mean overdensity is 200, its radius and mass are defined by

$$r_{\rm vir} = 0.1 H_0^{-1}(1+z)^{3/2} V_{\rm c}, \tag{94}$$

$$M_{\rm vir} = 0.1(GH_0)^{-1}(1+z)^{3/2} V_{\rm c}^3. \tag{95}$$

Let us distinguish two limiting cases. When $r_{\rm cool} \gg r_{\rm vir}$ (accretion limited case), cooling is so rapid that the infalling gas never comes to hydrostatic equilibrium. The supply of cold gas for star formation is then limited by the infall rate rather than by cooling. The accretion rate is obtained by differentiating (95) with respect to time and multiplying by the fraction of the mass of the universe that remains in gaseous form:

$$\dot{M}_{\rm acc} = f_{\rm g}\Omega_{\rm g}\frac{\rm d}{{\rm d}t} 0.1(GH_0)^{-1}(1+z)^{3/2} V_{\rm c}^3 = 0.15 f_{\rm g}\Omega_{\rm g} G^{-1} V_{\rm c}^3. \tag{96}$$

Note that, except for a weak time dependence of the fraction of the initial baryon density which remains in gaseous form, $f_{\rm g}$, this infall rate does not depend on redshift.

In the opposite limit, $r_{\rm cool} \ll r_{\rm vir}$ (quasi-static case), the accretion shock radiates only weakly, a quasi-static atmosphere forms, and the supply of cold gas for star formation is regulated by radiative losses near $r_{\rm cool}$. A simple expression for the inflow rate is

$$\dot{M}_{\rm qst} = 4\pi\rho_{\rm g}(r_{\rm cool}) r_{\rm cool}^2 \frac{\rm d}{{\rm d}t} r_{\rm cool}. \tag{97}$$

The gas supply rates predicted by (96) and (97) are illustrated in Fig. 11. In any particular halo, the rate at which cold gas becomes available for star formation is the minimum between $\dot{M}_{\rm acc}$ and $\dot{M}_{\rm qst}$.

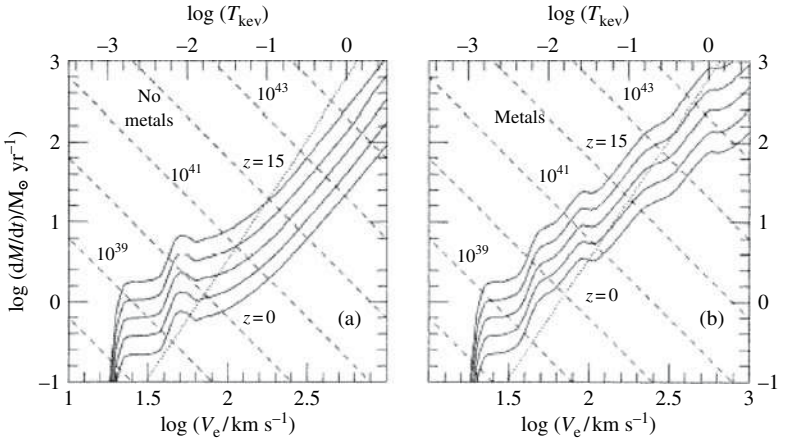

Fig. 11. Gas infall rate and cooling rates in dark matter halos as a function of circular velocity and redshift. The infall rate (*dotted line*) is essentially independent of redshift; the cooling rates (*solid lines*) are given for redshift $z = 0, 1, 3, 7$ and 15 (*from bottom to top*). Dashed lines give present-day X-ray luminosities in erg s^{-1} produced by gas cooling at the given rate in each halo. The predicted temperature of this emission is given on the upper abscissa. (**a**) A cooling function for gas of zero metallicity is assumed. (**b**) A cooling function for metal enriched gas. From White & Frenk (1991) [71]

The bolometric X-ray luminosity of the region within the cooling radius of a galactic cooling flow is

$$L_{\rm X} = 2.5 \dot{M}_{\rm cool} V_{\rm c}^2 .$$

The predictions of this formula are superposed on Fig. 11 (dashed lines). For large circular velocities the mass cooling rates correspond to quite substantial X-ray luminosities.

Integrating the gas supply rates $\dot{M}_{\rm cool}$ over redshift and halo mass distribution, we find that for $\Omega_{\rm b} = 0.04$ *most of the gas is used before the present time*. This is unacceptable, since the density contributed by the observed stars in galaxies is less than 1% of the critical density. So, the star formation results to be too rapid without a regulating process, i.e. feedback.

To solve this puzzle, [38] proposed that the energy input from young stars and supernovae could quench star formation in small protogalaxies before more than a small gas fraction has been converted into stars.

Stellar energy input would counteract radiative losses in the cooling gas and tend to reduce the supply of gas for further star formation. We can imagine that the star formation process is self-regulating in the sense that the star formation rate $\dot{M}_\star$ takes the value required for heating to balance dissipation in the material which does not form stars. This produces the following prescription for the star formation rate:

$$\dot{M}_\star(V_\mathrm{c}, z) = \epsilon(V_\mathrm{c}) \min(\dot{M}_\mathrm{acc}, \dot{M}_\mathrm{qst}), \tag{98}$$

$$\epsilon(V_\mathrm{c}) = [1 + \epsilon_0 (V_0/V_\mathrm{c})^2]^{-1}, \tag{99}$$

where V_0 and ϵ_0 are the typical velocity scale and efficiency, respectively. For large V_c the available gas turns into stars with high efficiency because the energy input is not sufficient to prevent cooling and fragmentation; for smaller objects the star formation efficiency ϵ is proportional to V_c^2. The assumption of self regulation at small V_c seems plausible because the time interval between star formation and energy injection is much shorter than either the sound crossing time or the cooling time in the gaseous halos. However, other possibilities can be envisaged. For example, [18] suggested that supernovae not only would suppress cooling in the halo gas but would actually expel it altogether. The conditions leading to such an event are discussed next.

6.3 Dwarf Galaxies: A Feedback Lab

The observation of dwarf galaxies has shown well-defined correlations between their measured properties, and in particular

1. Luminosity-Radius $L \propto R^r$ with $r = 4$
2. Luminosity-Metallicity $L \propto Z^z$ with $z = 5/2$
3. Luminosity-Velocity dispersion $L \propto V^v$ with $v = 4$

A simple model can relate the observed scaling parameters r, z, and v to each other ([18]). Consider a uniform cloud of initial mass $M_\mathrm{i} = M_\mathrm{g} + M_\star$ (M_g is the mass of gas driven out of the system, and $M_\star$ is the mass in stars) in a sphere of radius R_i which undergoes star formation. The metallicity for a constant yield in the instantaneous recycling approximation is given by $Z = y \ln(1 + M_\star/M_\mathrm{g})$ where y are the yields.

Let us impose simple scaling relations: $M_\mathrm{i} \propto M$ (gas mass proportional to the dark matter mass), $R \simeq R_\mathrm{i}$, and $V \simeq V_\mathrm{i}$ (gas loss has no dynamical effect). From the observed scaling relation we have $L \propto R^r \propto V^v$. Now we should write the analogous relations for structure and velocity that hold before the removal, i.e. $M_\mathrm{i} \propto R^\mathrm{ri}$ and $M_\mathrm{i} \propto V^\mathrm{vi}$.

Consider now the case in which the gas is embedded in a dark halo, and assume that when it forms stars the mass in gas is proportional to the dark matter inside R_i, $M_\mathrm{i} \propto M$. If the halo is dominant, the gas loss would have no dynamical effect on the stellar system that is left behind, so $R \simeq R_\mathrm{i}$ and $V \simeq V_\mathrm{i}$. The relations for structure and velocity that hold before the removal give

$$L \propto R^r \propto V^v,\ M \propto R^\mathrm{ri},\ V^2 \propto M/R, \tag{100}$$

so that we obtain

$$v = 2r/(\mathrm{ri} - 1). \tag{101}$$

In the limit $M_\star \ll M_\mathrm{g}$ and $L \propto M_\star$, from the metallicity relation we have

$$Z \propto \frac{M_\star}{M_\mathrm{g}} \propto \frac{L}{M} \propto L^{1/z}, \tag{102}$$

so that

$$r/r\mathrm{i} = (z-1)/z. \tag{103}$$

The final equation comes from the energy condition. For thermal energy, in the limit of substantial gas loss, $M_\mathrm{g} \simeq M_\mathrm{i}$, we have $L \propto MV^2$, and so

$$v = 2z. \tag{104}$$

For a CDM spectrum $P(k) = Ck^{n_\mathrm{s}}$, we obtain $r_\mathrm{i} = 6/(5+n_\mathrm{s})$ and introducing $z = 2r/(r-1)$ and $n_\mathrm{s} = 12/(r+1) - 5$. For $r = 4$ we find $n_\mathrm{s} = -2.6$ consistent with the CDM and the scaling relations are

$$L \propto R^4 \propto Z^{2.7} \propto V^{5.3}. \tag{105}$$

Let us now investigate the critical conditions, in terms of gas density n and virial velocity V, for a global supernova-driven gas removal from a galaxy while it is forming stars. Here, spherical symmetry and the presence of a central point source are assumed. The basic requirement for gas removal is that the energy that has been pumped into the gas is enough to expel it from the protogalaxy. The energy input in turn depends on the supernova rate, on the efficiency of energy transfer into the gas, and on the time it takes for the SNRs to overlap and hence affect a substantial fraction of the gas. The first is determined by the rate of star formation, the second by the evolution of the individual SNRs, and the third by both. When all these are expressed as a function of n and V, the critical condition for removal takes the form

$$E(n,V) \geq \frac{1}{2} M_\mathrm{g} V^2. \tag{106}$$

This relation defines a locus in the $n - V$ diagram shown in Fig. 12 within which substantial gas loss is possible. In Fig. 12, the cooling curve, above which the cooling time is less than the free-fall time, confines the region where the gas can contract and form stars. The almost vertical line V_crit divides the permissible region for galaxy formation in two; a protogalaxy with $V > V_\mathrm{crit}$ would not expel a large fraction of its original gas but rather turn most of its original gas into stars to form a 'normal' galaxy. A protogalaxy with $V < V_\mathrm{crit}$ can produce a supernova-driven wind out of the first burst of star formation, which would drive a substantial fraction of the protogalactic gas out, leaving a diffuse dwarf.

The short-dashed curve marked "1σ" corresponds to density perturbations $\delta M/M$ at their equilibrium configuration after a dissipationless collapse from a CDM spectrum, normalized to $\delta M/M$ at a comoving radius $8\,h^{-1}$ Mpc. The

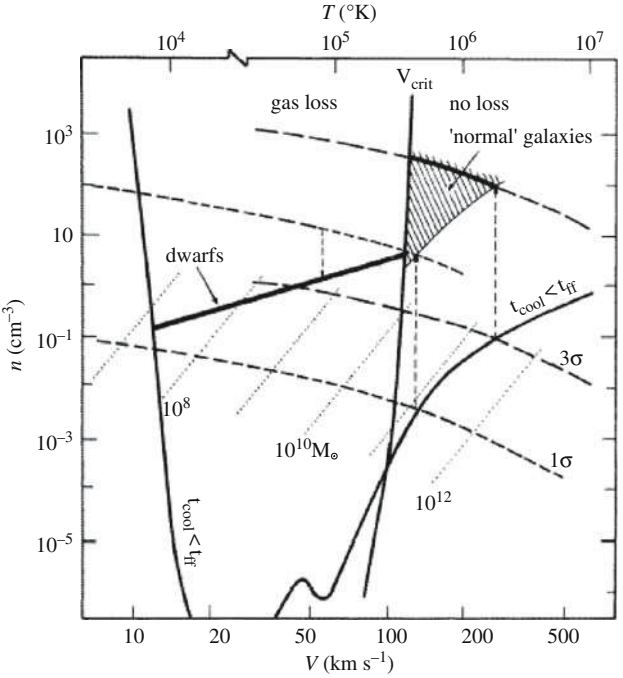

Fig. 12. Gas number density vs. virial velocity (or viral temperature), the formation of dwarf vs. "normal" galaxies in CDM halos, and the origin of biased galaxy formation. From Dekel & Silk (1986) [18]

density n is calculated for a uniformly distributed gas in the CDM halos, with a gas-to-total mass ratio $\chi = 0.1$. The corresponding parallel short-dashed curve corresponds to the protogalactic gas clouds, after a contraction by a factor $\chi^{-1} = 10$ inside isothermal halos, to densities such that star formation is possible. The vertical dashed arrow marks the largest galaxy that can form out of a typical 1σ peak in the initial distribution of density fluctuations. The vast majority of such protogalaxies, when they form stars, have $V < V_{\rm crit}$, so they would turn into dwarfs. The locus where "normal" galaxies are expected to be found is the shaded area. It is evident that most of them must originate form 2σ and 3σ peaks in the CDM perturbations.

So, the theory hence predicts two distinct types of galaxies which occupy two distinct loci in the $n - V$ diagram: the "normal" galaxies are confined to the region of larger virial velocities and higher densities, and they tend to be massive; while the diffuse dwarfs are typically of smaller velocities and lower densities, and their mass in star is less than 5×10^9 M$_\odot$.

This simple model, in spite of its enlightening power, has clear limitations. The most severe is that is assumes a spherical geometry and a single supernova explosion. These assumptions have been released by a subsequent study

Table 4. Mass ejection efficiency

Visible Mass	Mechanical Luminosity $[10^{38}\,\mathrm{erg\,s^{-1}}]$		
$[\mathrm{M}_\odot]$	0.1	1.0	10
10^6	0.18	1.0	1.0
10^7	3.5×10^{-3}	8.4×10^{-3}	4.8×10^{-2}
10^8	1.1×10^{-4}	3.4×10^{-4}	1.3×10^{-3}
10^9	0.0	7.6×10^{-6}	1.9×10^{-5}

based on a large set of numerical simulations ([41]). These authors modeled the effects of repeated supernova explosions from starbursts in dwarf galaxies on the interstellar medium of these galaxies, taking into account the gravitational potential of their dominant dark matter halos. They explored supernova rates from one every 30,000 yr to one every 3 Myr, equivalent to steady mechanical luminosities of $L = 0.1 - 10 \times 10^{38}\,\mathrm{ergs\,s^{-1}}$, occurring in dwarf galaxies with gas masses $M_\mathrm{g} = 10^6 - 10^9 \mathrm{M}_\odot$. Surprisingly, [41] found that the mass ejection efficiency is very low for galaxies with mass $M_\mathrm{g} \geq 10^7 \mathrm{M}_\odot$. Only galaxies with $M_\mathrm{g} \lesssim 10^6 \mathrm{M}_\odot$ have their interstellar gas blown away, and then virtually independently of L (see Table 4). On the other hand, metals from the supernova ejecta are accelerated to velocities larger than the escape speed from the galaxy far more easily than the gas. They found[1] that for $L_{38} = 1$, only about 30% of the metals are retained by a 10^9 $\mathrm{M}_\odot$ galaxy, and virtually none by smaller galaxies (see Table 5).

6.4 Blowout, Blowaway and Galactic Fountains

The results of the MacLow & Ferrara study served to clearly classify the various events induced by starburst in galaxies (Fig. 13). We can distinguish

Table 5. Metal ejection efficiency

Visible Mass	Mechanical Luminosity $[10^{38}\,\mathrm{erg\,s^{-1}}]$		
$[\mathrm{M}_\odot]$	0.1	1.0	10
10^6	1.0	1.0	1.0
10^7	1.0	1.0	1.0
10^8	0.8	1.0	1.0
10^9	0.0	0.69	0.97

[1] Throughout the text we use the standard notation $Y_X = Y/10^X$.

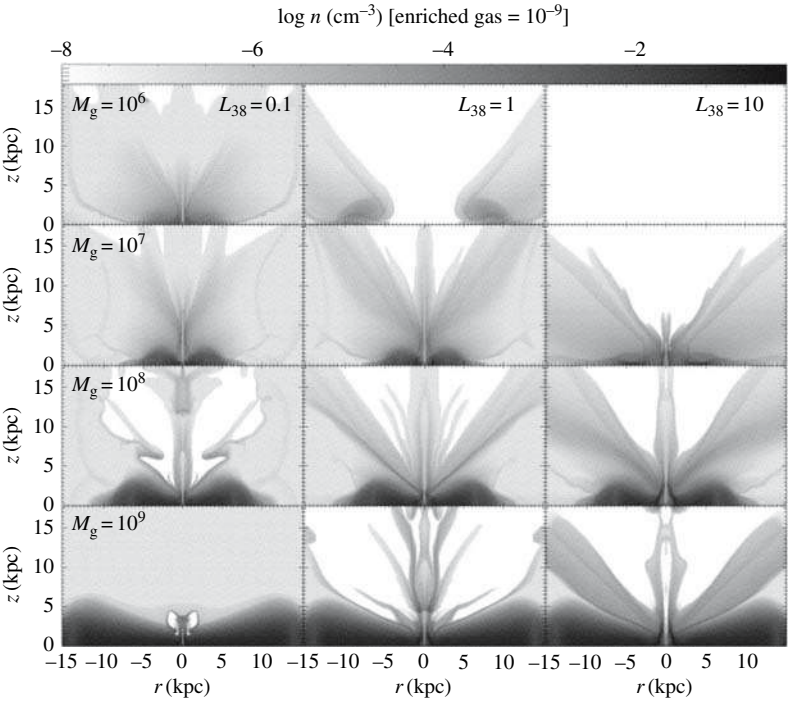

Fig. 13. Distribution of metal-enriched stellar outflow material and SN ejecta from the starburst energy source at time 200 Myr. Note that in most cases, no enriched gas remains in the disks of the galaxies. From Mac Low & Fenara (1999) [41]

between blowout and blowaway processes, depending on the fraction of the parent galaxy mass involved in the mass loss phenomena. Whereas in the blowout process the fraction of mass involved is the one contained in cavities created by the supernova or superbubble explosions, the blowaway is much more destructive, resulting in the complete expulsion of the gas content of the galaxy. The two processes lie in different regions of the $(1 + \phi) - M_{\rm g,7}$ plane shown in Fig. 14, where $\phi = M_{\rm h}/M_{\rm g}$ is the dark-to-visible mass ratio and $M_{\rm g,7} = M_{\rm g}/10^7\ {\rm M}_\odot$ ([23]). Galaxies with gas mass content larger than $10^9\ {\rm M}_\odot$ do not suffer mass losses, due to their large gravitational well. Of course, this does not rule out the possible presence of outflows with velocities below the escape velocity (fountain) in which material is temporarily stored in the halo and then returns to the main body of the galaxy. For galaxies with gas mass lower than this value, outflows cannot be prevented. If the mass is reduced further, and for $\phi \lesssim 20$, a blowavay, and therefore a complete stripping of the galactic gas, should occur. To exemplify, the expected value of ϕ as function of $M_{\rm g}$ empirically derived by [52] is also plotted, which should give an idea of a likely location of the various galaxies in the $(1 + \phi) - M_{\rm g,7}$ plane.

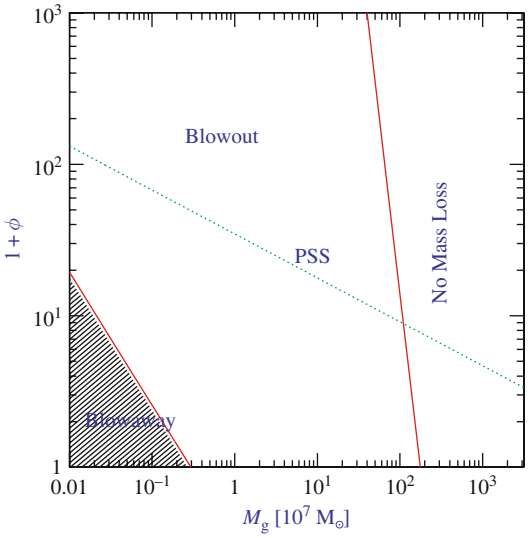

Fig. 14. Regions in the $(1+\phi)-M_{g,7}$ plane in which different dynamical phenomena may occur. Also shown is the locus point describing the Persic et al. (PSS) relation (*dotted line*). From Ferrara & Tolstoy (2000) [23]

Blowout

The evolution of a point explosion in an exponentially stratified medium can be obtained from dimensional analysis. Suppose that the gas density distribution is horizontally homogeneous and that $\rho(z) = \rho_0 \exp(-z/H)$. The velocity of the shock wave is $v \sim (P/\rho)^{1/2}$, where the pressure P is roughly equal to E/z^3, and E is the total energy if the explosion. Then it follows that

$$v(z) \simeq E^{1/2}\rho_0^{-1/2} \exp(z/2H) z^{-3/2}. \tag{107}$$

This curve has a minimum at $z = 3H$ and this value defines the height at which the shock wave, initially decelerating, is accelerated to infinity and a blowout takes place. Therefore, $3H$ can be used as the fiducial height where the velocity $v_\mathrm{b} = v(3H)$ is evaluated.

There are three different possible fates for SN-shocked gas, depending on the value of v_b. If $v_\mathrm{b} < c_\mathrm{s,eff}$, where c_s is the sound speed in the ISM, then the explosion will be confined in the disk and no mass-loss will occur; for $c_\mathrm{s,eff} < v_\mathrm{b} < v_\mathrm{e}$ (v_e is the escape velocity) the supershell will breakout of the disk into the halo, but the flow will remain bound to the galaxy; finally, $v_\mathrm{e} < v_\mathrm{b}$ will lead to a true mass-loss from the galaxy.

Blowaway

The requirement for blowaway is that the momentum of the shell is larger than the momentum necessary to accelerate the gas outside the shell at velocity

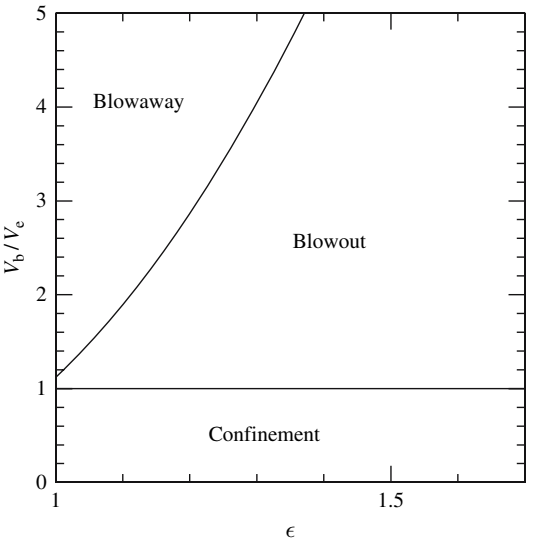

Fig. 15. Conditions for blowaway, blowout and confinement as a function of the major-to-minor axis ratio ϵ of dwarf galaxies; $\epsilon = 1$ corresponds to spherical bodies. From Ferrara & Tolstoy (2000) [23]

larger than the escape velocity, $M_s v_c \leq M_0 v_e$. Defining the disk axis ratio as $\epsilon = R/H \geq 1$, the blowaway condition can be rewritten as

$$\frac{v_b}{v_e} \geq (\epsilon - a)^2 a^{-2} e^{3/2}, \tag{108}$$

where $a = 2/3$. The above equation is graphically displayed in Fig. 15. Flatter galaxies (large ϵ values) preferentially undergo blowout, whereas rounder ones are more likely to be blown-away; as v_b/v_e is increased the critical value of ϵ increases accordingly. Unless the galaxy is perfectly spherical, blowaway is always preceded by blowout; between the two events the aspect of the galaxy may look extremely perturbed, with one or more huge cavities left after blowout.

6.5 Further Model Improvements

All models discussed so far make the so-called SEX bomb assumption, i.e. a Spherically EXpanding blastwave. This is a good assumption as long as a single burst regions exists whose size is small compared to the size of the system and it is located at the center of the galaxy. This however is a rather idealized situation.

A more detailed study about the possibility of mass loss due to distributed SN explosions at high redshift has been carried on by [46]. These authors presented results from three–dimensional numerical simulations of the dynamics

of SN–driven bubbles as they propagate through and escape the grasp of subgalactic halos with masses $M = 10^8\,h^{-1}\ \mathrm{M}_\odot$ at redshift $z = 9$. Halos in this mass range are characterized by very short dynamical timescales (and even shorter gas cooling times) and may therefore form stars in a rapid but intense burst before SN "feedback" quenches further star formation. This hydrodynamic simulations use a nested grid method to follow the evolution of explosive multi–SN events operating on the characteristic timescale of a few $\times 10^7$ yr, the lifetime of massive stars. The results confirm that, if the star formation efficiency of subgalactic halos is $\approx 10\%$, a significant fraction of the halo gas will be lifted out of the potential well ("blow–away"), shock the intergalactic medium, and pollute it with metal–enriched material, a scenario recently advocated by [44]. Depending on the stellar distribution, [46] found that less than 30% of the available SN energy gets converted into kinetic energy of the blown away material, the remainder being radiated away (Fig. 16).

However, it appears that realistic models lead to the conclusion that mechanical feedback is less efficient than expected from SEX bomb simple schemes. The reason is that off–nuclear SN explosions drive inward–propagating shocks that tend to collect and pile up cold gas in the central regions of the host halo. Low–mass galaxies at early epochs then may survive multiple SN events and continue forming stars.

Figures 17 and 18 show the fraction of the initial halo baryonic mass contained inside (1, 0.5, 0.1) times the virial radius r_{vir} as a function of time for two different runs: an extended stellar distribution (case 1) and a more concentrated one (case 2). The differences are striking: in case 1, the amount of gas at the center is constantly increasing, finally collecting inside $0.1r_{\mathrm{vir}}$ about 30% of the total initial mass. On the contrary, in case 2, the central regions remain practically devoided of gas until 60 Myr, when the accretion process starts. The final result is a small core containing a fraction of only 5% of the initial mass. In the former case, 50% of the halo mass is ejected together with the shell, whereas in case 2 this fraction is $\sim 85\%$, i.e., the blow-away is nearly complete.

As a first conclusion, it seems that quenching star formation in galaxies by ejecting large fraction of their gas seems very difficult and hence unviable as a feedback scheme (although this might be possible in very small galaxies). Heavy elements can instead escape much more easily as they are carried away by the mass-unloaded, hot, SN-shocked gas; energy is carried away efficiently as well. This conclusion raises the issue if star formation must be rather governed by more gentle and self-regulating processes, i.e. a "true" feedback.

6.6 Additional Implications

The mechanical feedback discussed so far might have important additional consequences which can profoundly affect the subsequent process of structure formation. The most prominent ones are discussed in the following.

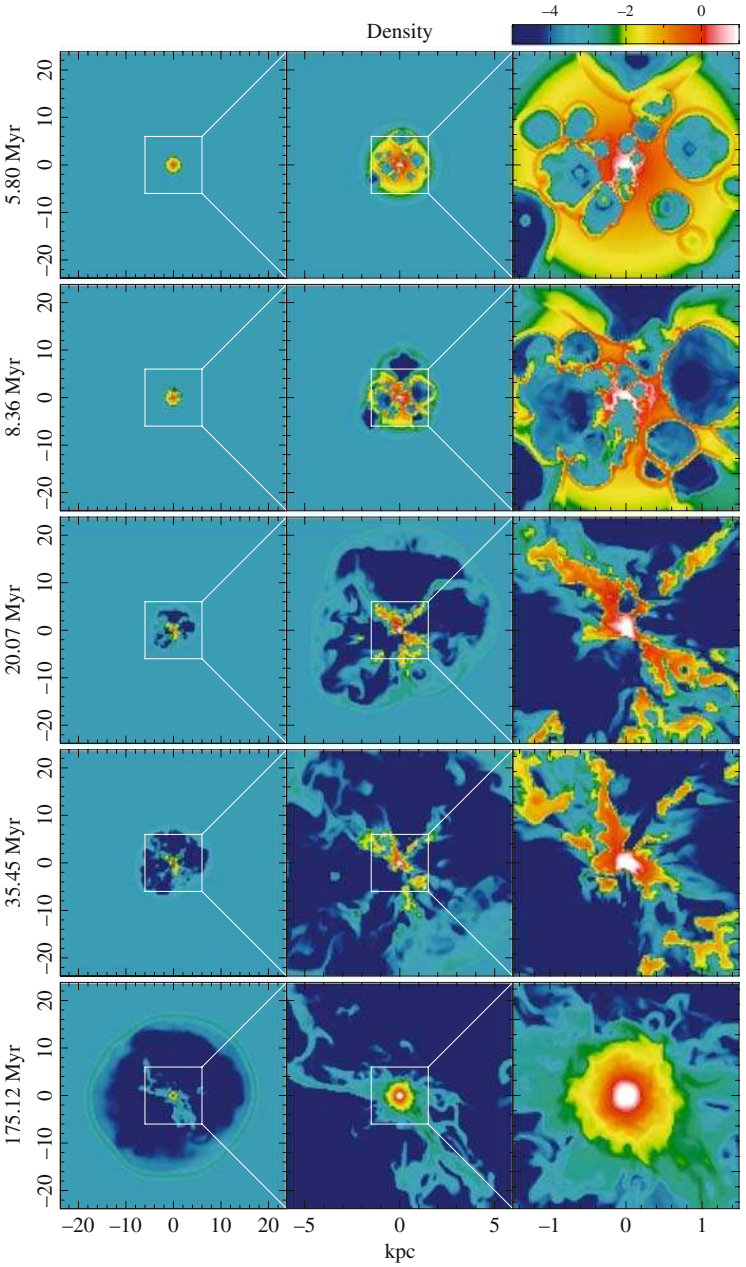

Fig. 16. Snapshots of the logarithmic number density of the gas at five different elapsed times for case 1. The three panels in each row show the spatial density distribution in the x-y plane on the nested grids. The density range is $-5 \leqslant \log(n/\mathrm{cm}^{-3}) \leqslant 1$. From Mori et al. (2002) [46]

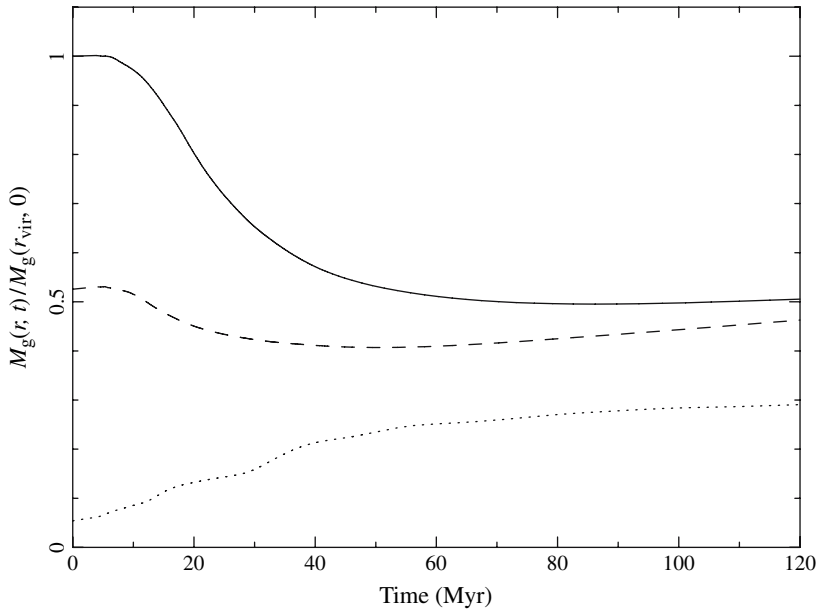

Fig. 17. The evolution of the gas mass inside the gravitational potential well of the CDM halo for an extended stellar distribution (case 1) as a function of time. Curves correspond to the gas mass inside the virial radius r_{vir} (*solid line*), $0.5r_{vir}$ (*dashed line*), and $0.1r_{vir}$ (*dotted line*). From Mori et al. 2002 [46]

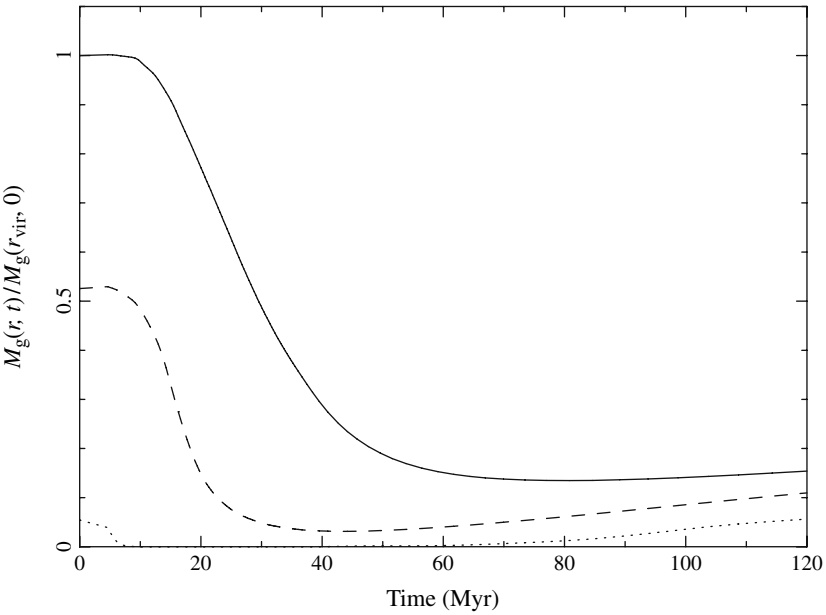

Fig. 18. The same of Fig. 17 but for a more concentrated stellar distribution (case 2). From Mori et al. (2002) [46]

IGM Metal Enrichment and Heating

One of the most obvious signatures of mechanical feedback is the metal enrichment of the intergalactic medium. If powerful winds are driven by supernova explosions, one would expect to see widespread traces of heavy elements away from their production sites, i.e. galaxies. Reference [61] have reported the detection of O VI in the low-density IGM at high redshift. They perform a pixel-by-pixel search for O VI absorption in eight high quality quasar spectra spanning the redshift range $z = 2.0 - 4.5$. At $2 \lesssim z \lesssim 3$ they detect O VI in the form of a positive correlation between the HI Lyα optical depth and the optical depth in the corresponding O VI pixel, down to $\tau_{\rm HI} \sim 0.1$ (underdense regions), that means that metals are far from galaxies (Fig. 19). Moreover, the observed narrow widths of metal absorption lines (CIV, SiIV) lines lines imply low temperatures $T_{\rm IGM} \sim$ few $\times\, 10^4$ K.

A natural hypothesis would be that the Lyα forest has been enriched by metals ejected by Lyman Break Galaxies at moderate redshift. The density of these objects is $n_{\rm LBG} = 0.013\, h^3\, {\rm Mpc}^3$. A filling factor of $\sim 1\%$ is obtained for a shock radius $R_{\rm s} = 140\, h^{-1}\, {\rm kpc}$, that corresponds at $z = 3$ ($h = 0.5$)

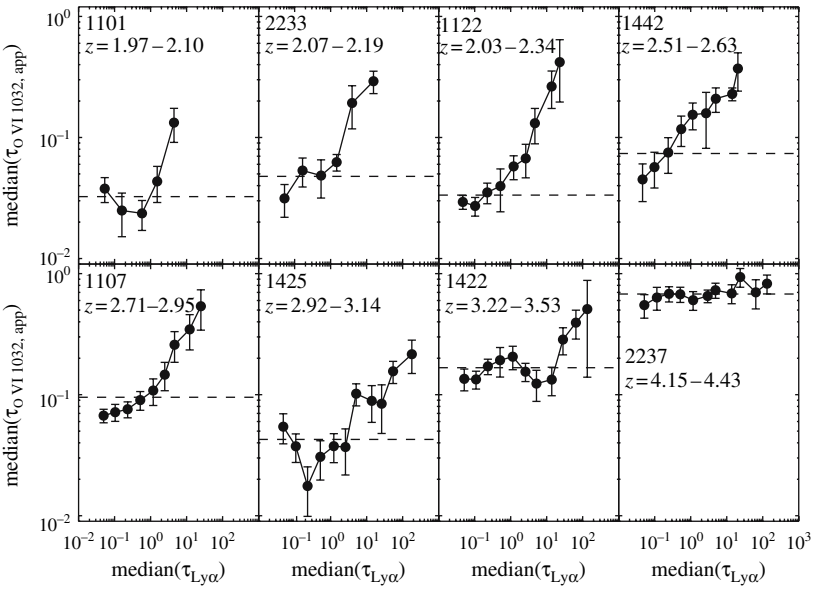

Fig. 19. IGM Lyα-O VI optical depth correlation from the pixel analysis of the spectra of 8 QSOs. Vertical error bars are 1σ errors, horizontal error bars are smaller than the symbols and are not shown. For $z \lesssim 3$ $\tau_{\rm O\ VI\ ,app}$ and $\tau_{\rm HI}$ are clearly correlated, down to optical depths as low as $\tau_{\rm HI} \sim 10^{-1}$. A correlation between $\tau_{\rm O\ VI\ ,app}$ and $\tau_{\rm HI}$ implies that O VI absorption has been detected in the Lyα forest. From Schaye et al. (2002) [61]

to a shock velocity $v_s = 600\,\mathrm{km\,s^{-1}}$. In this case, we expect a post shock gas temperature larger than $2 \times 10^6\,\mathrm{K}$, that is around hundred times what observed. So the metal pollution must have occurred earlier than redshift 3, resulting in a more uniform distribution and thus enriching vast regions of the intergalactic space. This allows the Lyα forest to be hydrodynamically "cold" at low redshift, as intergalactic baryons have enough time to relax again under the influence of dark matter gravity only ([57]).

In Fig. 20 shows the thermal history of the IGM as a function of redshift as computed by [44]. The gas is allowed to interact with the CMB through Compton cooling and either with a time-dependent quasar-ionizing background as computed by [30] or with a time-dependent metagalactic flux of intensity $10^{-22}\,\mathrm{erg\,cm^{-2}\,s^{-1}\,Hz^{-1}\,sr^{-1}}$ at 1 Ryd and power–law spectrum with energy slope $\alpha = 1$. The temperature of the medium at $z = 9$ has been either computed self-consistently from photoheating or fixed to be in the range $10^{4.6} - 10^5\,\mathrm{K}$, as expected in SN-driven bubbles with significant filling factors.

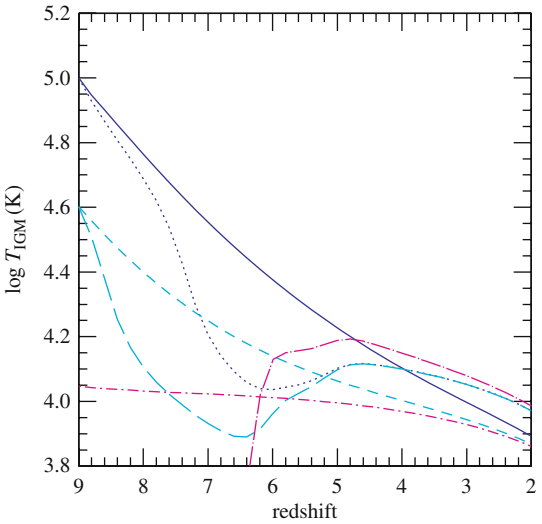

Fig. 20. Thermal history of intergalactic gas at the mean density. *Short dash–dotted line:* temperature evolution when the only heating source is a constant UV background of intensity $10^{-22}\,\mathrm{erg\,cm^{-2}\,s^{-1}\,Hz^{-1}\,sr^{-1}}$ at 1 Ryd and power–law spectrum with energy slope $\alpha = 1$. *Long dash–dotted line:* same for the time–dependent quasar ionizing background as computed by Haardt & Madau (1996) [30] (HM). *Short dashed line:* heating due to a constant UV background but with an initial temperature of $4 \times 10^4\,\mathrm{K}$ at $z = 9$ as expected from an early era of pregalactic outflows. *Long dashed line:* same but for a HM background. *Solid line:* heating due to a constant UV background but with an initial temperature of $10^5\,\mathrm{K}$ at $z = 9$. *Dotted line:* same but for a HM background. From Madau et al. (2001) [44]

The various curves show that the temperature of the IGM at $z = 3 - 4$ will retain little memory of an early era of pregalactic outflows.

The large increase of the IGM temperature at high redshift connected with such an era of pregalactic outflows, causes a larger Jeans mass, thereby preventing gas from accreting efficiently into small dark matter halos ([7]). For typical preheating energies, the IGM is driven to temperatures just below the virial temperature of halos hosting $L_\star$ galaxies. Thus we may expect preheating to have a strong effect on the galaxy luminosity function at $z = 0$ (Fig. 21). Moderate preheating scenarios, with $T_{\rm IGM} \geq 10^5\,{\rm K}$ at $z \sim 10$, are able to flatten the faint-end slope of the luminosity function, producing excellent agreement with observations, without the need for any local strong feedback within galaxies.

Gas Stripping from Neighbor Galaxy Shocks

The formation of a galaxy can be inhibited also by the outflows from neighboring dwarfs as the result of two different mechanisms. In the "mechanical evaporation" scenario, the gas associated with an overdense region is heated by a shock above its virial temperature. The thermal pressure of the gas then overcomes the dark matter potential and the gas expands out of the halo,

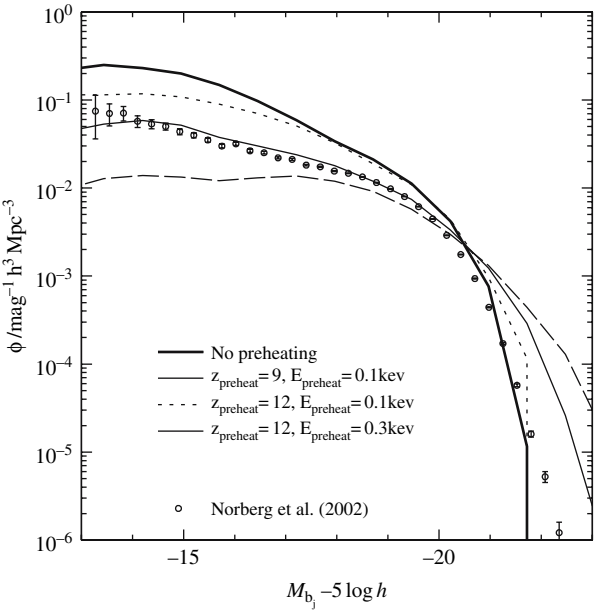

Fig. 21. B-band luminosity functions of galaxies at $z = 0$, as predicted by the semi-analytic model of Benson et al. (2002) [6]. The observational data are shown as circles

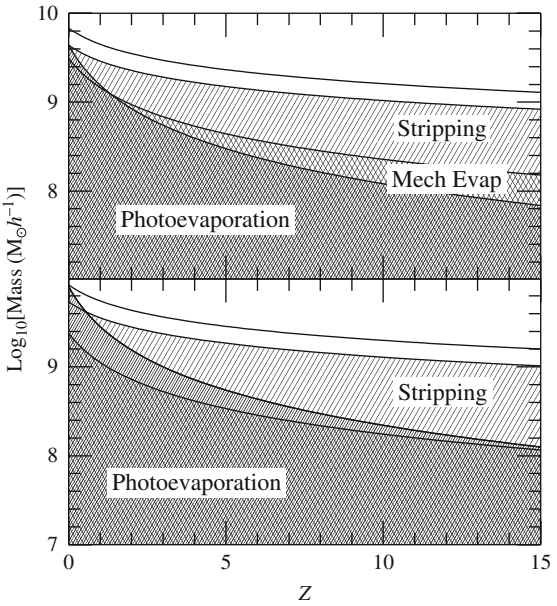

Fig. 22. Relevant mass scales for suppression of dwarf galaxy formation. The upper lines are the masses below which halos will be heated beyond their virial temperatures, although cooling prevents mechanical evaporation from occurring for halos with masses above the cross hatched regions. The second highest set of lines, bounding the lightly shaded regions, show the masses below which baryonic stripping is effective. Finally the heavily shaded regions show objects that are susceptible to photoevaporation. The upper panel is a flat CDM model and the lower panel is a flat ΛCDM model with $\Omega_0 = 0.3$. In all cases $(\epsilon N h) = 5000\ \Omega_0^{-1}$, the overdensity $\delta = \rho/\rho_0 = 2.0$, $\Omega_{\rm b} = 0.05$, $h = 0.65$. Note that photoevaporation affects a larger mass range than mechanical evaporation in the ΛCDM cosmology. From Scannapieco et al. (2002) [57]

preventing galaxy formation. In this case, the cooling time of the collapsing cloud must be shorter than its sound crossing time, otherwise the gas will cool before it expands out of the gravitational well and will continue to collapse.

Alternatively, the gas may be stripped from a collapsing perturbation by a shock from a nearby source. In this case, the momentum of the shock is sufficient to carry with it the gas associated with the neighbor, thus devoiding the halo of its baryons and preventing a galaxy from forming.

In principle, outflows can suppress the formation of nearby galaxies both by shock heating/evaporation and by stripping of the baryonic matter from collapsing dark matter halos; in practice, the short cooling timesfor most dwarf-scale collapsing objects suggest that the baryonic stripping scenario is almost always dominant. This mechanism has the largest impact in forming dwarfs in the $\lesssim 10^9$ M$_\odot$ range which is sufficiently large to resist photoevap-

oration by UV radiation, but too small to avoid being swept up by nearby dwarf outflows.

7 Additional Feedback Processes

Once the first sources have formed, their mass deposition, energy injection and emitted radiation can deeply affect the subsequent galaxy formation process and influence the evolution of the IGM via additional "feedback" effects. The occurrence of the first supernovae will evacuate/heat the gas thus suppressing the subsequent SF process. This feedback is then acting back on the energy source (star formation); it is of a negative type, and it could drive the SF activity in such a way that only a sustainable amount of stars is formed. However, feedback can fail to produce such regulation either in small galaxies, where the gas can be ejected by a handful of SNe, or in cases when the star formation timescale is too short compared to the feedback timescale. In the spirit of the present review we will only discuss feedback occurring at high redshifts and hence shaping the first structures.

Although a rigorous classification of the various effects is not feasible, they can be divided into three broad classes: *radiative*, *mechanical* and *chemical* feedback. In the first class fall all those effects associated, in particular, with ionization/dissociation of hydrogen atoms/molecules; the second class (discussed at length in the previous section) is produced by the mechanical energy injection of massive stars in form of winds or SN explosions; and chemical feedback is instead related to the postulated existence of a critical metallicity governing the cosmic transition from very massive stars to "normal" stars.

Attempting a classification of all proposed feedback effects is almost desperate, due to the large number of applications and definitions, often discordant, present in the literature. Nevertheless, we offer in Table 6 a working classification aimed essentially at organizing the material presented in these Lectures. Whereas the radiative feedback is intimately connected with cosmic reionization, chemical feedback is instead strongly dependent on the history of metal enrichment of the universe.

7.1 Radiative Feedback

Radiative feedback is related to the ionizing/dissociating radiation produced by massive stars or quasars. This radiation can have local effects (i.e. on the same galaxy that produces it) or long-range effects, either affecting the formation and evolution of nearby objects or joining the radiation produced by other galaxies to form a background. In spite of the different scenarios implied, the physical processes are very similar.

Photoionization/evaporation

The collapse and formation of primordial objects exposed to a UV radiation field can be inhibited or halted for two main reasons: (i) cooling is

considerably suppressed by the decreased fraction of neutral hydrogen, and (ii) gas can be photoevaporated out of the host halo. In fact, the gas incorporated into small mass objects that were unable to cool efficiently, can be boiled out of the gravitational potential well of the host halo if it is heated by UV radiation above the virial temperature. Such effects are produced by the same radiation field and act simultaneously. For this reason, it is hard to separate their individual impact on the final outcome. The problem has been extensively studied by several authors (e.g. for details see [15] and references therein) In particular, [35], assuming spherical symmetry, solve self-consistently radiative transfer of photons, non-equilibrium H_2 chemistry and gas hydrodynamics of a collapsing halo. They find that at weak UV intensities ($J < 10^{-23}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,sr^{-1}\,Hz^{-1}}$), objects as small as $v_c \sim 15\,\mathrm{km\,s^{-1}}$ are able to collapse, owing to both self-shielding of the gas and H_2 cooling. At stronger intensities though, objects as large as $v_c \sim 40\,\mathrm{km\,s^{-1}}$ can be photoevaporated and prohibited from collapsing, in agreement with previous investigations based on the optically thin approximation. More refined three-dimensional hydrodynamic simulations confirm the result that the presence of UV radiation delays or suppresses the formation of low mass objects. In contrast with these studies, [19], applying a 1D code to $z > 10$ objects, find that objects as small as $v_c \sim 10\,\mathrm{km\,s^{-1}}$ can self-shield and collapse because the collisional cooling processes at high redshift are more efficient and the amplitude of the ionizing background is lower. The second condition might not always apply, as the ionizing flux at high redshift is dominated by the direct radiation from neighboring halos rather than the background ([14]).

The photoevaporation effect might be particularly important for Pop III objects, as their virial temperatures are below the typical temperatures achieved by a primordial photoionized gas ($T \approx 10^4\,\mathrm{K}$). For such an object, the ionization front gradually burns its way through the collapsed gas,

Table 6. Classification of different feedback effects

Radiative	NEGATIVE *Mechanical*	*Chemical*
1. Photoionization/evaporation	1. Blowout/blowaway	1. Fragmentation
2. H_2 Photodissociation	2. Impinging shocks	
3. Photoheating filtering	3. Preheating	
Radiative	POSITIVE *Mechanical*	*Chemical*
1. In front of H II regions	1. Behind shocks	
2. Inside relic H II regions	2. Shell fragmentation	
3. X-ray background		

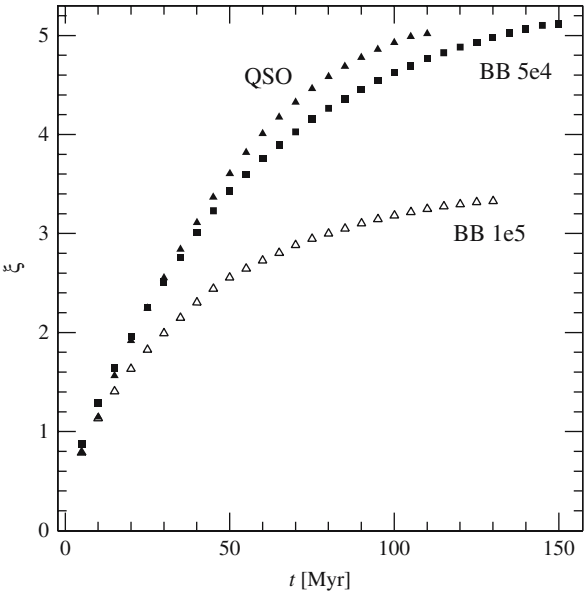

Fig. 23. Evolution of the cumulative number of ionizing photons absorbed per initial minihalo atom for a QSO and two black-body spectra with $T = 5 \times 10^4$ K and $T = 10^5$ K, as labeled (see Shapiro et al. 2004 [64] for detailed definitions)

producing a wind that blows backwards into the IGM and that eventually evaporates all the gas content. According to recent studies, these sub-kpc galactic units were so common to dominate the absorption of ionizing photons. This means that estimates of the number of ionizing photons per H atom required to complete reionization should not neglect their contribution to absorption. As Fig. 23 ([64]) shows, the number of ionizing photons absorbed per initial minihalo atom, ξ, increases gradually with time; in addition, it depends on the ionizing spectrum assumed. For the hard QSO spectrum, the ionization front is thicker and penetrates deeper into the denser and colder parts of the halo, increasing the rate of recombinations per atom, compared to stellar type sources. However, this same pre-heating effect shortens the evaporation time, ultimately leading to a rough cancellation of the two effects and the same total ξ as for a black-body spectrum with $T = 5 \times 10^4$ K (mimicking a low metallicity Pop II stellar emission). An even lower ξ is needed for a black-body spectrum with $T = 10^5$ K (more typical of a Pop III stellar emission), because of an increased evaporation vs. penetration ability. Thus, overall, Pop III stellar sources appear significantly more efficient than Pop II or QSO sources in terms of the total number of ionizing photons needed to complete the photoevaporation process.

H_2 *photodissociation*

As intergalactic H_2 is easily photodissociated, a soft-UV background in the Lyman-Werner bands could quickly build up and have a negative feedback on the gas cooling and star formation inside small halos (see Chap. 2). In addition to an external background, the evolution of structures can also be affected by internal dissociating radiation. In fact, once the first generation of stars has formed in an object, it can affect the subsequent star formation process by photodissociating molecular hydrogen in star forming clouds; for example, if the molecular cloud has a metallicity smaller than about $10^{-2.5}$ $Z_\odot$, a single O star can seriously deplete the H_2 content so that subsequent star formation is almost quenched. Thus, it seems plausible that stars do not form efficiently before the metallicity becomes larger than about 10^{-2} solar.

Reference [42] find that the fraction of gas available for star formation in Pop III objects of mass M exposed to a flux with intensity $J_{\rm LW}$ in the Lyman-Werner band is $\sim 0.06 \ln(M/M_{\rm th})$, where the mass threshold, $M_{\rm th}$, is given by:

$$(M_{\rm th}/{\rm M}_\odot) = 1.25 \times 10^5 + 8.7 \times 10^5 \left(\frac{J_{\rm LW}}{10^{-21}\,{\rm erg}^{-1}{\rm cm}^{-2}{\rm Hz}^{-1}} \right). \qquad (109)$$

The same problem has been analyzed by [65] by means of three-dimensional SPH calculations, where radiative transfer is solved by a direct method and the non-equilibrium chemistry of primordial gas is included. They find that star formation is suppressed appreciably by UVB, but baryons at high-density peaks are self-shielded, eventually forming some amount of stars.

The negative feedback described above could be counterbalanced by the *positive feedback* of H_2 re-formation, e.g. in front of H II regions, inside relic H II regions, once star formation is suppressed in a halo and ionized gas starts to recombine ([54]), in cooling gas behind shocks produced during the ejection of gas from these objects ([21]). Thus, a second burst of star formation might take place also in the small objects where it has been suppressed by H_2 dissociation. H_2 production could also be promoted by an X-ray background, which would increase the fractional ionization of protogalactic gas. Such a positive feedback, though, is not able to balance UV photodissociation in protogalaxies with $T_{\rm vir} < 2000\,{\rm K}$ Similar arguments apply to high energy cosmic rays.

Photoheating filtering

Cosmic reionization might have a strong impact on subsequent galaxy formation, particularly affecting low-mass objects. In fact, the heating associated with photoionization causes an increase in the temperature of the IGM gas which will suppress the formation of galaxies with masses below the Jeans mass. As [29] has pointed out, one can expect that the effect of reionization depends on the reionization history, and thus is not universal at a given redshift. More precisely, one should introduce a "filtering" scale, $k_{\rm F}$, (or, equivalently, filtering mass $M_{\rm F}$) over which the baryonic perturbations

are smoothed as compared to the dark matter, yielding the approximate relation $\delta_{\rm b} = \delta_{\rm dm} {\rm e}^{-k^2/k_{\rm F}^2}$. The filtering mass as a function of time is related to the Jeans mass by:

$$M_{\rm F}^{2/3} = \frac{3}{a} \int_0^a {\rm d}a' M_{\rm J}^{2/3}(a') \left[1 - \left(\frac{a'}{a} \right)^{1/2} \right]. \tag{110}$$

Note that at a given moment in time the two scales can be very different. Also, in contrast to the Jeans mass, the filtering mass depends on the full thermal history of the gas instead of the instantaneous value of the sound speed, so it accounts for the finite time required for pressure to influence the gas distribution in the expanding universe. The filtering mass increases from roughly 10^7 $\rm M_\odot$ at $z \approx 10$ to about 10^9 $\rm M_\odot$ at redshift $z \approx 6$, thus efficiently suppressing the formation of objects below that mass threshold. Of course such result is somewhat dependent on the assumed reionization history.

An analogous effect is found inside individual H II regions around the first luminous sources. Once an ionizing source turns off, its surrounding H II region Compton cools and recombines. Nonetheless, the "fossil" H II regions left behind remain at high adiabats, prohibiting gas accretion and cooling in subsequent generations of Pop III objects.

7.2 Chemical Feedback

The concept of chemical feedback is relatively recent, having been first explored by [62] and subsequently discussed by [63] and [43].

According to the scenario outlined in Sect. 2, the first stars forming out of gas of primordial composition might be very massive, with masses $\approx 10^2 - 10^3$ $\rm M_\odot$. The ashes of these first supernova explosions pollute with metals the gas out of which subsequent generations of low-mass Pop II/I stars form, driving a transition from a top-heavy IMF to a "Salpeter-like" IMF when locally the metallicity approaches the critical value $Z_{\rm cr} = 10^{-5\pm1}$ $\rm Z_\odot$ ([62, 63]). Thus, the cosmic relevance of Pop III stars and the transition to a Pop II/I star formation epoch depends on the efficiency of metal enrichment from the first stellar explosions, the so-called chemical feedback, which is strictly linked to the number of Pop III stars that explode as PISN, the metal ejection efficiency, transport and mixing in the IGM. It is very likely that the transition occurred rather smoothly because the cosmic metal distribution is observed to be highly inhomogeneous: even at moderate redshifts, $z \approx 3$, the clustering properties of C IV and Si IV QSO absorption systems are consistent with a metal filling factor $< 10\%$, showing that metal enrichment is incomplete and inhomogeneous.

As a consequence, the use of the critical metallicity as a global criterion is somewhat misleading because chemical feedback is a *local process*, with regions close to star formation sites rapidly becoming metal-polluted and overshooting

$Z_{\rm cr}$, and others remaining essentially metal-free. Thus, Pop III and Pop II star formation modes could have been coeval, and detectable signatures from Pop III stars could be found well after the volume-averaged metallicity has become larger than critical.

Reference [58] have studied, using an analytical model of inhomogeneous structure formation, the separate evolution of Pop III/Pop II stars as a function of star formation and wind efficiencies. They parametrized the chemical feedback through a single quantity, $E_{\rm g}$, which represents the kinetic energy input from SNe per unit gas mass into stars. This quantity governs the transport of metals in regions away from their production site and therefore the metallicity distribution; it is related to the number of exploding Pop III stars and encodes the dependence on the assumed IMF. For all values of the feedback parameter, $E_{\rm g}$, [58] found that, while the peak of Pop III star formation occurs at $z \approx 10$, such stars continue to contribute appreciably to the star formation rate density at much lower redshifts (Fig. 24), even though the mean IGM metallicity has moved well past the critical transition metallicity. This finding has important implications for the development of efficient strategies for the detection of Pop III stars in primeval galaxies, as discussed previously.

7.3 A Final View

Given the large number of processes so far discussed, it is probably worth to summarize them briefly. Figure 25 illustrates all possible evolutionary tracks and final fates of primordial objects, together with the mass scales determined by the various physical processes and feedbacks. We recall that there are four critical mass scales in the problem: (i) $M_{\rm crit}$, the minimum mass for an object to be able to cool in a Hubble time; (ii) $M_{\rm H}$, the critical mass for which hydrogen Lyα line cooling is dominant; (iii) $M_{\rm sh}$, the characteristic mass above which the object is self-shielded, and (iv) $M_{\rm by}$ the characteristic mass for stellar feedback, below which blowaway can not be avoided. Starting from a virialized dark matter halo, condition (i) produces the first branching, and objects failing to satisfy it will not collapse and form only a negligible amount of stars. In the following, we will refer to these objects as *dark objects*. Protogalaxies with masses in the range $M_{\rm crit} < M < M_{\rm H}$ are then subject to the effect of radiative feedback, which could either impede the collapse of those of them with mass $M < M_{\rm sh}$, thus contributing to the class of dark objects, or allow the collapse of the remaining ones ($M > M_{\rm sh}$) to join those with $M > M_{\rm H}$ in the class of *luminous objects*. This is the class of objects that convert a considerable fraction of their baryons in stars. Stellar feedback causes the final bifurcation by inducing a blowaway of the baryons contained in luminous objects with mass $M < M_{\rm by}$; this separates the class in two subclasses, namely "normal" galaxies (although of masses comparable to present day dwarfs) that we dub *gaseous galaxies* and tiny stellar aggregates with negligible traces (if any) of gas to which consequently we will refer to as *naked stellar clusters*.

The role of these distinct populations for the reionization of the universe and their density evolution will be clarified by the following results.

Reference [14] find that the majority of the luminous objects that are able to form at high redshift will experience blowaway, becoming naked stellar clusters, while only a minor fraction, and only at $z \lesssim 15$, when larger objects start to form, will survive and become gaseous galaxies. An always increasing number of luminous objects is forming with decreasing redshift, until $z \approx 15$, where a flattening is seen. This is due to the fact that the dark matter halo mass function is still dominated by small mass objects, but a large fraction of them cannot form due to the following combined effects: (i) towards lower redshift the critical mass for the collapse ($M_{\rm crit}$) increases and fewer objects satisfy the condition $M > M_{\rm crit}$; (ii) the radiative feedback due to either the direct dissociating flux or the soft UVB (SUVB) increases at low redshift as the SUVB intensity reaches values significant for the negative feedback

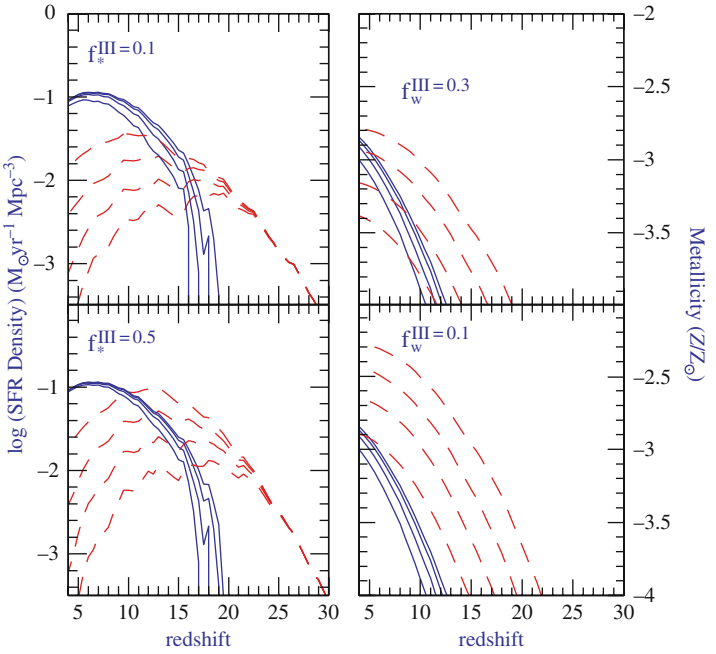

Fig. 24. *Left Panels:* Cosmic star formation rate densities for Pop III (*dashed lines*) and Pop II (solid) stars. A Pop III star formation efficiency $f_\star^{\rm III} = 0.1$ and chemical feedback parameter values, from top to bottom, $\log E_{\rm g} = -4.0, -3.5, -3.0, -2.5$ are assumed (top panel); in the bottom panel $f_\star^{\rm III} = 0.5$, and $\log E_{\rm g} = -3.5, -3.0, -2.5, -2.0$. *Right Panels:* IGM average metallicity from Pop III (*dashed lines*) and Pop II (*solid*) star formation. The wind efficiency is $f_{\rm w} = 0.3$ and $\log E_{\rm g} = -4.0, -3.5, -3.0, -2.5$, from bottom to top (top panel); in the bottom panel, same as above with $f_{\rm w} = 0.1$. From Scannapieco et al. (2003) [58]

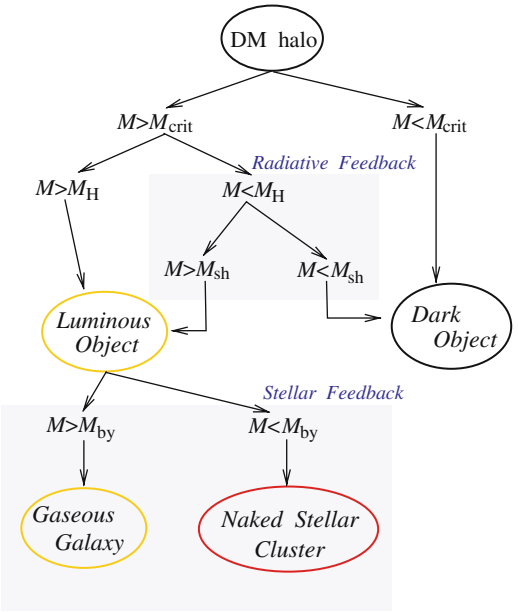

Fig. 25. Schematic evolutionary tracks for primordial objects. From Ciardi et al. (2000) [14]

effect. When the number of luminous objects becomes dominated by objects with $M > M_{\rm H}$, by $z \approx 10$ the population of luminous objects grows again, basically because their formation is now unaffected by negative feedback. A steadily increasing number of objects is prevented from forming stars and remains dark; this population is about $\approx 99\%$ of the total population of dark matter halos at $z \approx 8$. This is also due to the combined effect of points (i) and (ii) mentioned above. This population of halos which have failed to produce stars could be identified with the low mass tail distribution of the dark galaxies that reveal their presence through gravitational lensing of quasars. It has been argued that this population of dark galaxies outnumbers normal galaxies by a substantial amount. At the same time, CDM models predict that a large number of satellites observed should be present around normal galaxies. Many of them would form at redshifts higher than five and would survive merging and tidal stripping inside larger halos to the present time. Their existence is then linked to that of small mass primordial objects and the natural question arises if these objects can be reconciled with the internal properties of halos of present day-galaxies.

A question that naturally arises is which is the fate of the naked stellar objects. As the maximum total mass of a naked stellar object at $z \approx 8$ is $\approx 10^9$ $M_\odot$, given a Salpeter IMF, we find that the upper limit for the number of the above relic stars is $\approx 4 \times 10^6$. As in the Galaxy, given the same IMF,

$\approx 2 \times 10^{11}$ stars with masses in the range $0.1 - 1$ $M_\odot$ are present, one out of $\approx 5 \times 10^4$ stars comes from naked stellar objects.

8 Early Cosmic Dust

In the recent years, dust has been recognized to have an increasingly important role in our understanding of the near and distant Universe. The dramatic effect of dust at low and moderate redshifts has been noticed when a reconstruction of the cosmic star formation history from rest-frame UV/visible emission was first attempted: dust grains absorb stellar light and re-emit it in the FIR. Thus, even a tiny amount of dust extinction can lead to a severe underestimate of the actual star formation rate. New IR, FIR and submm facilities have revealed the existence of populations of sources, such as SCUBA $z \geq 1$ sources, that are thought to be dust-enshrouded star forming galaxies or AGNs, and the Extremely Red Objects, which are at least partly populated by dusty star-forming systems at $z \sim 1$. Finally, dust plays a critical role in galaxy evolution, accelerating the formation of molecular hydrogen (H_2), dominating the heating of gas through emission of photoelectrons in regions where UV fields are present and contributing to gas cooling through IR emission.

Evidences for the presence of dust at high redshifts come from observations of damped Lyα systems and from the detection of dust thermal emission from high redshift QSOs selected from the SDSS survey out to redshifts 5.5 and re-observed at mm wavelengths. Very recently, [8] have reported the observations of three $z > 6$ SDSS QSOs at 1.2 mm, detecting thermal dust radiation. From the IR luminosities, the estimated dust masses are huge ($> 10^8$ $M_\odot$) implying a high abundance of heavy elements and dust at redshifts as high as 6.4 that can not be accounted by low-mass stars. Thus, dust enrichment must have occurred primarily on considerably shorter timescales in the ejecta of supernova explosions [36] and [68] (TF). Reference [22] obtained limits on dust and metallicity evolution of Lyα forest clouds using COBE data by relating the dust content to the metal evolution of the absorbers and assuming that dust is heated by the ultraviolet background radiation and by the CMB. The expected CMB spectral distortions due to high-z dust in Lyα clouds is $\sim$1.25–10 smaller than the current COBE upper limit, depending on the metallicity evolution of the clouds. They also find that the Lyα cloud dust opacity to redshift ~ 5 sources around the observed wavelength $\lambda_0 \sim 1\,\mu$m is ~ 0.13 and that the corresponding CMB spectral distortion is $\sim$1.25–10 smaller than current COBE upper limit on the y-parameter, depending on the metallicity evolution.

The questions that we pose here are the following. *When was dust first formed? Is grain formation possible starting from a metal-free environment? What are the dust properties and amount produced? How are these quantities affected by metallicity changes?*

At high redshift, the contribution to dust production due to evolved stars (M and carbon stars, Wolf-Rayet stars, red giants and supergiants, novae) is

even more negligible or absent. The reason is that the typical evolutionary timescale of these stars ($\geq$ 1 Gyr) is longer than the age of the universe, $t_{\rm H} = 6.6\,h^{-1}(1+z)^{-3/2}$ Gyr in a EdS cosmology, if $(1+z) \geq 5$ (adopting $h = 0.7$). Thus, it seems clear that if high redshift dust exists, it must have been produced by SNe, to which we then devote the rest of this study.

The strongest evidence for dust formation in supernova explosions was seen in SN 1987A. In this event, the increased IR emission was accompanied by a a corresponding decrease in the optical emission and the emission-line profiles were observed to shift toward the blue. In another supernova explosion, SN 1998S, the observed evolution of the hydrogen and helium line profiles argues in favor of dust formation within the ejecta as the redshifted side of the profile steadily faded while the blueshifted side remained constant. Dust emission has been seen in the supernova remnant Cassiopea A : Similarly to SN 1987A, the total dust mass derived from IR luminosities is less than expected from theory, suggesting that a colder population of dust grains may be present that emit at longer wavelengths and it is not detectable in the IR . Finally, in a considerable number of cases, supernovae have shown IR emission that was stronger toward longer wavelengths. This IR excess has been generally interpreted as due to thermal emission from dust forming in the ejecta but alternative explanations exist; in particular, new IR observations of five Type II SNe have shown late-time emission that remains bright many years after the maximum and that it is hard to reconcile with emission from newly formed dust; IR echos from pre-existing dust in the circumstellar medium heated by the supernova flash might represent an alternative interpretation.

Theoretical studies have started to investigate the process of dust formation in expanding SN ejecta. Most of the available models are based on classical nucleation theory and grain growth. The model developed by TF is able to predict the dust mass and properties as a function of the initial stellar progenitor mass and metallicity. In spite of the many uncertainties and approximations, this model has been shown to satisfactorily reproduce the observed properties of SN 1987A and of the young dwarf galaxy SBS 0335-052. In addition to normal SNe, the formation of dust grains in the ejecta of PISN has been investigated [48, 63].

8.1 Dust Formation Model

Dust Nucleation and Accretion

The formation of solid materials from the gas phase can occur only from a vapor in a supersaturated state. Because of the existence of a well defined condensation barrier, expressed by a corresponding "critical cluster" size, the formation of solid particles in a gaseous medium is described as a two-step process: (i) the formation of critical clusters; (ii) the growth of these clusters into macroscopic dust grains. The classical theory of nucleation gives an expression for the nucleation current, J, i.e. the number of clusters of critical

size formed per unit volume and unit time in the gas:

$$J = \alpha\Omega\left(\frac{2\sigma}{\pi m_1}\right)^{1/2} c_1^2 \exp\left\{-\frac{4\mu^3}{27(\ln S)^2}\right\}, \tag{111}$$

where $\mu = 4\pi a_o^2\sigma/k_{\rm B}T$ with a_o the radius of molecules (or atoms, depending on chemical species) in the condensed phase; σ is the specific surface energy (corresponding to surface tension in liquids), $k_{\rm B}$ is the Boltzmann constant and T the gas temperature; m_1 and c_1 are the mass and the concentration of the monomers in the gas phase, respectively; $\Omega = (4/3)\pi a_o^3$ is the volume of the single molecules in the condensed phase, α is the sticking coefficient and S the supersaturation ratio, defined below. The subsequent growth of the clusters occurs by accretion and is described by:

$$\frac{{\rm d}r}{{\rm d}t} = \alpha\Omega v_1 c_1(t), \tag{112}$$

with the condition:

$$r(0) = r_* = \frac{2\sigma\Omega}{k_{\rm B}TlnS}, \tag{113}$$

where v_1 is the mean velocity of monomers, $r(t)$ is the cluster radius at time t, and r_* is the cluster critical radius. Equations (111) and (112) describe nucleation and growth of solid particles in a gas composed of a single chemical species (i.e. reactions of the type Fe(gas) $\rightarrow$ Fe(solid) or SiO(g) $\rightarrow$ SiO(s)). However, there are some compounds (like forsterite, Mg_2SiO_4) whose nominal molecule does not exist in the gas phase. These compounds form directly in the solid phase by means of a chemical reaction with the reactants in the gas phase. We need to extend the theory described above to this situation. Let us consider a vapor in a supersaturated state. In this vapor grains condense homologously via the reaction:

$$\sum_i \nu_i A_i = \text{solid compound}, \tag{114}$$

where A_i's represent the chemical species of reactants and products in the gas phase and ν_i's are stoichiometric coefficients, which are positive for reactants and negative for products respectively. We make the following assumptions: (i) the rates of nucleation and grain growth are controlled by a single chemical species, referred to as a key species. (ii) the key species corresponds to the reactant with the least collisional frequency onto a target cluster. In this case, (111) and (112) become:

$$J = \alpha\Omega\left(\frac{2\sigma}{\pi m_{1k}}\right)^{1/2} c_{1k}^2 \exp\left\{-\frac{4\mu^3}{27(\ln S)^2}\right\}, \tag{115}$$

and

$$\frac{{\rm d}r}{{\rm d}t} = \alpha\Omega v_{1k} c_{1k}(t), \tag{116}$$

where m_{1k}, c_{1k} and v_{1k} are the mass, concentration and mean velocity of monomers of key species, respectively. In this case the supersaturation ratio is expressed by:

$$\ln S = -\frac{\Delta G_r}{RT} + \sum_i \nu_i \ln P_i, \tag{117}$$

where P_i is the partial pressure of the i-th specie, R is the gas constant and ΔG_r is the Gibbs free energy for the reaction (4).
We investigate the formation of the following solid compounds: Al_2O_3 (corundum), iron, Fe_3O_4 (magnetite), $MgSiO_3$ (enstatite), Mg_2SiO4 (forsterite) and amorphous carbon (ACG) grains. These compounds are constituted by the most abundant heavy elements in the ejecta. It is important to note that, as we will see, ACG grains can form also if the ejecta composition is richer in oxygen than in carbon ($O > C$). The formation of CO and SiO molecules in SN ejecta can be very important for dust formation, because carbon atoms bound in CO molecules are not available to form ACG grains and SiO molecules take part in the reactions which lead to the formation of $MgSiO_3$ and Mg_2SiO_4. Thus, the process of molecule formation in the expanding ejecta is followed at temperatures $T \leq 2\,10^4\,\mathrm{K}$ together with the $^{56}\mathrm{Co} \rightarrow ^{56}$ Fe radioactive decay as the impact with energetic electrons produced during this decay represents the main destruction process of CO molecules.

Supernova Model

We now describe the adopted model for the SN ejecta. Before the explosion the progenitor develops the standard "onion skin" stratified structure, with a hydrogen-rich envelope, a helium layer, and several thinner heavy element layers up to a $\mathrm{Fe} - \mathrm{Ni}$ core. During the explosion a shock wave propagates through the layers, reheats the gas and triggers the explosive nucleosynthesis phase. This phase lasts for few hours, then expansion cools the gas and the thermonuclear reactions turn off. After the explosion the SN starts to expand homologously, with velocity $v \propto R$, where R is the distance from the center. During the first weeks Rayleigh-Taylor instabilities cause the mixing of the internal layers. The early emergence of X-rays and γ-rays observed in SN 1987A can be explained if radioactive ^{56}Co is mixed from the internal regions of the star into the external ones; more precisely, observations suggest mixing of the materials in the ejecta at least up to the outer edge of the helium layer. Dust grains are formed by heavy elements so we focus on the volume containing them, i.e. the sphere of radius R, defined as the radius of the outer edge of the He-rich layer. It is thought that mixing forms clumps of heavy elements embedded in the He-rich layer. As a first approximation, we assume that mixing is complete, and that the gas has uniform density and temperature in the considered volume at any given time. Stated differently, we assume that all the chemical species are mixed at the molecular level in the ejecta.

Photometric observations have shown that a SN emits typically 10^{49} erg in electromagnetic energy, but current theoretical models predict kinetic energies $E_{\rm kin} \approx 10^{51}$ erg. The expansion velocity v is then given by: $v \simeq \sqrt{E_{\rm kin}/M_{\rm tot}}$, where $M_{\rm tot}$ is the total mass ejected by the SN. We take the chemical composition of the expelled gas from the results of [72] (WW95), apart from the specific case of SN 1987A, see below. They determine the nucleosynthetic yields of isotopes lighter than A = 66 (Zinc) for a grid of stellar masses and metallicities including stars in the mass range $11-40$ $M_\odot$ and metallicities $(Z/Z_\odot) = 0, 10^{-4}, 0.01, 0.1, 1$. They also give the values for $E_{\rm kin}$ and $M_{\rm tot}$ for all the SN models considered. The range $11-40$ $M_\odot$ is the most relevant mass range for the production of heavy elements. In fact, stars with mass between 8 and 11 $M_\odot$ are characterized by very thin heavy element layers, whereas stars heavier than 40 $M_\odot$ might be rare and give rise to a black hole partially swallowing the nucleosynthetic products.

Expansion of the ejecta leads to cooling of the gas. We have already mentioned that the radiation losses are only a few percent of the total internal energy; their contribution to cooling is even smaller in the first week after the explosion due to the high opacity which prevents photons from escaping from the inner regions. Therefore, it is a good approximation to assume that the expansion is adiabatic, although this hypothesis becomes less correct in the advanced evolutionary stages. Radiation losses are also partly balanced by the heating provided by radioactive decay (especially of ^{56}Ni $\rightarrow^{56}$ Co $\rightarrow^{56}$ Fe). We neglect here these complications and assume that the expansion is adiabatic (for a more detailed treatment of the radiative losses see [39]. In this case (for a perfect gas) the temperature evolution is given by

$$T = T_{\rm i}\left(1 + \frac{v}{R_{\rm i}}t\right)^{3(1-\gamma)};$$

γ is the adiabatic index, $T_{\rm i}$ and $R_{\rm i}$ are the temperature and the radius at the beginning of the computation, t is the time elapsed from this initial epoch. The choice of such parameters for normal SNII and PISN are discussed in detail in TF and in SFS. We will generally use these fiducial values in what follows.

8.2 Dust in Primordial Type II Supernovae

In this section we present the main results concerning dust formation in primordial SNe. Additional cases and non-zero metallicity dust yields can be found in TF. Kinetic energies of 10^{51} erg are not sufficient to completely expel the heavy elements external to the $Ni-Fe$ core of the most massive SNe, and a variable amount of material falls back onto the core, probably forming a neutron star or a black hole. The fallback will mostly affect the inner layers, containing the heaviest elements; as a result, progenitors with masses larger than $\approx$ 20 $M_\odot$ will be prevented from forming dust. For essentially

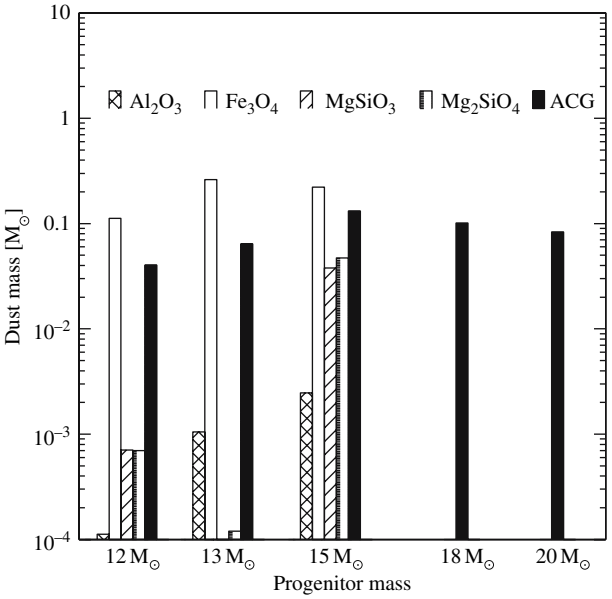

Fig. 26. Dust mass formed in metal-free SNII (in the range 12 $M_\odot < M < 20\ M_\odot$); also shown is the grain composition From Todini & Ferrara (2001) [68]

the same reason, above $M \approx 15\ M_\odot$, only ACG grains are formed. Figure 26 shows the amount of dust formed as a function of progenitor mass, and the grain composition. ACG grains are typically the first solid particles to condense, depending on the models. The formation of these grains is quite fast with respect to the cooling time scale of the ejecta: most of the ACG dust mass forms in a narrow range of $30 \div 40\,\mathrm{K}$ around $T = 1800\,\mathrm{K}$. Subsequently, at a temperature of $\approx 1600\,\mathrm{K}$ Al_2O_3 starts to condense, followed by Fe_3O_4, $MgSiO_3$ and Mg_2SiO_4 at $T \approx 1100\,\mathrm{K}$. Clearly this sequence is governed by the condensation temperature (higher for carbon than silicates) of a given material. Because of this, ACG grains form when the density is still high; as a result the accretion rate proceeds rapidly until complete carbon depletion. Under these conditions few, large grains are formed. Silicate grains, instead, form later and the accretion rate is lower and comparable to the nucleation rate; this leads to the formation to a large number of small grains. Silicate grain growth is also inhibited by the relative paucity of the key species SiO molecule in SN ejecta.

The typical size of ACG dust grains is $a = 300\,\text{Å}$, whereas Fe_3O_4 grains have typically $a = 20\,\text{Å}$ and Al_2O_3, $MgSiO_3$ and Mg_2SiO_4 grains are even smaller ($\approx 10\,\text{Å}$). In spite of the high condensation temperature, Al_2O_3 grains do not grow to sizes comparable to those of ACG, as their growth is limited by the low abundance of Al. The two silicates (enstatite and forsterite) start to condense almost simultaneously; however, Mg_2SiO_4 enters the supersaturation

regime earlier than $MgSiO_3$. For this reason, forsterite grains grow quickly, strongly depleting the Si (or Mg) available. Thus, metal-free SNe can contribute a significant amount of dust: about 0.08 $M_\odot \lesssim M_d \lesssim 0.3$ $M_\odot$ of dust/SN are produced.

8.3 Dust in Pair-instability Supernovae

Figure 27, summarizing results for PISN, shows that in the expanding ejecta of PISN, a significant amount of dust is synthesized out of the heavy elements (adopted yields are from [31]) produced in the progenitor stellar interiors during the main sequence lifetime and at the onset of the explosion due to explosive nucleosynthetic processes.

The fraction of the original progenitor mass which is converted into dust depends on the thermodynamics of the explosion and on the progenitor mass, with values ranging between 7 and 20%, resulting in $10 - 50$ $M_\odot$ of dust produced per SN. These values are much larger than those found for $Z = 0$ Type II SNe resulting from stars with masses between 12 and 40 $M_\odot$ ($0.1 - 1$ $M_\odot$), even if these stars are assumed to be of solar metallicity ($0.1 - 2$ $M_\odot$).

The composition and size of the dominant compounds depends critically on the thermodynamics of the ejecta, i.e. on the assumed value of the adiabatic

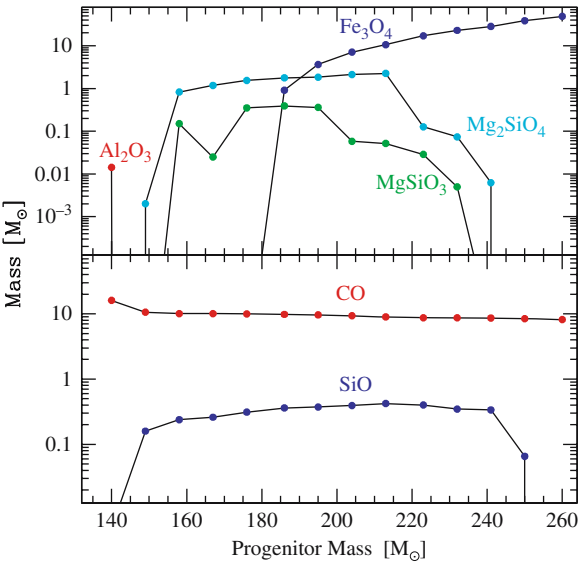

Fig. 27. *Top panel:* Final dust mass of different solid compounds formed in PISN ejecta as a function of the mass of the stellar progenitors. Iron and amorphous carbon grains are never formed and Al_2O_3 is formed only in the ejecta of the lowest mass progenitor. *Bottom panel:* total mass of CO and SiO molecules synthesized in the ejecta as a function of the initial mass of the stellar progenitors

index which controls the temperature evolution of the ejecta and thus determines the epochs and density regimes favorable to dust condensation. In particular, for the two limiting cases explored we find that:

- for $\gamma = 1.25$, grain condensation starts only 40–60 years after the explosions: as a result, ACG grains never form and the dominant compounds are silicates (Mg_2SiO_4 and $MgSiO_3$) for initial stellar masses < 200 $M_\odot$ and magnetite grains for progenitors with larger mass. Because of the large volume of the ejecta, the process of grain accretion is rather inefficient and the typical grain sizes never exceed 10 Å.
- for $\gamma = 4/3$ the grain properties are closer to those found for Type II SNe by TF: grain condensation starts only 50–150 days after the explosion and silicates, magnetite and ACG grains are formed. The dominant compound for all progenitor masses is forsterite Mg_2SiO_4 while the contribution of ACG grains and magnetite grows with progenitor mass. In this case, grain accretion is much more efficient, leading to grain sizes which are never smaller than a few 10 Åand that can be as large as 10^4 Å.

As a by product of the computation, we have estimated the total mass of SiO and CO molecules synthesized in the explosions: a SiO mass which ranges between 0.1 and 0.4 $M_\odot$ is produced, nearly independent of the assumed value for the adiabatic index. Conversely, the mass of CO molecules formed depends on the adiabatic index: for $\gamma = 1.25$ all C initially available in the ejecta is depleted onto CO molecules and roughly 10 $M_\odot$ of CO are formed, independently of the progenitor mass. For $\gamma = 4/3$, the total mass of CO molecules formed decreases with progenitor mass, with values ranging from $0.1 - 10$ $M_\odot$ (Table 7).

The above results have been obtained under the assumption that the heavy elements present in the ejecta at the onset of the explosion are uniformly mixed up to the outer edge of the helium core. In dust formation models developed so far this approximation was motivated by the early emergence of X-rays and γ-rays in SN 1987A, which suggested mixing in the ejecta at least up to the helium core edge. The use of multi-dimensional hydrodynamic codes to model the observed light curves has made clear that mixing occurs on macroscopic scales through the development of Rayleigh-Taylor instabilities: these instabilities arise at the interface between elemental zones and grow

Table 7. Characteristic grain sizes (in Å) synthesized in the ejecta of a 200 $M_\odot$ PISN for different dust compounds and assuming two possible values for the adiabatic index, $\gamma = 1.25$ and $\gamma = 4/3$

γ	Al_2O_3	Fe_3O_4	$MgSiO_3$	Mg_2SiO_4	ACG
1.25		[4–10]	[1–8]	[1–10]	
4/3	[30–400]	[70–500]		[500–4000]	[2000–20000]

non-linearly to produce (i) fingers of heavy elements projected outwards with high velocities and (ii) mixing of lighter elements down to regions that have lower velocities. As a result, the gas in the ejecta is mixed into regions which are still chemically homogeneous and which cool with different timescales, whereas only small clumps in the ejecta are microscopically mixed. It is reasonable to expect that such a structure would affect the process of dust formation, changing both the total amount of dust formed and the relative abundance of different solid compounds.

A related aspect of the model which requires deeper investigation is the assumed temperature structure of the gas within the ejecta, i.e. the assumed adiabatic index. As we have emphasized above, the resulting dust mass and composition depends critically on the cooling timescales of the gas. The temperature and density structure of the stellar interior at the onset of the 250 $M_\odot$ PISN explosion in the simulation of [26] seems to favor a value for the adiabatic index of $\gamma = 4/3$ but a self-consistent model which takes into account the impact of the decays of radioactive elements would be highly desirable.

Finally, to quantify the cosmic relevance of dust formation in the early Universe, we should restrict the analysis to the fraction of the newly formed dust which is able to survive the impact of the reverse shock, following thereafter the fate of the surrounding metal-enriched gas. The process of dust sputtering dissociates the grains into their metal components and might have an important role in cosmic metal enrichment. However, new theoretical models seem to indicate that the post-shock temperature enters the regime suitable for dust condensation inside the oxygen layer, because of the high cooling efficiency of this element, but remains substantially higher in the outer He and H layers. If this is the case, then the reverse shock might lead to an increase in the total amount of dust formed rather than decreasing it.

8.4 Depletion Factors

The cosmological relevance of dust synthesized in PISN explosions will depend mainly on the global properties of dust rather than on the nature and size of the different compounds.

In Fig. 28 we show the total dust depletion factor, defined as the dust-to-metal mass ratio released in the explosions, $f_{\rm dep} = M_{\rm dust}/M_{\rm met}$ as a function of PISN progenitor mass assuming $\gamma = 1.25$ and $\gamma = 4/3$. In the first case, $f_{\rm dep}$ depends sensibly on the initial progenitor mass and ranges between 1% for a 155 $M_\odot$ PISN up to 40% for the highest progenitor masses whereas in the second case we find an almost constant depletion factor of $f_{\rm dep} \approx 0.1$. In the bottom panel of the same figure, we show the ratio of the total dust mass and the initial progenitor stellar mass. In the model with adiabatic index $\gamma = 4/3$, approximately 7% of the initial progenitor mass is synthesized in dust grains. Conversely, if an adiabatic index $\gamma = 1.25$ is assumed, this ratio ranges between 0.7% up to 20% for the highest progenitor masses.

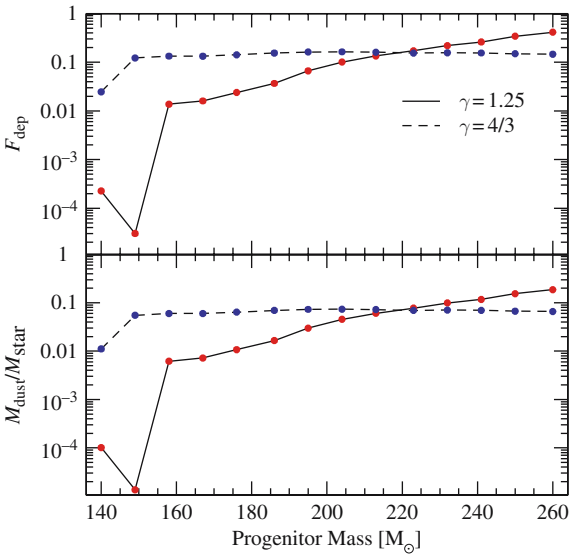

Fig. 28. *Top panel*: total dust depletion factor defined as the dust-to-metal mass ratio as a function of the initial progenitor mass. The two curves refers to two different values of the adiabatic index, $\gamma = 1.25$ and $\gamma = 4/3$. *Bottom panel*: corresponding values of the ratio between the total dust mass synthesized and the initial progenitor mass

The moderate metal yields in Type II SNe and the effect of fallback of material after the explosion onto the compact remnant are responsible for depletion factors which are significantly higher than for PISN, with values ranging from 20% up to 70% for $Z = 0$ progenitors with masses < 22 $M_\odot$. For larger masses, the depletion factor decreases in case B scenarios (higher kinetic energy) because of the larger amount of metals released. For case A scenarios (lower kinetic energy), instead, 25 $M_\odot$ and 30 $M_\odot$ Type II SNe are predicted to have an $f_{dep} = 1$. This is due to the fact that because of fallback, these stars eject only a rather small amount of metals, mostly in the form of carbon, which is completely depleted into ACG grains. If the stellar progenitors have solar metallicities, the depletion factor ranges between 10 and 50% with a much reduced scatter between case A and B.

In spite of these large depletion factors, the total mass of dust synthesized by PISN is significantly higher than that produced by Type II SNe. Indeed, as already discussed in TF, $Z = 0$ Type II SNe synthesize a total dust mass which corresponds to a fraction between 0.3 and 4% of the original stellar progenitor mass (thus, $\sim 0.1 - 0.3$ $M_\odot$ of dust per SN). These values are slightly larger if case B models are considered. When the initial stellar progenitors have solar metallicity, the resulting dust mass is typically a factor ~ 3 larger than for the metal-free case but it is always less than 8% of the initial stellar mass.

8.5 Cosmological Implications

Our analysis shows that if the first stars formed according to a top-heavy IMF and a fraction of them exploded as PISN, a large amount of dust is produced in the early Universe. In this section, we discuss some of the main cosmological implications of this result, with particular emphasis on its role in the thermodynamics of the gas that will be later incorporated into subsequent generations of objects.

Recent numerical and semi-analytical models for the collapse of star-forming gas clouds in the early Universe have shown that because of the absence of metals and the reduced cooling ability of the gas, the formation of low-mass stars is strongly inhibited. In particular, below a critical threshold level of metallicity of $Z_{\rm cr} = 10^{-5\pm1}$ $Z_\odot$ cooling and fragmentation of the gas clouds stop when the temperature reaches a few hundreds K (minimum temperature for H_2 cooling) and the corresponding Jeans mass is of the order of $10^3 - 10^4$ $M_\odot$ ([62]). Gas clouds with mass comparable to the Jeans mass start to gravitationally collapse without further fragmentation, until a central protostellar core is formed which rapidly grows in mass through gas accretion from the surrounding envelope. The absence of metals and dust in the accretion flow and the high gas temperature favor very high accretion efficiencies and the resulting stars can be as massive as 600 $M_\odot$ ([51]).

As the gas becomes more and more enriched with heavy elements, the cooling rate increases because of metal (especially C and O) line emission. More importantly, if a fraction of the available metals is depleted onto dust grains, dust-gas thermal exchanges activate a new phase of cooling and fragmentation which enables the formation of gas clumps with low-mass ([63]). In particular, if the metallicity of the star forming gas clouds exceeds 10^{-4} $Z_\odot$, this dust-driven cooling pathway is irrelevant because cooling via metal-line emission by itself is able to fragment the gas down to characteristic Jeans masses in the range $10^{-2}-1$ $M_\odot$. At the same time, if the metallicity of the gas is below 10^{-6} $Z_\odot$, even if all the available metals are assumed to be depleted onto dust grains, the resulting cooling efficiency is too low to activate fragmentation and the resulting Jeans masses are as large as in the metal-free case (10^3-10^4 $M_\odot$). Thus, we can conclude that the presence of dust is crucial to determine the final mass of stars forming out of gas clouds with metallicities in the critical range $Z_{\rm cr} = 10^{-5\pm1}$ $Z_\odot$. As a reference value, for a metallicity of $Z = 10^{-5}$ $Z_\odot$, if 20% of the metals are depleted onto dust grains ($f_{\rm dep} = 0.2$) the resulting mass of the protostellar core is reduced to $10^{-2}-10^{-1}$ $M_\odot$ and can lead to low-mass stars. This is marginally consistent with the dust formation model with $\gamma = 4/3$ and progenitor masses larger than 150 $M_\odot$ and fully consistent with the model with $\gamma = 1.25$ and progenitor masses larger than 220 $M_\odot$.

Finally, we remind that models predict the formation of a significant amount of CO molecules, which can contribute to gas cooling. This aspect needs to be investigated further though we expect that the presence of CO

molecules will be complementary to that of dust in the critical range of metallicities.

Therefore, in the emerging picture of galaxy evolution, the first episodes of star formation were characterized by very massive stars forming according to a top-heavy IMF. It is only when the metals and dust ejected by the first PISN are able to pollute a substantial fraction of the IGM, that an overall transition to a normal IMF forming stars with masses comparable to those that we presently observe in the nearby Universe occurs. The epoch of this transition depends crucially on the filling factor of the emitted metals: within metal-enriched regions of the Universe, if the metallicity lies within the critical range, the presence of dust becomes critical and can no longer be neglected in the thermodynamics of the star forming gas. It is very likely that the transition will not occur at a single redshift because of the highly inhomogeneous process of metal enrichment ([58]) and that there will be epochs when the two modes of star formation will be coeval in different regions of the Universe.

This might be very important for the reionization history of the IGM. Indeed, very massive metal-free stars are powerful sources of ionizing photons and an early epoch of star formation with a top-heavy IMF can easily match the required optical depth ($\tau_{\rm e} = 0.09 \pm 0.03$) to electron scattering measured by 3-year WMAP. However, an early epoch of reionization is difficult to reconcile with the observed Gunn-Peterson (see Sect. 9) effect in the spectra of $z > 6$ quasars which implies a mass averaged neutral fraction of $\sim 1\%$. These two observational constraints seem to indicate that the reionization history of the Universe might have been more complex than previously thought, possibly with an extended period of incomplete reionization ([12]), probably due to a decrease in the ionizing power of luminous sources caused by their metal enrichment and consequent IMF transition.

Finally, an early epoch of reionization, as required by WMAP data, poses another critical issue: after reionization, the temperature of the IGM starts to decline as a consequence of cosmic expansion. By the time it reaches the observable range of redshifts $z < 6$ the temperature of the IGM might be too low to match the observed thermal history (see Sect. 9). The presence of a significant amount of dust synthesized by the first very massive supernovae might be extremely important to raise the IGM temperature through dust photoelectric heating. To estimate the amount of dust in the IGM required to match the observed thermal history, it is necessary to make specific assumptions about the UV background radiation, the reionization history of the IGM and the grain composition and size distribution. Furthermore, the properties of dust grains in the IGM may differ from those directly predicted by dust formation models as a consequence of specific selection rules in the transfer of grains from the host galaxies to the IGM, who find that only grains larger than $\approx 0.1\,\mu$m are preferentially ejected in the IGM. These complications are beyond the scope of the present analysis.

Neglecting selection rules in the transfer of dust from host galaxies to the IGM, our model allows us to predict depletion factors for specific elements

that may be used in the interpretation of abundance data in the Lyα forest and damped Lyα systems. In Fig. 28 we plot the dust depletion factors for various elements as a function of the progenitor mass assuming two different values for the adiabatic index. In the model with $\gamma = 1.25$ O and Fe depletion are particularly important (more than 10%) for progenitor masses $\geq 180 - 200\ M_\odot$. If γ is taken to be equal to 4/3, a larger variety of elements show significant depletion, particularly Al, C, Si and O.

Finally, it is well known that the presence of dust at high redshifts offers an alternative formation channel for molecular hydrogen, the dominant coolant in the early Universe, which, in the absence of dust, can form only from the gas phase. In small protogalaxies, the H_2 formation rate on grain surface becomes dominant with respect to the formation rate from the gas phase, when the dust-to-gas ratio exceeds roughly 5% of the galactic value. This, in turn, might have very important consequences for the star formation activity at high redshift.

9 The Intergalactic Medium

The study of the IGM has been closely linked to observations related to spectra of distant quasars, in particular to so-called the Lyα forest, though it was not obvious in the beginning whether the Lyα forest traces baryons of cosmological significance. In particular, models in which the Lyα forest arises from some kind of "confined clouds" predicted that the amount of baryons within the forest may not be of cosmological significance. To stress this in slightly more detail, let us briefly review some of the major ideas in the development of this field.

9.1 Historical Background

In a classic paper, Gunn & Peterson showed that the hydrogen in a diffuse uniform IGM must have been highly ionized at $z \approx 2$ in order to avoid complete absorption of the transmitted flux at wavelengths blue-wards of the Lyα emission line of the QSO; this is now commonly known as the Gunn- Peterson (GP) effect. Following that, it was proposed that this GP effect can be used to probe the ionization state of hydrogen within the IGM at various redshifts (and also for other elements). At the same time, it was also realized that gas which was not uniformly distributed would produce discrete Lyα absorption lines. In the beginning, the most natural structures considered were gas clumped into groups of galaxies or low mass protogalaxies. However, these were soon found to be unrealistic when different groups discovered a large number of discrete absorption lines in the QSO spectra, which are usually known as the "Lyα forest". It was shown that these forest lines could not be associated with galaxy clusters, rather they have an intergalactic origin and arise in discrete intergalactic clouds at various cosmological redshifts along

the line of sight (for a recent review and references, see [12]). Various arguments (like the apparent lack of rapid evolution in the properties of the forest, the short relaxation time scales for electrons and protons and short mean free paths) led to the notion that the clouds were "self-contained entities in equilibrium". A two-phase medium was postulated, with the diffuse, very hot, intercloud medium (ICM) in pressure equilibrium with the cooler and denser Lyα clouds. In this two-phase scenario, the ICM was identified with the IGM, while the Lyα clouds were treated as separate entities.

According to the pressure confinement model, the Lyα clouds are supposed to be in photoionization equilibrium with an ionizing ultraviolet (UV) background. The gas is heated by photoionization and cools via thermal bremsstrahlung, Compton cooling, and the usual recombination and collisional excitation processes. Since the ICM is highly ionized, the photoheating is not efficient and hence the medium cools adiabatically through cosmic expansion. The denser clouds embedded in the hot ICM have a nearly constant temperature fixed by thermal ionization equilibrium ($\approx 3 \times 10^4\,\mathrm{K}$). The available range of cloud masses is constrained by the requirement that the clouds must be small enough not to be Jeans-unstable but large enough not to be evaporated rapidly when heated by thermal conduction from the ambient ICM. According to such constraints, clouds formed at high redshifts would survive down to observed redshifts only if their masses range between $10^{5-6}\ \mathrm{M}_\odot$.

The neutral hydrogen within the confining ICM is expected to cause a residual GP absorption trough between the absorption lines (clouds). However, observations at higher spectral resolution revealed no continuous absorption between the discrete lines, placing strong limits on the GP effect, which in turn, puts a strict upper limit on the density of the ICM. The ICM temperature has a lower limit from the absorption line width, while the condition that the cloud must be large enough not to evaporate gives an upper limit on the temperature. Another independent upper limit on the temperature of the ICM comes from the lack of inverse Compton distortions in the spectrum of the cosmic microwave background through the Sunyaev-Zeldovich effect. In fact, the upper limit of the so-called y-parameter is able to rule out any cosmologically distributed component of temperature $> 10^6\,\mathrm{K}$. When all the limits are combined, only a relatively small corner of allowed density-temperature parameter space remains for the ICM. It turns out that, according to the pressure-confinement model, the density of the ICM is too small to be cosmologically significant. Hence, during these early days, the connection between the cosmic reionization and the IGM was not at all obvious as most of the baryons was expected to lie somewhere else. The pressure-confinement model ran into severe problems while trying to match the observed column density distribution. For example, in order to reproduce the low column density systems between, say, $13 < \log(N_{\mathrm{HI}}/\mathrm{cm}^{-2}) < 16$ (where N_{HI} is the column density of neutral hydrogen), the mass has to vary by 9 orders of magnitude. On the other hand, the mass is severely constrained in order to ensure cloud survival. Therefore, the only escape route is to invoke pressure inho-

mogeneities. However, the Lyα absorbers are found to be weakly clustered over a large range of scales, which thus excludes any significant pressure fluctuations. Similarly, detailed hydrodynamical simulations show that the small mass range of the clouds leads to a failure in producing the column density distribution at high $N_{\rm HI}$. In addition, pressure-confinement models predict small cloud sizes which are incompatible with the observations of multiple lines of sight. It was thus concluded that the pure pressure confinement model is unlikely to explain the Lyα forest as a whole though it is possible that some lines of sight must go through sites where gas is locally confined by external pressure (say, the galactic haloes, the likely hosts of the dense Lyman limit absorbing clouds). Even from a theoretical point of view, there are no physical reasons for preferring pressure to gravitational confinement or to no confinement at all. Because of this, self-gravitating baryonic clouds were suggested as an alternative to the pressure confinement model. In this model, the appearance of the IGM as a forest of lines is because of the variations in the neutral hydrogen density rather than a sharp transition between separate entities. In this sense, there is no real difference between an ICM and the clouds in the gravitational confinement model. This scenario of self-gravitating clouds predicts larger sizes of the absorbing clouds ($\approx 1\,$Mpc) compared to the pressure-confinement scenario. However, this model, too, runs into problems while trying to match the observed column density distribution as it predicts larger number of high column density systems than is observed. Secondly, the large absorber sizes seemed to contradict observations. Furthermore, gravitationally confined clouds are difficult to explain theoretically since the mass of such clouds must lie in a restricted range to maintain the gas in equilibrium against free expansion or collapse. As a further alternative, the properties of gas clouds confined by the gravitational field of dark matter have been investigated, more specifically in terms of the "minihalo" model. In this picture, Lyα clouds are a natural byproduct of the cold dark matter (CDM) structure formation scenario. Photoionized gas settles in the potential well of an isothermal dark matter halo. The gas is stably confined if the potential is sufficiently shallow to avoid gravitational collapse but deep enough to prevent the warm gas from escaping. CDM minihalos are more compact than the self-gravitating baryonic clouds because of the larger dark matter gravity, thus alleviating the size problem. The detailed structure of the halo depends on the relative spatial distribution of baryons and CDM. However, the virial radii of the confining objects ($\approx 10\,$kpc) are much lower than the coherence lengths of the Lyα systems as obtained from constraints on absorption line observations of lensed or paired QSOs. It was thus natural to extend the minihalo model to non-static systems. A non-static minihalo model was proposed, who examined the hydrodynamics of a collapsing spherical top-hat perturbation and suggested that clouds were in a free expansion phase.

IGM as a Fluctuating Density Field

Following the non-static models, it was realized that an IGM with the density fluctuation variance of the order of unity could also produce line-like absorptions in quasar spectra. According to such models, the IGM becomes clumpy and acquires peculiar motions under the influence of gravity, and so the Lyα (or GP) optical depth should vary even at the lowest column densities. In a CDM-dominated structure formation scenario, the accumulation of matter in overdense regions reduces the optical depth for Lyα absorption considerably below the average in most of the volume of the universe, leading to what has been called the fluctuating GP phenomenon. Traditional searches for the GP effect that try to measure the amount of matter between the absorption lines were no longer meaningful, as they were merely detecting absorption from matter left over in the most underdense regions. If this is not taken into account, the amount of ionizing radiation necessary to keep the neutral hydrogen GP absorption below the detection limits can be overestimated, which would then have severe implications for reionization studies. In this scenario, the density, temperature and thermal pressure of the medium were described as continuous fields and could not be attributed simply to gravitational confinement or pressure confinement. These studies led to a shift in the paradigm of IGM theories, especially since they implied that the IGM contains most of the baryons at high redshifts, thus making it cosmologically significant and hence quite relevant to cosmic reionization. The actual fluctuation picture can be derived from cosmological N-body and hydrodynamical simulations. It was possible to solve hydrodynamical equations from first principles and set up an evolutionary picture of the IGM in these simulations. Although different techniques and cosmological models were used by different groups, all the simulations indicate a fluctuating IGM instead of discrete clouds. Since in this new paradigm, the Lyα forest arises from a median-fluctuated quasi-linear IGM, it is possible to ignore the high nonlinearities. This made it possible to study the IGM through semi-analytical techniques too. The issue of dealing with quasi-linear densities were dealt in two ways. In the first method, it was showed that a quasi-linear density field, described by a lognormal distribution, can reproduce almost all the observed properties of the Lyα forest. In fact, this was motivated by earlier ideas for dark matter distribution. In an alternate method, it was also possible to obtain the density distribution of baryons from simulations which could then be used for semi-analytical calculations. Given the baryonic distribution, the neutral hydrogen fraction was calculated assuming photoionization equilibrium between the baryons and the ionizing radiation field. It was also realized that the equilibrium between photoheating and adiabatic cooling implies a tight relation between the temperature and density of the gas, described by a power-law equation of state, which was used for determining the temperature of the gas. Given such simplifying and reasonable assumptions, it was possible to make detailed predictions about the Lyα forest. For example, a relation

between column density peaks ("absorption lines") and the statistics of density peaks was proposed, and analytical expressions for the dependence of the shape of the column density distribution on cosmological parameters were obtained. The simulations and the semi-analytical calculations both have been quite successful in matching the overall observed properties of the absorption systems. The shape of the column density distribution and the Doppler parameter distribution are reasonably well reproduced by the simulations as well as semi-analytical calculations over a wide redshift range. The large transverse sizes of the absorbers seen against background paired and lensed QSOs are well explained by the coherence length of the sheets and filaments. In addition, the probability distribution function and power spectrum of the transmitted flux in the Lyα forest is reproduced very well by the models. The Lyα optical depth fluctuations were used for recovering the power spectrum of matter density fluctuations at small scales and also to obtain various quantities related to the IGM. Given the fact that the Lyα can be modeled so accurately, it has become the most useful tool in studying the thermal and ionization history of the universe ever since. Subsequently it was realized that this simple description of the IGM could be coupled to the properties of the ionizing sources and hence it was possible to compute the reheating and reionization history. Since the modelling of the sources is a highly non-linear problem and much more non-trivial to solve that the quasi-linear IGM, it was more natural to make some simple assumptions about the sources, calculate their effect on the IGM and then constrain the properties of the sources themselves.

9.2 Physical Properties

The most basic observable of the Lyα forest is the flux decrement $D_{\rm A}$, or the mean fraction of the QSO continuum absorbed, typically measured between the Lyα and Lyβ emission lines. This is defined as

$$D_{\rm A} = \left\langle 1 - \frac{f_{\rm obs}}{f_{\rm cont}} \right\rangle = \langle 1 - e^{-\tau} \rangle = 1 - e^{-\tau_{\rm eff}} \tag{118}$$

where $f_{\rm obs}$ is the observed (=residual) flux, $f_{\rm cont}$ the estimated flux of the unabsorbed continuum, and τ is the resonance line optical depth as a function of wavelength or redshift ([53]). The absorption is measured against a continuum level usually taken to be a power law in wavelength extrapolated from the region red-ward of the Lyα emission line.

With $D_{\rm A}$ measurements available over a range of redshifts the redshift evolution of the Lyα forest can be investigated. If we characterize a Lyα forest as a random distribution of absorption systems in column density N, Doppler parameter b, and redshift z space, such that the number of lines per interval dN, db and dz is given by $F(N, b, z)$dNdbdz, then

$$\tau_{\rm eff} = \int_{z_1}^{z_2} \int_{b_1}^{b_2} \int_{N_1}^{N_2} {\rm d}N{\rm d}b{\rm d}z (1 - e^{-\tau(N,b)}) F(N, b, z) \tag{119}$$

Assuming that the N and b distribution functions are independent of redshift, and the redshift evolution of the number density of lines can be approximated by a power law, we can write $F(N, b, z) = (1 + z)F(N, b)$, and

$$\tau_{\rm eff}(z) = (1+z)^{\gamma+1}\lambda_0^{-1}\int_{b_1}^{b_2}\int_{N_1}^{N_2} {\rm d}N{\rm d}b \quad e^{-\tau(N,b)}F(N,b)W(N,b) \tag{120}$$

where W is the rest frame equivalent width. This relation enables us to measure the redshift evolution of the number density forest clouds, ${\rm d}N/{\rm d}z \propto (1+z)^{\gamma}$, from the redshift dependence of the effective optical depth ($\tau_{\rm eff} \propto (1+z)^{\gamma+1}$) even if individual absorption lines cannot be resolved.

In order to proper resolve the lines high resolution spectroscopy is required. The standard approach to Voigt profile fitting relies on χ^2 minimization to achieve a complete decomposition of the spectrum into as many independent Voigt profile components as necessary to make the χ^2 probability consistent with random fluctuations. For stronger Lyα lines the higher order Lyman lines can provide additional constraints when fitted simultaneously. Given sufficient spectral resolution, and assuming that Lyα clouds are discrete entities (in the sense of some of the models discussed in the previous section)the profile fitting approach is the most physically meaningful way of extracting information from the Lyα forest. If the absorber is a gas cloud with a purely Gaussian velocity dispersion (a thermal Maxwell-Boltzmann distribution, plus any Gaussian contributions from turbulence) a Voigt profile provides an exact description of the absorption line shape. The Doppler parameter can then be written as the quadratic sum of its individual contributions:

$$b = \sqrt{\frac{2kT}{m} + b_{\rm turb}^2} \tag{121}$$

Unfortunately, in more realistic models of the absorbing gas finite velocity and density gradients invalidate the assumptions underlying Voigt profile fitting, and the line parameters may have less immediate physical meaning. Departures of the absorption line shape from a Voigt profile may contain valuable information about the underlying nature of the absorption systems, and different scenarios may have quite different observational signatures.

Having succinctly described the main investigation techniques we now briefly discuss the current knowledge of the physical properties of the Lyα forest from observations ([17]).

The most impressive feature of the Lyα forest is the rapid increase of the number of lines with redshift shown in Fig. 29. The maximum-likelihood fit to the data at $z > 1.5$ with the power-law parametrization discussed above gives $N(z) = N_0(1+z)^{\gamma} = (6.5 \pm 3.8)(1+z)^{2.4\pm0.2}$. The UVES observations imply that the turn-off in the evolution does occur at $z \approx 1$. While the opacity is varying so fast, the column density distribution stays almost unchanged. The differential density distribution function measured by UVES, that is, the

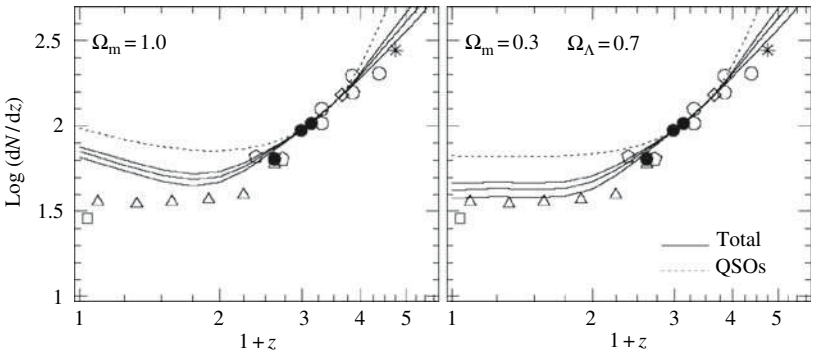

Fig. 29. Number density evolution of the Lyα forest with $N_{\rm HI} = 10^{13.64-16}{\rm cm}^{-2}$. *Dotted lines* refer to the evolution compatible with an ionizing UV background due only to QSOs. *Solid lines* show the expected evolution when both QSOs and galaxies contribute to the background, for models with fesc=0.05 (*upper line*), 0.1 and 0.4 (*lower line*). Data points come from several observations in the literature (see Cristiani et al. 2003 [17])

number of lines per unit redshift path and per unit $N_{\rm HI}$ as a function of $N_{\rm HI}$, follows a power-law $f(N_{\rm HI}) \propto {N_{\rm HI}}^{-1.5}$ extending over 10 orders of magnitude with little, but significant deviations: the slope of the power-law in the range $14 < \log N_{\rm HI} < 16$ goes from about -1.5 at $\langle z \rangle = 3.75$ to -1.7 at $z < 2.4$. This trend continues at lower redshift, with a slope of -2.0 at $z < 0.3$. Physically, the evolution of the $N(z)$ is governed by two main factors: the Hubble expansion and the metagalactic UV background (UVB). At high z both the expansion, which decreases the density and tends to increase the ionization, and the UVB, which is increasing or non-decreasing with decreasing redshift, work in the same direction and cause a steep evolution of the number of lines. At low z, the UVB starts to decrease with decreasing redshift, due to the reduced number and intensity of the ionizing sources, counteracting the Hubble expansion. As a result the evolution of the number of lines slows down. Up to date, numerical simulations have been remarkably successful in qualitatively reproducing the observed evolution, although more precise predictions are jeopardized by the yet poor knowledge of the evolution of the UV background (UVB), and in particular of the relative contribution of different sources (i.e. quasars vs. galaxies). One of the most widely used techniques used so far to determine the amplitude of the UVB is the so-called proximity effect ([3]).

9.3 Proximity Effect

The UV radiation from QSOs has been considered as the most natural origin for the ionization of the intergalactic gas. The finite number density of QSOs suggests that there may be inhomogeneities in the ionization state of the Lyα clouds near each QSO. The term "Proximity Effect" refers to a relative lack

of Lyα absorption in the vicinity of the background QSO. The effect was first discussed in the early 80s, when the currently accepted explanation of increased ionization of the clouds by the nearby QSO was also suggested. It was then realized that the general increase of the absorption line density $\mathrm{d}N/\mathrm{d}z$ with redshift was accompanied by a simultaneous decrease of $\mathrm{d}N/\mathrm{d}z$ in each individual QSO spectrum when approaching the QSO's emission redshift.

If the proximity effect is indeed caused by enhanced ionization measuring the intensity of the ionizing UV background becomes possible from observations of the density of lines, $\mathrm{d}N/\mathrm{d}z$, as a function of the distance from the QSO. Let us assume that in the vicinity of a QSO $\mathrm{d}N/\mathrm{d}z$ is reduced, presumably owing to the excess ionization of the gas. With increasing distance from the emission redshift the QSO's ionizing flux decreases until the UV background intensity begins to dominate the ionization of the intergalactic gas. For example, at the point where the background intensity equals the QSO flux, $L_{\mathrm{Q}}/(4\pi r_{\mathrm{L}})^2$ (known from photometry), the neutral column density of a cloud should be lower by a factor of one half, with a corresponding decrease in $\mathrm{d}N/\mathrm{d}z$ for lines above a given detection threshold. In this way the first crude measurements of the UV background radiation field were carried on, obtaining $J_{21} = 3$, where $J = J_{21} \times 10^{-21}$ erg cm^{-2} s^{-1} Hz^{-1} sr^{-1} is the intensity at the Lyman limit, 912 Å. This result was later confirmed using larger low resolution samples, obtaining $J_{21} = 1^{+3.2}_{-0.7}$. Their measurement procedure (adopted by most later studies) consists of fitting the number density of lines per unit redshift distance $X = \int (1+z)^{\gamma} \mathrm{d}z$

$$\frac{\mathrm{d}N}{\mathrm{d}X} = \left(\frac{\mathrm{d}N}{\mathrm{d}X}\right)_0 \left(1 + \frac{L_{\mathrm{Q}}}{16\pi^2 r_{\mathrm{L}}^2 J}\right)^{1-\beta} \tag{122}$$

as a function of the luminosity distance r_{L}, where the background intensity J is the quantity desired. The quantity β is again the exponent of the power law distribution of column densities. The largest compilations of high resolution data gave $J_{21} = 0.5 \pm 0.1$. None of the studies has found evidence for a significant change with redshift (for $1.6 < z < 4.1$). However, more recent data (see review by [20]) indicate a sharp drop of the UVB intensity for redshift above $z = 5$, perhaps signaling the approach to the end of the reionization epoch.

A more modern version of the proximity effect has been pioneered in the last few years by Adelberger and his group ([2]). In brief such technique uses a survey of the relative spatial distributions of galaxies and intergalactic neutral hydrogen at high redshift. By obtaining high-resolution spectra of bright quasars at intermediate redshifts ($z = 3$–4) and spectroscopic redshifts for a large number of Lyman break galaxies (LBGs) at slightly lower redshifts, and by comparing the locations of galaxies to the absorption lines in the QSO spectra shows that the intergalactic medium (at least for a good fraction of the galaxies belonging to the sample) contains less neutral hydrogen than the

global average within about 1 comoving Mpc of LBGs and more than average at slightly larger distances distances.

Although the interpretation of the lack of HI absorption at small distances from LBGs as a result of a galaxy (as opposed to the previously discussed QSO) proximity effect is quite tempting. However, such explanation might be affected by the presence the galaxies' supernova-driven winds, which could produce a similar effect by heating the gas via their powerful shock waves. In any case this technique is likely to open new avenues for the future better understanding of the very dynamic interplay between galaxies and the intergalactic medium.

References

1. Abel, T., Bryan, G.L., Norman, M.L.: Astrophys. J. **540**, 39 (2000)
2. Adelberger, K.L., Steidel, C.C., Shapley, A.E., Pettini, M.: Astrophys. J. **584**, 45 (2003)
3. Bajtlik, S., Duncan, R.C., Ostriker, J.P.: Astrophys. J. **327**, 570 (1988)
4. Barkana, R., Loeb, A.: Phys. Rep. **349**, 125 (2001)
5. Beers, T.C., et al.: In From Lithium to Uranium: Elemental Tracers of Early Cosmic Evolution, IAU Symposium 228, Hill, V., Francois P., Primas, F. (eds.), (2005) astro-ph/0508423
6. Benson, A.J., Lacey, C.G., Baugh, C.M., Cole, S., Frenk, C.S. Monthly Notices Roy. Astron. Soc. **333**, 156 (2002)
7. Benson A.J., Madau P.: MNRAS **344**, 835 (2003)
8. Bertoldi, F., et al.: Astron. Astrophys. **409**, 47 (2003)
9. Bouwens R.J., Illingworth G.D., Thomson R.I., Franx M.: Astrophys. J. **624**, L5 (2005)
10. Bromm, V., Kudritzki, R.P., Loeb, A.: Astrophys. J. **552**, 464 (2001)
11. Bromm, V., Larson, R.: ARA & A **42**, 79 (2004)
12. Choudhury, T., Ferrara, A.: MNRAS **371**, 55 (2006)
13. Christlieb, N., et al.: Nature **419**, 904 (2002)
14. Ciardi, B., Ferrara, A., Governato, F., Jenkins, A.: Monthly Notices Roy. Astron. Soc. **314**, 611 (2000)
15. Ciardi, B., Ferrara, A.: SSRv **116**, 625 (2005)
16. Cioffi, D.F., McKee, C.F., Bertschinger, E. Astrophys. J. **334**, 252 (1988)
17. Cristiani, S., Bianchi, S., D'Odorico, S., Kim, T.-S.: RMxAC **17**, 275 (2003)
18. Dekel A., Silk J., ApJ **303**, 39 (1986)
19. Dijkstra, M., Haiman, Z., Rees, M.J., Weinberg, D.H.: Astrophys. J. **601**, 666 (2004)
20. Fa., X., Carilli, C., Keating, B., ARA&A, 44, 415 (2006) astro-ph/0602375
21. Ferrara, A.: Astrophys. J. **499**, 17L (1998)
22. Ferrara A., Nath B., Sethi S.K., Shchekinov Y.: MNRAS **303**, 301 (1999)
23. Ferrara A., Tolstoy E.: MNRAS **313**, 291 (2000)
24. Fosbury, R.A.E.: Astrophys. J. **569**, 797 (2003)
25. Frebel A., Christlieb N., Norris J.E., Aoki W., Asplund M.: Nature, **434**, 871 (2005)
26. Fryer, C.L., Woosley, S.E., Heger, A.: Astrophys. J. **550**, 372 (2001)

27. Galli, D., Palla, F.: Astron. Astrophys. **335**, 403 (1998)
28. Gibson, B.K., Loewenstein, M. Mushotzky, R.F., MNRAS, **289**, 632 (1997)
29. Gnedin, N.Y.: Astrophys. J. **542**, 535 (2000)
30. Haardt, F., Madau, P.: Astrophys. J. **461**, 20 (1996)
31. Heger, A., Woosley, S.E.: Astrophys. J. **567**, 532 (2002)
32. Hernandez, X., Ferrara, A.: Monthly Notices Roy. Astron. Soc. **324**, 484 (2001)
33. Hollenbach, D., McKee, C.F.: Astrophys. J. **342**, 306 (1989)
34. Johnson, J.L., Bromm, V.: MNRAS **366**, 247 (2006)
35. Kitayama, T., Susa, H., Umemura, M., Ikeuchi, S.: Monthly Notices Roy. Astron. Soc. **326**, 1353 (2001)
36. Kozasa, T., Hasegawa, H., Nomoto, K.: Astrophys. J. **344**, 325 (1989)
37. Larson, R.B.: MNRAS **145**, 271 (1969)
38. Larson R.B.: MNRAS **169**, 229 (1974)
39. Lattimer, J.M., Schramm, D.N., Grossman, L.: Astrophys. J. **219**, 230 (1978)
40. Lepp, S., Shull, J.L.: Astrophys. J. **280**, 465 (1984)
41. Mac Low, M.-M., Ferrara, A.: ApJ. **513**, 142 (1999)
42. Machacek, M.M., Bryan, G.L., Abel, T.: Astrophys. J. **548**, 509 (2001)
43. Mackey, J., Bromm, V., Hernquist, L.: Astrophys. J. **586**, 1 (2003)
44. Madau, P., Ferrara, A., Rees, M.J.: Astrophys. J. **555**, 92 (2001)
45. McKee, C.F., Draine, B.T.: Science, **252**, 397 (1991)
46. Mori M., Ferrara A., Madau P.: ApJ. **571**, 40 (2002)
47. Nomoto, K., et al.: IAUS **212**, 395 (2003)
48. Nozawa, T., Kozasa, T., Umeda, H., Maeda, K., Nomoto, K.: Astrophys. J. **598**, 785 (2003)
49. Oh, S.P., Nollett, K.M., Madau, P., Wasserburg, G.J.: Astrophys. J. **562**, 1 (2001)
50. Omukai, K., Nishi, R.: Astrophys. J. **518**, 64 (1999)
51. Omukai, K., Palla, F.: Astrophys. J. **589**, 677 (2003)
52. Persic M., Salucci P., Stel F.: MNRAS **281**, 27 (1996)
53. Rauch, M.: ARA&A **36**, 267 (1998)
54. Ricotti, M., Gnedin, N.Y., Shull, J.M.: Astrophys. J. **575**, 49 (2002)
55. Ripamonti, E., Haardt, F., Ferrara, A., Colpi, M.: MNRAS **334**, 401 (2002)
56. Salvaterra, R., Ferrara, A.: MNRAS **367**, 11 (2006)
57. Scannapieco E., Ferrara A., Madau P.: ApJ. **574**, 590 (2002)
58. Scannapieco, E., Schneider, R., Ferrara, A.: Astrophys. J. **589**, 35 (2003)
59. Schaerer, D.: Astron. Astrophys. **382**, 28 (2002)
60. Schaerer, D.: Astron. Astrophys. **397**, 527 (2003)
61. Schaye J., Rauch M., Sargent W.L.W., Kim T.-S.: ApJL, **541**, 1 (2002)
62. Schneider, R., Ferrara, A., Natarajan, P., Omukai, K.: Astrophys. J. **571**, 30. (2002)
63. Schneider, R., Ferrara, A., Salvaterra, R.: MNRAS **351**, 1379 (2003)
64. Shapiro, P.R., Iliev, I.T., Raga, A.C.: Monthly Notices Roy. Astron. Soc. **348**, 753 (2004)
65. Susa, H., Umemura, M.: Astrophis. J. **600**, 1 (2004)
66. Tan, J.C., McKee, C.: 2003, astro-ph/0307414
67. Tegmark, M., Silk, J., Rees, M.J., Blanchard, A., Abel, T., Palla, F.: Astrophys. J. **474**, 1 (1997)
68. Todini, P., Ferrara, A.: MNRAS **325**, 726 (2001)
69. Tsujimoto, T., Nomoto, K., Yoshii, Y., Hashimoto, M., Yanagida, S., Thielemann, F.K., MNRAS, **277**, 945 (1995)

70. Villar-Martn, M., Cerviño, M., Gonzlez Delgado, R.M.: MNRAS **355**, 1132 (2004)
71. White, S.D.M., Frenk, C.: Astrophys. J. **379**, 52 (1991)
72. Woosley, S.E., Weaver, T.A.: ApJSS **101**, 181 (1995)

Observations of the High Redshift Universe

R. S. Ellis

Abstract. In this series of lectures, aimed for non-specialists, I review the considerable progress that has been made in the past decade in understanding how galaxies form and evolve. Complementing the presentations of my theoretical colleagues, I focus primarily on the impressive achievements of observational astronomers. A credible framework, the ΛCDM model, now exists for interpreting these observations: this is a universe with dominant dark energy whose structure grows slowly from the gravitational clumping of dark matter halos in which baryonic gas cools and forms stars. The standard model fares well in matching the detailed properties of local galaxies, and is addressing the growing body of detailed multi-wavelength data at high redshift. Both the star formation history and the assembly of stellar mass can now be empirically traced from redshifts $z \simeq 6$ to the present day, but how the various distant populations relate to one another and precisely how stellar assembly is regulated by feedback and environmental processes remains unclear. In the latter part of my lectures, I discuss how these studies are being extended to locate and characterize the earliest sources beyond $z \simeq 6$. Did early star-forming galaxies contribute significantly to the reionization process and over what period did this occur? Neither theory nor observations are well-developed in this frontier topic but the first results are exciting and provide important guidance on how we might use more powerful future facilities to fill in the details.

1 Role of Observations in Cosmology & Galaxy Formation

1.1 The Observational Renaissance

These are exciting times in the field of cosmology and galaxy formation! To justify this claim it is useful to review the dramatic progress made in the subject over the past $\simeq 25$ years. I remember vividly the first distant galaxy conference I attended: the IAU Symposium 92 *Objects of High Redshift*, held in Los Angeles in 1979. Although the motivation was strong and many observers were pushing their 4 m telescopes to new limits, most imaging detectors were

R. S. Ellis: *Observations of the High Redshift Universe*. In: D. Schaerer et al.: First Light in the Universe, Saas-Fee Advanced Courses, pp. 259–364 (2008)
DOI 10.1007/978-3-540-74163-3_3

still photographic plates with efficiencies of a few percent and there was no significant population of sources beyond a redshift of $z = 0.5$, other than some radio galaxies to $z \simeq 1$ and more distant quasars.

In fact, the present landscape in the subject would have been barely recognizable even in 1990. In the cosmological arena, convincing angular fluctuations had not yet been detected in the cosmic microwave background nor was there any consensus on the total energy density $\Omega_{\rm TOT}$. Although the role of dark matter in galaxy formation was fairly well appreciated, neither its amount nor its power spectrum were particularly well-constrained. The presence of dark energy had not been uncovered and controversy still reigned over one of the most basic parameters of the Universe: the current expansion rate as measured by Hubble's constant. In galaxy formation, although evolution was frequently claimed in the counts and colors of galaxies, the physical interpretation was confused. In particular, there was little synergy between observations of faint galaxies and models of structure formation.

In the present series of lectures, aimed for non-specialists, I hope to show that we stand at a truly remarkable time in the history of our subject, largely (but clearly not exclusively) by virtue of a growth in observational capabilities. By the standards of all but the most accurate laboratory physicist, we have "precise" measures of the form and energy content of our Universe and a detailed physical understanding of how structures grow and evolve. We have successfully charted and studied the distribution and properties of hundreds of thousands of nearby galaxies in controlled surveys and probed their luminous precursors out to redshift $z \simeq 6$ – corresponding to a period only 1 Gyr after the Big Bang. Most importantly, a standard model has emerged which, through detailed numerical simulations, is capable of detailed predictions and interpretation of observables. Many puzzles remain, as we will see, but the progress is truly impressive.

This gives us confidence to begin addressing the final frontier in galaxy evolution: the earliest stellar systems and their influence on the intergalactic medium. When did the first substantial stellar systems begin to shine? Were they responsible for reionizing hydrogen in intergalactic space and what physical processes occurring during these early times influenced the subsequent evolution of normal galaxies?

Let's begin by considering a crude measure of our recent progress. Figure 1 shows the rapid pace of discovery in terms of the *relative fraction* of the refereed astronomical literature in two North American journals pertaining to studies of galaxy evolution and cosmology. These are cast alongside some milestones in the history of optical facilities and the provision of widely-used datasets. The figure raises the interesting question of whether more publications in a given field means most of the key questions are being answered. Certainly, we can conclude that more researchers are being drawn to work in the area. But some might argue that new students should move into other, less well-developed, fields. Indeed, the progress in cosmology, in particular, is

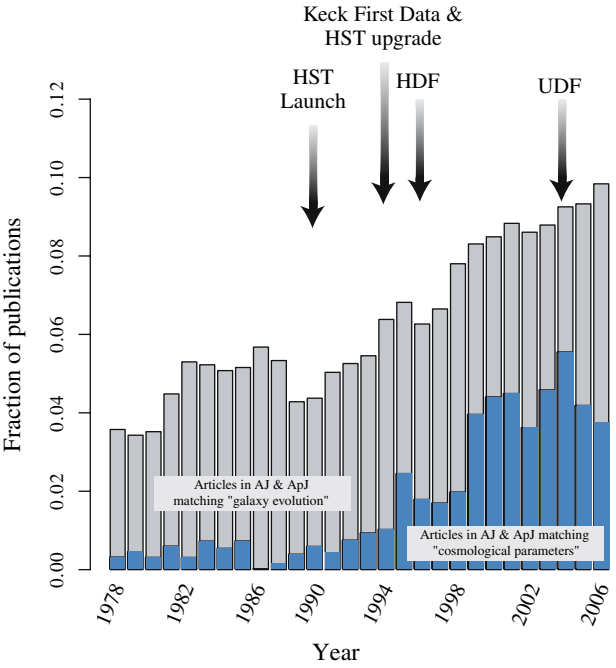

Fig. 1. Fraction of the refereed astronomical literature in two North American journals related to galaxy evolution and the cosmological parameters. The survey implies more than a doubling in fractional share over the past 15 years. Some possibly-associated milestones in the provision of unique facilities and datasets are marked (Courtesy of J. Brinchmann.)

so rapid that some have raised the specter that the subject may soon reaching some form of natural conclusion (c.f. [112]).

I believe, however, that the rapid growth in the share of publications is largely a reflection of new-found observational capabilities. We are witnessing an expansion of *exploration* which will most likely be followed with a more detailed *physical phase* where we will be concerned with *understanding* how galaxies form and evolve.

1.2 Observations Lead to Surprises

It's worth emphasizing that many of the key features which define our current view of the Universe were either not anticipated by theory or initially rejected as unreasonable. Here is my personal short list of surprising observations which have shaped our view of the cosmos:

1. The cosmic expansion discovered by Slipher and Hubble during the period 1917–1925 was not anticipated and took many years to be accepted.

Despite the observational evidence and the prediction from General Relativity for evolution in world models with gravity, Einstein maintained his preference for a static Universe until the early 1930s.

2. The hot Big Bang picture received widespread support only in 1965 upon the discovery of the cosmic microwave background ([184]). Although many supported the hypothesis of a primeval atom, Hoyle and others considered an unchanging "Steady State" universe to be a more natural solutuion.
3. Dark matter was inferred from the motions of galaxies in clusters over seventy years ago ([267]) but no satisfactory explanation of this puzzling problem was ever presented. The ubiquity of dark matter on galactic scales was realized much later ([195]). The dominant role that dark matter plays in structure formation only followed the recent observational evidence [27][1].
4. The cosmic acceleration discovered independently by two distant supernovae teams [185, 192] was a complete surprise (including to the observers, who set out to measure the deceleration). Although the cosmological constant, Λ, had been invoked many times in the past, the presence of dark energy was completely unforeseen.

Given the observational opportunities continue to advance. it seems reasonable to suppose further surprises may follow!

1.3 Recent Observational Milestones

Next, it's helpful to examine a few of the most significant observational achievements in cosmology and structure formation over the past $\simeq 15$ years. Each provides the basis of knowledge from which we can move forward, eliminating a range of uncertainty across a wide field of research.

The Rate of Local Expansion: Hubble's Constant

The Hubble Space Telescope (HST) was partly launched to resolve the puzzling dispute between various observers as regards to the value of Hubble's constant H_0, normally quoted in $\mathrm{km\,sec^{-1}\,Mpc^{-1}}$, or as h, the value in units of $100\,\mathrm{km\,sec^{-1}\,Mpc^{-1}}$. During the planning phases, a number of scientific *key projects* were defined and proposals invited for their execution.

A very thorough account of the impasse reached by earlier ground-based observers in the 1970s and early 1980s can be found in [193] who reviewed the field and concluded a compromise of $67 \pm 15\,\mathrm{km\,sec^{-1}\,Mpc^{-1}}$, surprisingly close to the presently-accepted value. Figure 2 nicely illustrates the confused situation.

[1] For an amusing musical history of the role of dark matter in cosmology suitable for students of any age check out `http://www-astronomy.mps.ohio-state.edu/~dhw/Silliness/silliness.html`.

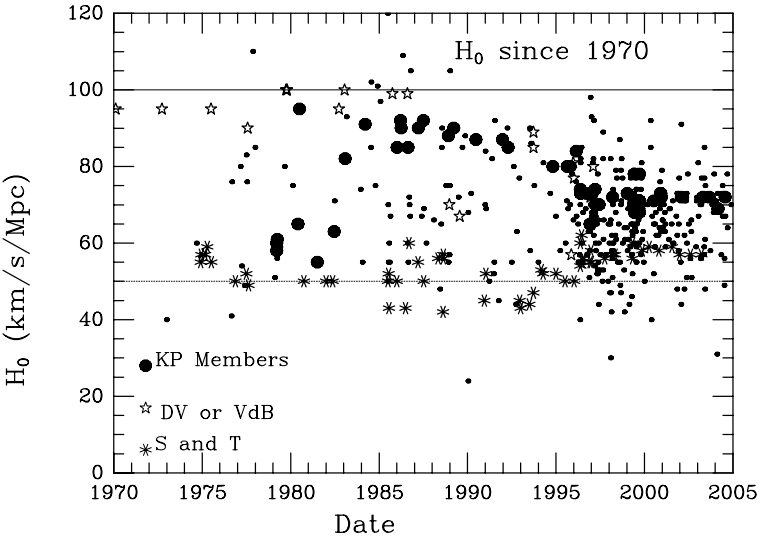

Fig. 2. Various values of Hubble's constant in units of $\mathrm{km\,sec^{-1}\,Mpc^{-1}}$ plotted as a function of the date of publication. Labels refer to estimates by Sandage & Tammann, de Vaucouleurs, van den Bergh and their respective collaborators. Estimates from the HST Key Project group Freedman et al. (2001) [91] are labeled KP. From an initial range spanning $50 < H_0 < 100$, a gradual convergence to the presently-accepted value is apparent. (Plot compiled and kindly made available by J. Huchra)

Figure 3 shows the two stage "step-ladder" technique used by [91] who claim a final value of $67 \pm 15\,\mathrm{km\,sec^{-1}\,Mpc^{-1}}$. "Primary" distances were estimated to a set of nearby galaxies via the measured brightness and periods of luminous Cepheid variable stars located using HST's WFPC-2 imager. Over the distance range across which such individual stars can be seen (<25 Mpc), the leverage on H_0 is limited and seriously affected by the peculiar motions of the individual galaxies. At $\simeq 20$ Mpc, the smooth cosmic expansion would give $V_{\mathrm{exp}} \simeq 1400\,\mathrm{km\,sec^{-1}}$ and a 10% error in H_0 would provide a comparable contribution, at this distance, to the typical peculiar motions of galaxies of $V_{\mathrm{pec}} \simeq 50$–$100\,\mathrm{km\,sec^{-1}}$. Accordingly, a secondary distance scale was established for spirals to 400 Mpc distance using the empirical relationship first demonstrated by Tully & Fisher [250] between the I-band luminosity and rotational velocity. At 400 Mpc, the effect of V_{pec} is negligible and the leverage

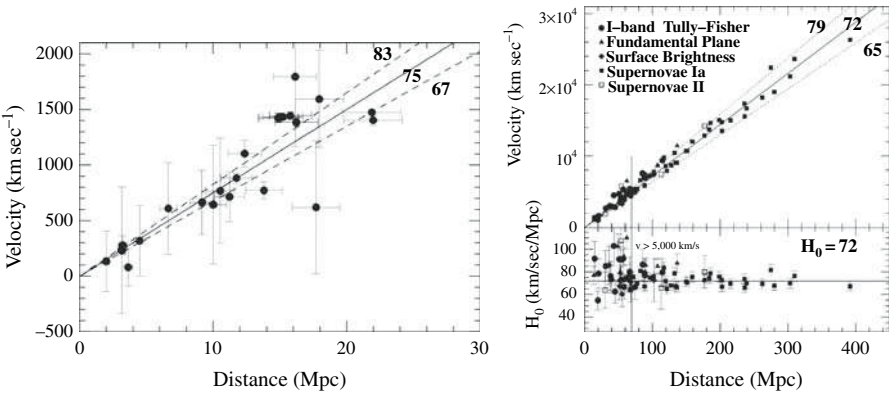

Fig. 3. Two step approach to measuring Hubble's constant H_0 – the local expansion rate (Freedman et al. 2001 [91]). (**Left**) Distances to nearby galaxies within 25 Mpc were obtained by locating and monitoring Cepheid variables using HST's WFPC-2 camera; the leverage on H_0 is modest over such small distances and affected seriously by peculiar motions. (**Right**) Extension of the distance-velocity relation to 400 Mpc using the I-band Tully-Fisher relation and other techniques. The absolute scale has been calibrated using the local Cepheid scale

on H_0 is excellent. Independent distance estimators utilizing supernovae and elliptical galaxies were used to verify possible systematic errors.

Cosmic Microwave Background: Thermal Origin and Spatial Flatness

The second significant milestone of the last 15 years is the improved understanding of the cosmic microwave background (CMB) radiation, commencing with the precise black body nature of its spectrum ([162]) indicative of its thermal origin as a remnant of the cosmic fireball, and the subsequent detection of fluctuations ([217]), both realized with the COBE satellite data. The improved angular resolution of later ground-based and balloon-borne experiments led to the isolation of the acoustic horizon scale at the epoch of recombination [21, 106]. Subsequent improved measures of the angular power spectrum by the Wilkinson Microwave Anisotropy Probe (WMAP, [224, 225]) have refined these early observations. The location of the primary peak in the angular power spectrum at a multiple moment $l \simeq 200$ (corresponding to a physical angular scale of $\simeq 1$ degree) provides an important constraint on the total energy density $\Omega_{\rm TOT}$ and hence spatial curvature.

The derivation of spatial curvature from the angular location of the first acoustic (or "Doppler") peak, $\theta_{\rm H}$, is not completely independent of other cosmological parameters. There are dependences on the scale factor via H_0 and the contribution of gravitating matter $\Omega_{\rm M}$, viz:

$$\theta_{\rm H} \propto (\Omega_{\rm M}\, h^{3.4})^{0.14}\, \Omega_{\rm TOT}^{1.4} \qquad (1)$$

where h is H_0 in units of $100\,\mathrm{km\,sec^{-1}\,Mpc^{-1}}$.

However, in the latest WMAP analysis, combining with distant supernovae data, space is flat to within 1%.

Clustering of Galaxies: Gravitational Instability

Galaxies represent the most direct tracer of the rich tapestry of structure in the local Universe. The 1970s saw a concerted effort to introduce a formalism for describing and interpreting their statistical distribution through angular and spatial two point correlation functions ([183]). This, in turn, led to an observational revolution in cataloging their distribution, first in 2-D from panoramic photographic surveys aided by precise measuring machines, and later in 3-D from multi–object spectroscopic redshift surveys.

The angular correlation function $w(\theta)$ represents the excess probability $\delta\, P$ of finding a pair of galaxies separated by an angular separation θ (degrees).

In a catalog averaging N galaxies per square degree, the probability of finding a pair separated by θ can be written:

$$\delta\, P = N[1 + w(\theta)]\delta\, \Omega \tag{2}$$

where $\delta\, \Omega$ is the solid angle of the counting bin, (i.e. θ to $\theta + \delta\, \theta$).

The corresponding spatial equivalent, $\xi(r)$ in a catalog of mean density ρ per Mpc^3 is thus:

$$\delta\, P = \rho[1 + \xi(r)]\delta\, V \tag{3}$$

One can be statistically linked to the other if the overall redshift distribution of the sources is available.

Figure 4 shows a pioneering detection of the angular correlation function $w(\theta)$ for the Cambridge APM Galaxy Catalog ([158]). This was one of the first well-constructed panoramic 2-D catalogs from which the large scale nature of the galaxy distribution could be discerned. A power law form is evident:

$$w(\theta) = A\theta^{-0.8} \tag{4}$$

where, for example, θ is measured in degree. The amplitude A decreases with increasing depth due to both increased projection from physically-uncorrelated pairs and the smaller projected physical scale for a given angle.

Highly-multiplexed spectrographs such as the 2 degree field instrument on the Anglo-Australian Telescope [53] and the Sloan Digital Sky Survey ([265]) have led to the equivalent progress in 3-D surveys (Fig. 5). In the early precursors to these grand surveys, the 3-D equivalent of the angular correlation function, was also found to be a power law:

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-1.8} \tag{5}$$

where r_o (Mpc) is a valuable *clustering scale length* for the population.

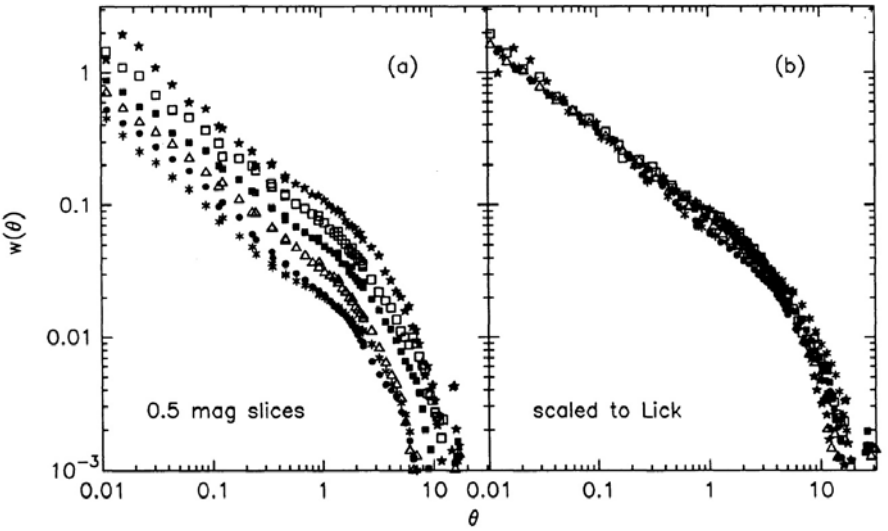

Fig. 4. Angular correlation function for the APM galaxy catalog – a photographic survey of the southern sky (Maddox et al. 1990 [158]) – partitioned according to limiting magnitude (**left**). The amplitude of the clustering decreases with increasing depth due to an increase in the number of uncorrelated pairs and a smaller projected physical scale for a given angle. These effects can be corrected in order to produce a high signal/noise function scaled to a fixed depth clearly illustrating a universal power law form over nearly 3 dex (**right**)

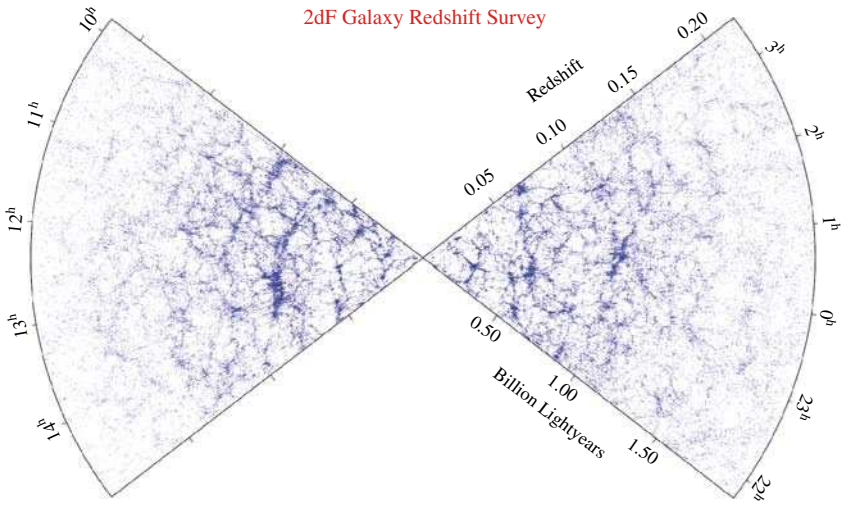

Fig. 5. Galaxy distribution from the completed 2dF redshift survey Colless et al. 2001 [53]

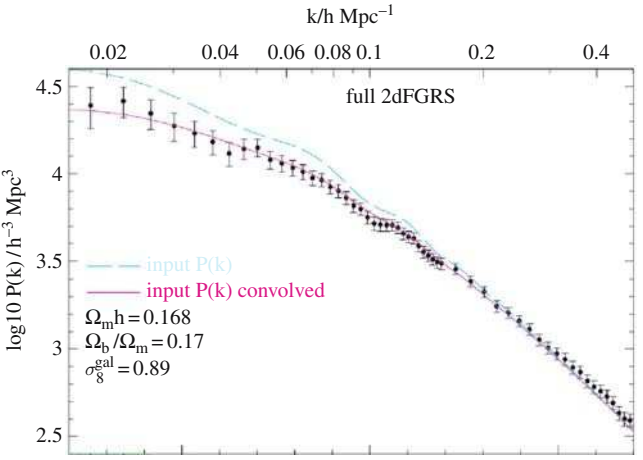

Fig. 6. Power spectrum from the completed 2dF redshift survey Cole et al. (2005) [52]. Solid lines refer to the input power spectrum for a dark matter model with the tabulated parameters and that convolved with the geometric "window function" which affects the observed shape on large scales

As the surveys became more substantial, the power spectrum $P(k)$ has become the preferred analysis tool because its form can be readily predicted for various dark matter models. For a given density field $\rho(\mathbf{x})$, the fluctuation over the mean is $\delta = \rho \,/\, \overline{\rho}$ and for a given wavenumber k, the power spectrum becomes:

$$P(k) = \langle |\delta_k^2| \rangle = \int \xi(r) \exp(\mathrm{i}\mathbf{k}.\mathbf{r}) \mathrm{d}^3 r \tag{6}$$

The final power spectrum for the completed 2dF survey is shown in Fig. 6 [52] and is in remarkably good agreement with that predicted for a cold dark matter spectrum consistent with that which reproduces the CMB angular fluctuations.

Dark Matter and Gravitational Instability

We have already mentioned the ubiquity of dark matter on both cluster and galactic scales. The former was recognized as early as the 1930s from the high line of sight velocity dispersion $\sigma_{\rm los}$ of galaxies in the Coma cluster ([267]). Assuming simple virial equilibrium and isotropically-arranged galaxy orbits, the cluster mass contained with some physical scale $R_{\rm cl}$ is:

$$M = 3\langle \sigma_{\rm los}^2 \rangle R_{\rm cl} / G \tag{7}$$

which far exceeds that estimated from the stellar populations in the cluster galaxies. High cluster masses can also be confirmed completely independently

from *gravitational lensing* where a background source is distorted to produce a "giant arc" – in effect a partial or incomplete "Einstein ring" whose diameter $\theta_{\rm E}$ for a concentrated mass M approximates:

$$\theta_E = \frac{4GM^{\frac{1}{2}}}{c^2}\, D^{\frac{1}{2}} \tag{8}$$

and $D = D_{\rm s}\, D_{\rm l}, / D_{\rm d}s$ where the subscripts s and l refer to angular diameters distances of the background source and lens respectively.

On galactic scales, extended rotation curves of gaseous emission lines in spirals (see review by [194]) can trace the mass distribution on the assumption of circular orbits, viz:

$$\frac{GM(< R)}{R^2} = \frac{V^2}{R} \tag{9}$$

Flat rotation curves ($V \sim$constant) thus imply $M(< R) \propto R$. Together with arguments based on the question on the stability of flattened disks ([177]), such observations were critical to the notion that all spiral galaxies are embedded in dark extensive "halos".

The evidence for halos around local elliptical galaxies is less convincing largely because there are no suitable tracers of the gravitational potential on the necessary scales (see [98]). However, by combining gravitational lensing with stellar dynamics for intermediate redshift ellipticals, [141] and [249] have mapped the projected dark matter distribution and show it to be closely fit by an isothermal profile $\rho(r) \propto r^{-2}$.

The presence of dark matter can also be deduced statistically from the distortion of the galaxy distribution viewed in *redshift space*, for example in the 2dF survey ([182]). The original idea was discussed by [121]. The spatial correlation function $\xi(r)$ is split into its two orthogonal components, $\xi(\sigma, \pi)$ where σ represents the projected separation perpendicular to the line of sight (unaffected by peculiar motions) and π is the separation along the line of sight (inferred from the velocities and hence used to measure the effect). The distortion of $\xi(\sigma, \pi)$ in the π direction can be measured on various scales and used to estimate the line of sight velocity dispersion of pairs of galaxies and hence their mutual gravitational field. Depending on the extent to which galaxies are biased tracers of the density field, such tests indicate $\Omega_{\rm M} = 0.25$.

On the largest scales, *weak gravitational lensing* can trace the overall distribution and dark matter content of the Universe ([25, 190]). Recent surveys are consistent with these estimates ([109]).

Dark Energy and Cosmic Acceleration

Prior to the 1980s observational cosmologists were obsessed with two empirical quantities though to govern the cosmic expansion history – $R(t)$: Hubble's constant $H_0 = {\rm d}R/{\rm d}t$ and a second derivative, the deceleration parameter q_0, which would indicate the fate of the expansion:

$$q_0 = -\frac{\mathrm{d}^2 R/\mathrm{d}t^2}{(\mathrm{d}R/\mathrm{d}t)^2} \tag{10}$$

In the presence only of gravitating matter, Friedmann cosmologies indicate $\Omega_{\mathrm{M}} = 2\,q_0$. The distant supernovae searches were begun in the expectation of measuring q_0 independently of Ω_{M} and verifying a low density Universe.

As we have discussed, Type Ia supernovae (SNe) were found to be *fainter* at a given recessional velocity than expected in a Universe with a low mass density; Fig. 7 illustrates the effect for the latest results from the Canada-France SN Legacy Survey [4]. In fact the results cannot be explained even in a Universe with *no* gravitating matter! A formal fit for q_0 indicates a negative value corresponding to a cosmic acceleration.

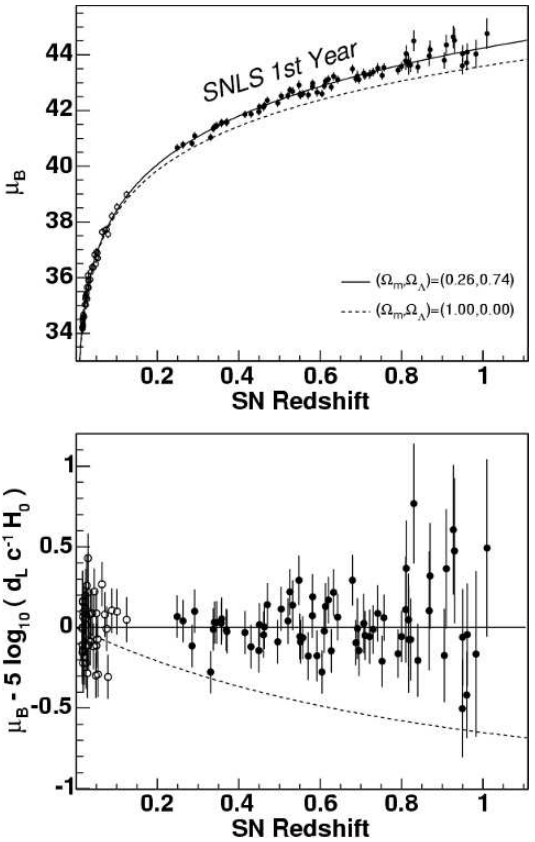

Fig. 7. Hubble diagram (distance-redshift relation) for calibrated Type Ia supernovae from the first year data taken by the Canada France Supernova Legacy Survey Astier et al. (2006) [4]. Curves indicate the relation expected for a high density Universe without a cosmological constant and that for the concordance cosmology (see text)

Acceleration is permitted in Friedmann models with a non-zero cosmological constant Λ. In general ([43]):

$$q_0 = \frac{\Omega_{\rm M}}{2} - 3\,\frac{\Omega_\Lambda}{2} \tag{11}$$

where $\Omega_\Lambda = \Lambda/8\pi G$ is the energy density associated with the cosmological constant.

The appeal of resurrecting the cosmological constant is not only its ability to explain the supernova data but also the spatial flatness in the acoustic peak in the CMB through the combined energy densities $\Omega_{\rm M} + \Omega_\Lambda$ - the so-called *Concordance Model* [5, 178].

However, the observed acceleration raises many puzzles. The absolute value of the cosmological constant cannot be understood in terms of physical descriptions of the vacuum energy density, and the fact that $\Omega_{\rm M} \simeq \Omega_\Lambda$ implies the accelerating phase began relatively recently (at a redshift of $z \simeq 0.7$). Alternative physical descriptions of the phenomenon (termed "dark energy") are thus being sought which can be generalized by imagining the vacuum obeys an equation of state where the negative pressure p relates to the energy density ρ via an index w,

$$p = w\,\rho \tag{12}$$

in which case the dependence on the scale factor R goes as

$$\rho \propto R^{-3(1+w)} \tag{13}$$

The case $w = -1$ would thus correspond to a constant term equivalent to the cosmological constant, but in principle any $w < -1/3$ would produce an acceleration and conceivably w is itself a function of time. The current SNLS data indicate $w = -1.023 \pm 0.09$ and combining with the WMAP data does not significantly improve this constraint.

1.4 Concordance Cosmology: Why is such a Curious Model Acceptable?

According to the latest WMAP results ([225]) and the analysis which draws upon the progress reviewed above (the HST Hubble constant Key Project, the large 2dF and SDSS redshift surveys, the CFHT supernova survey and the first weak gravitational lensing constraints), we live in a universe with the constituents listed in Table 1.

Table 1. Cosmic Constituents

Total Matter	$\Omega_{\rm M}$	0.24 ± 0.03
Baryonic Matter	$\Omega_{\rm B}$	0.042 ± 0.004
Dark Energy	Ω_Λ	0.73 ± 0.04

Given only one of the 3 ingredient is physically understood it may be reasonably questioned why cosmologists are triumphant about having reached the era of "precision cosmology"! Surely we should not confuse measurement with understanding?

The underlying reasons are two-fold. Firstly, many independent probes (redshift surveys, CMB fluctuations and lensing) indicate the low matter density. Two independent probes not discussed (primordial nucleosynthesis and CMB fluctuations) support the baryon fraction. Finally, given spatial flatness, even if the supernovae data were discarded, we would deduce the non-zero dark energy from the above results alone.

Secondly, the above parameters reconcile the growth of structure from the CMB to the local redshift surveys in exquisite detail. Numerical simulations based on 10^{10} particles (e.g. Springel et al. 2005) have reached the stage where they can predict the non-linear growth of the dark matter distribution at various epochs over a dynamic range of 3–4 dex in physical scales. Although some input physics is needed to predict the local galaxy distribution, the agreement for the concordance model (often termed ΛCDM) is impressive. In short, a low mass density and non-zero Λ both seem necessary to explain the present abundance and mass distribution of galaxies. Any deviation would either lead to too much or too little structure.

This does not mean that the scorecard for ΛCDM should be considered perfect at this stage. As discussed, we have little idea what the dark matter or dark energy might be. Moreover, there are numerous difficulties in reconciling the distribution of dark matter with observations on galactic and cluster scales and frequent challenges that the mass assembly history of galaxies is inconsistent with the slow hierarchical growth expected in a Λ-dominated Universe. However, as we will see in later lectures, most of these problems relate to applications in environments where dark matter co-exists with baryons. Understanding how to incorporate baryons into the very detailed simulations now possible is an active area where interplay with observations is essential. It is helpful to view this interplay as a partnership between theory and observation rather than the oft-quoted "battle" whereby observers challenge or call into question the basic principles.

1.5 Lecture Summary

I have spent my first lecture discussing largely cosmological progress and the impressive role that observations have played in delivering rapid progress.

All the useful cosmological functions – e.g. time, distance and comoving volume versus redshift, are now known to high accuracy which is tremendously beneficial for our task in understanding the first galaxies and stars. I emphasize this because even a decade ago, none of the physical constants were known well enough for us to be sure, for example, the cosmic age corresponding to a particular redshift.

I have justified ΛCDM as an acceptable standard model, despite the unknown nature of its two dominant constituents, partly because there is a concordance in the parameters when viewed from various observational probes, and partly because of the impressive agreement with the distribution of galaxies on various scales in the present Universe.

Connecting the dark matter distribution to the observed properties of galaxies requires additional physics relating to how baryons cool and form stars in dark matter halos. Detailed observations are necessary to "tune" the models so these additional components can be understood.

All of this will be crucial if we are correctly predict and interpret signals from the first objects.

2 Galaxies & The Hubble Sequence

2.1 Introduction: Changing Paradigms of Galaxy Formation

We now turn to the interesting history of how our views of galaxy formation have changed over the past 20–30 years. It is convenient to break this into 3 eras

1. The classical era (pre-1985) as articulated for example in the influential articles by Beatrice Tinsley and others. Galaxies were thought to evolve in isolation with their present-day properties governed largely by one function – the time-dependent star formation rate $\psi(t)$. Ellipticals suffered a prompt conversion of gas into stars, whereas spirals were permitted a more gradual consumption rate leading to a near-constant star formation rate with time.
2. The dark matter-based era (1985-): in hierarchical models of structure formation involving gravitational instability, the ubiquity of dark matter halos means that merger driven assembly is a key feature. If mergers redistribute angular momentum, galaxy morphologies are transformed.
3. Understanding feedback and the environment (1995-): In the most recent work, the evolution of the morphology-density relation [66] and the dependence of the assembly history on galactic mass ("downsizing", [49]) have emphasized that star formation is regulated by processes other than gas cooling and infall associated with DM-driven mergers.

2.2 Galaxy Morphology – Valuable Tool or Not?

In the early years, astronomers placed great stock on understanding the origin of the *morphological* distribution of galaxies, sometimes referred to as the Hubble sequence ([115]). Despite this simple categorization 70 years ago, the scheme is evidently still in common use. In its support, Sandage (e.g. [199]) has commented on this classification scheme as describing "a true order among

the galaxies, not one imposed by the classifier". However, many contemporary modelers and observers have paid scant attention to morphology and placed more emphasis on understanding stellar population differences. What value should we place on accurately measuring and reproducing the morphological distribution?

The utility of Hubble's scheme, at least for local galaxies, lies in its ability to distinguish dynamically distinct structures – spirals and S0s are rotating stellar disks, whereas luminous spheroids are pressure-supported ellipsoidal or triaxial systems with anisotropic velocity fields. This contains key information on the degree of dissipation in their formation ([81]).

There are also physical variables that seem to underpin the sequence, including (i) gas content and color which relate to the ratio of the current to past average star formation ratio $\psi(t_0)/\overline{\psi}$ (Fig. 8) and (ii) inner structures including the bulge-to-disk ratio. Various modelers [8] have argued that the bulge-to-disk ratio is closely linked to the merger history and attempted to reproduce the present distribution as a key test of hierarchical assembly.

Much effort has been invested in attempting to classify galaxies at high redshift, both visually and with automated algorithms. This is a challenging task because the precise appearance of diagnostic features such as spiral arms and the bulge/disk ratio depends on the rest-wavelength of the observations. An effect termed the "morphological k-correction" can thus shift galaxies to

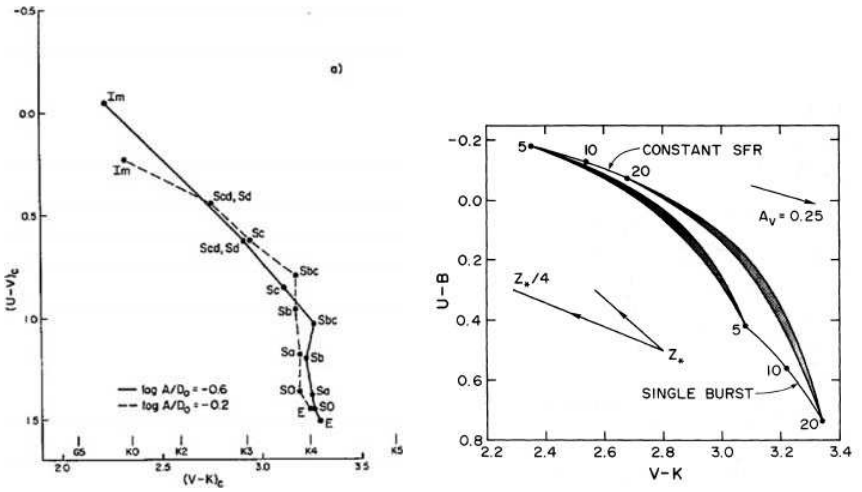

Fig. 8. A succinct summary of the classical view of galaxy formation (pre-1985): (**Left**) The monotonic distribution of Hubble sequence galaxies in the $U-V$ vs $V-K$ color plane Aaronson (1978) [1]. (*Right*) A simple model which reproduces this trend by changing only the ratio of the current to past average star formation rate (Struck-Marcell & Tinsley 1978 [238]). Galaxies with constant star formation permanently occupy the top left (*blue*) corner; galaxies with an initial burst rapidly evolve to the bottom right (*red*) corner

apparently *later* types as the redshift increases for observations conducted in a fixed band. A further limitation, which works in the opposite sense, is surface brightness dimming, which proceeds as $\propto (1+z)^4$, rendering disks less prominent at high redshift and shifting some galaxies to apparent *earlier* types.

The most significant achievements from this effort has been the realization that, despite the above quantitative uncertainties, faint star-forming galaxies are generally more irregular in their appearance than in local samples [67, 100]. Moreover, HST images suggest on-going mergers with an increasing frequency at high redshift ([147]) although quantitative estimates of the merging fraction as a function of redshift remain uncertain (see Bundy et al. 2004) [37].

The idea that morphology is driven by mergers took some time for the observational community to accept. Numerical simulations by [245] provided the initial theoretical inspiration, but the observational evidence supporting the notion that spheroidal galaxies were simple collapsed systems containing old stars was strong [31]. Tell-tale signs of mergers in local ellipticals include the discovery of orbital shells ([161]) and multiple cores revealed only with 2-D dynamical studies [57].

2.3 Semi-Analytical Modeling

As discussed by the other course lecturers and briefly in Sect. 1, our ability to follow the distribution of dark matter and its growth in numerical simulations is well-advanced (e.g. [226]). The same cannot be said of understanding how the baryons destined, in part, to become stars are allocated to each DM halo. This remains the key issue in interfacing theory to observations.

Progress has occurred in two stages – according to the eras discussed in Sect. 2.1. *Semi analytic* codes were first developed in the 1990s to introduce baryons into DM n-body simulations using prescriptive methods for star formation, feedback and morphological assembly (Fig. 9). These codes were initially motivated to demonstrate that the emerging DM paradigm was consistent with the abundance of observational data [50, 123, 218]. Prior to development of these codes, evolutionary predictions were based almost entirely on the "classical" viewpoint with stellar population modeling based on variations in the star formation history $\psi(t)$ for galaxies evolving in isolation e.g. [36].

Initially these feedback prescriptions were adjusted to match observables such as the luminosity function (whose specific details we will address below), as well as specific attributes of various surveys (counts, redshift distributions, colors and morphologies). In the recent versions, more elaborate physically-based models for feedback processes are being considered (e.g. [55])

The observational community was fairly skeptical of the predictions from the first semi-analytical models since it was argued that the parameter space implied by Fig. 9 enabled considerable freedom even for a fixed primordial fluctuation spectrum and cosmological model. Moreover, where different codes could be compared, considerably different predictions emerged [19]. Only as the observational data has moved from colors and star formation rates to

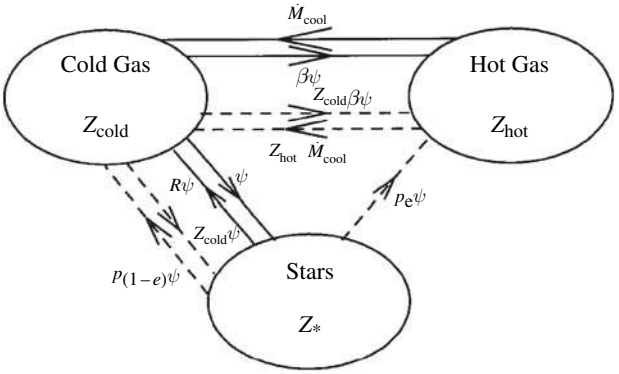

Fig. 9. Schematic of the ingredients inserted into a semi-analytical model (from Cole et al. 2000 [50]). *Solid lines* refer to mass transfer, *dashed lines* to the transfer of metals according to different compositions Z. Gas cooling ($\dot{M}$) and star formation (ψ) is inhibited by the effect of supernovae (β). Stars return some fraction of their mass to the interstellar medium (R) and to the hot gas phase (e) according to a metal yield p

physical variables more closely related to galaxy assembly (such as stellar masses) have the limitations of the early semi-analytical models been exposed.

2.4 A Test Case: The Galaxy Luminosity Function

One of the most straightforward and fundamental predictions a theory of galaxy formation can make is the present distribution of galaxy luminosities – the luminosity function (LF) $\Phi(L)$ whose units are normally per comoving Mpc^3.[2]

As the contribution of a given luminosity bin $\mathrm{d}L$ to the integrated luminosity density per unit volume is $\propto \Phi(L)\, L\, \mathrm{d}L$, an elementary calculation shows that all luminosity functions (be they for stars, galaxies or QSOs) must have a bend at some characteristic luminosity, otherwise they would yield an infinite total luminosity (see [86] for a cogent early discussion of the significance and intricacies of the LF). Recognizing this, [205] proposed the product of a power law and an exponential as an appropriate analytic representation of the LF, viz:

$$\Phi(L)\frac{\mathrm{d}L}{L^*} = \Phi^* \left(\frac{L}{L^*}\right)^{-\alpha} \exp\left(-\frac{L}{L^*}\right) \frac{\mathrm{d}L}{L^*} \tag{14}$$

where Φ^* is the overall normalization corresponding to the volume density at the turn-over (or characteristic) luminosity L^*, and α is the faint end slope which governs the relative abundance of faint and luminous galaxies.

[2] If Hubble's constant is not assumed, it is quoted in units of $h^{-3}\,\mathrm{Mpc}^{-3}$.

The total abundance of galaxies per unit volume is then:

$$N_{\rm TOT} = \int \Phi(L)\,{\rm d}L = \Phi^*\,\Gamma(\alpha+1) \tag{15}$$

and the total luminosity density is:

$$\rho_{\rm L} = \int \Phi(L)\,L\,{\rm d}L = \Phi^*\,\Gamma(\alpha+2) \tag{16}$$

where Γ is the incomplete gamma function which can be found tabulated in most books with integral tables (e.g. [102]). Gradshteyn & Ryzhik 2000). Note that $N_{\rm TOT}$ diverges if $\alpha < -1$, whereas $\rho_{\rm L}$ diverges only if $\alpha < -2$.

Recent comprehensive surveys by the 2dF team ([175]) and by the Sloan Digital Sky Survey [26] have provided definitive values for the LF in various bands. Encouragingly, when allowance is made for the various photometric techniques, the two surveys are in excellent agreement. Fig. 10 shows the Schechter function is a reasonably good (but not perfect) fit to the 2dF data limited at apparent magnitude $b_{\rm J} < 19.7$. Moreover there is no significant difference between the LF derived independently for the two Galactic hemispheres. The slight excess of intrinsically faint galaxies in the northern cap is attributable to a local inhomogeneity in the nearby Virgo supercluster.

Fundamental though this function is, despite ten years of semi-analytical modeling, reproducing its form has proved a formidable challenge (as discussed by [20, 55, 58]). Early predictions also failed to reproduce the color distribution along the LF. The halo mass distribution does not share the sharp bend at L^* and too much star formation activity is retained in massive galaxies. These early predictions produced too many luminous blue galaxies and too many faint red galaxies [32].

As a result, more specific feedback recipes have been created to resolve this discrepancy. Several physical processes have been invoked to regulate star formation as a function of mass, viz:

- Reionization feedback: radiative heating from the first stellar systems at high redshift which increases the Jeans mass, inhibiting the early formation of low mass systems,
- Supernova feedback: this was considered in the early semi-analytical models but is now more precisely implemented so as to re-heat the interstellar medium, heat the halo gas or even eject the gas altogether from low mass systems,
- Feedback from active galactic nuclei: the least well-understood process with various modes postulated to transfer energy from an active nucleus to the halo gas.

References [20] and [55] illustrate the effects of these more detailed prescriptions for these feedback modes on the predicted LF and find that supernova and reionization feedback largely reduce the excess of intrinsically faint

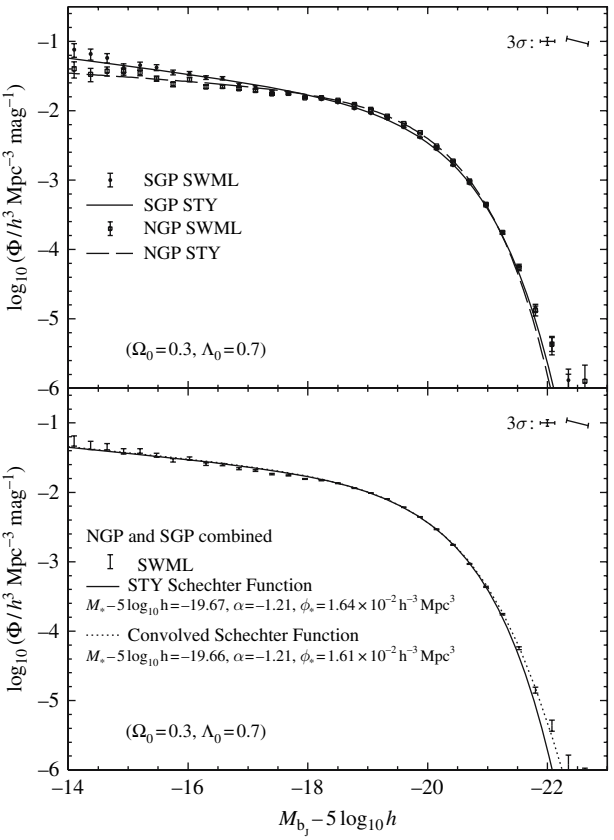

Fig. 10. Rest-frame b_J luminosity function from the 2dF galaxy redshift survey (Norberg et al. 2002 [175]). (**Top**) a comparison of results across the northern and southern Galactic caps; there is only a marginal difference in the abundance of intrinsically faint galaxies. (**Bottom**) combined from both hemispheres indicating the Schechter function is a remarkably good fit except at the extreme ends of the luminosity distribution

galaxies but, on grounds of energetics, only AGN can inhibit star formation and continued growth in massive galaxies. There remains an excess at the very faint end (Fig. 11).

2.5 The Role of the Environment

In addition to recognizing that more elaborate modes of feedback need to be incorporated in theoretical models, the key role of the environment has also emerged as an additional feature which can truncate star formation and alter galaxy morphologies.

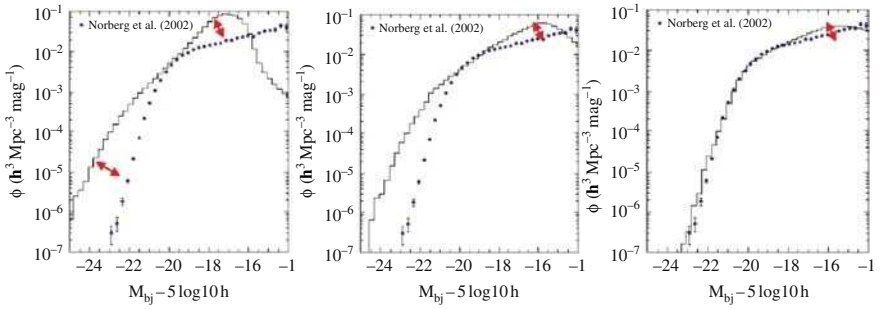

Fig. 11. The effect of various forms of feedback (dots) on the resultant shape of the blue 2dF galaxy luminosity function (Norberg et al. 2002 [175]). From left to right: no feedback, supernova feedback only, supernova, reionization & AGN feedback (Courtesy of Darren Croton.)

The preponderance of elliptical and S0 galaxies in rich clusters was noticed in the 1930's but the first quantitative study of this effect was that of [64] who correlated the fraction of galaxies of a given morphology T above some fixed luminosity with the projected galaxy density, Σ, measured in galaxies Mpc^{-2}.

The local $T - \Sigma$ relation was used to justify two rather different possibilities. In the first, the *nature* hypothesis, those galaxies which formed in high density peaks at early times were presumed to have consumed their gas efficiently, perhaps in a single burst of star formation. Galaxies in lower density environments continued to accrete gas and thus show later star formation and disk-like morphologies. In short, segregation was established at birth and the present relation simply represents different ways in which galaxies formed according to the density of the environment at the time of formation. In the second, the *nurture* hypothesis, galaxies are transformed at later times from spirals into spheroidals by environmentally-induced processes.

Work in the late 1990s, using morphologies determined using Hubble Space Telescope, confirmed a surprisingly rapid evolution in the $T - \Sigma$ relation over $0 < z < 0.5$ [54, 66] strongly supporting environmentally-driven evolution along the lines of the *nurture* hypothesis. Impressive Hubble images of dense clusters at quite modest redshifts ($z \simeq 0.3$–0.4) showed an abundance of spirals in their cores whereas few or none exist in similar environs today.

What physical processes drive this relation and how has the $T - \Sigma$ relation evolved in quantitative detail? Recent work ([187, 216], Fig. 12) has revealed that the basic relation was in place at $z \simeq 1$, but that the fraction $f_{\rm E+S0}$ of Es and S0s has doubled in dense environments since that time. Smith et al. suggest that a continuous, density-dependent, transformation of spirals into S0s would explain the overall trend. Reference [247] likewise see a strong dependence of the fraction as a function of Σ (and to a lesser extent with

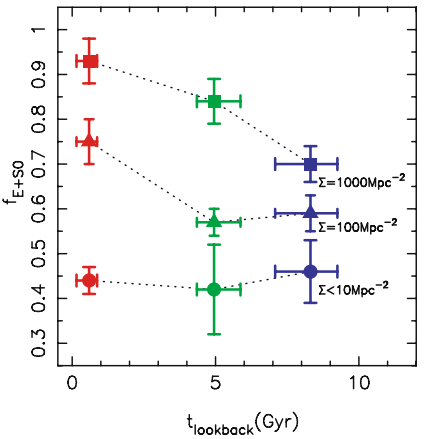

Fig. 12. The fraction of observed E and S0 galaxies down to a fixed rest-frame luminosity as a function of lookback time and projected density Σ from the study of Smith et al. 2005 [216]. Although the morphology density relation was already in place at $z \simeq 1$, there has been a continuous growth in the fraction subsequently, possibly as a result of the density-dependent transformation of spirals into S0s

cluster-centric radius) within a well-studied cluster at $z \simeq 0.4$; they review the various physical mechanisms that may produce such a transformation.

Figure 12 encapsulates much of what we now know about the role of the environment on galaxy formation. The early development of the $T - \Sigma$ relation implies dense peaks in the dark matter distribution led to accelerated evolution in gas consumption and stellar evolution and this is not dissimilar to the *nature* hypothesis. However, the subsequent development of this relation since $z \simeq 1$ reveals the importance of environmentally-driven morphological transformations.

2.6 The Importance of High Redshift Data

This glimpse of evolving galaxy populations to $z \simeq 1$ has emphasized the important role of high redshift data. In the case of the local morphology-density relation, Hubble morphologies of galaxies in distant clusters have given us a clear view of an evolving relationship, partly driven by environmental processes. Indeed, the data seems to confirm the *nurture* hypothesis for the origin of the morphology-density relation.

Although we can place important constraints on the past star formation history from detailed studies of nearby galaxies, as the standard model now needs several additional ingredients (e.g. feedback) to reproduce even the most basic local properties such as the luminosity function (Fig. 11), data

at significant look-back times becomes an essential way to test the validity of these more elaborate models.

Starting in the mid-1990s, largely by virtue of the arrival of the Keck telescopes – the first of the new generation of 8–10 m class optical/infrared telescopes – and the refurbishment of the Hubble Space Telescope, there has been an explosion of new data on high redshift galaxies.

It is helpful at this stage to introduce three broad classes of distant objects which will feature significantly in the next few lectures. Each gives a complementary view of the galaxy population at high redshift and illustrates the challenge of developing a unified vision of galaxy evolution.

- *Lyman-break galaxies: color-selected luminous star forming galaxies at* $z > 2$. First located spectroscopically by [234, 235, 236, 237], these sources are selected by virtue of the increased opacity shortward of the Lyman limit ($\lambda = 912$ Å) arising from the combined effect of neutral hydrogen in hot stellar atmospheres, the interstellar gas and the intergalactic medium. When redshifted beyond $z \simeq 2$, the characteristic "drop out" in the Lyman continuum moves into the optical (Fig. 13). We will review the detailed properties of this, the most well-studied, distant galaxy population over $2 < z < 5$ in subsequent lectures.
- *Sub-millimeter star forming sources:* The SCUBA 850 μm array on the 15 m James Clerk Maxwell Telescope and other sub-mm imaging devices have also been used to locate distant star forming galaxies ([116, 215]). In this case, emission is detected from dust, heated either by vigorous star formation or an active nucleus. Remarkably, their visibility does not fall off significantly with redshift because they are detected in the Rayleigh-Jeans tail of the dust blackbody spectrum [23].

Progress in understanding the role and nature of this population has been slower because sub-mm sources are often not visible at optical and near-infrared wavelengths (due to obscuration) and the positional accuracy of the sub-mm arrays is too coarse for follow-up spectroscopy. The importance of sub-mm sources lies in the fact that they contribute significantly to the star formation rate at high redshift. Regardless of their redshift, the source density at faint limits is 1000 times higher than a no-evolution prediction based on the local abundance of dusty IRAS sources. For several years the key issue was to nail the redshift distribution.

Progress has been made by securing accurate positions using radio interferometers such as the VLA ([90]). About 70% of those brighter than 5 mJy have VLA detections and spectroscopic redshift have now been determined for a significant fraction of this population ([45, 46], Fig. 14).

- *Passively-Evolving Sources:* The Lyman-break and sub-mm sources are largely star-forming galaxies. The arrival of panoramic near-infrared cameras has opened the possibility of locating quiescent sources that are no longer forming stars. Such sources would not normally be detected via the

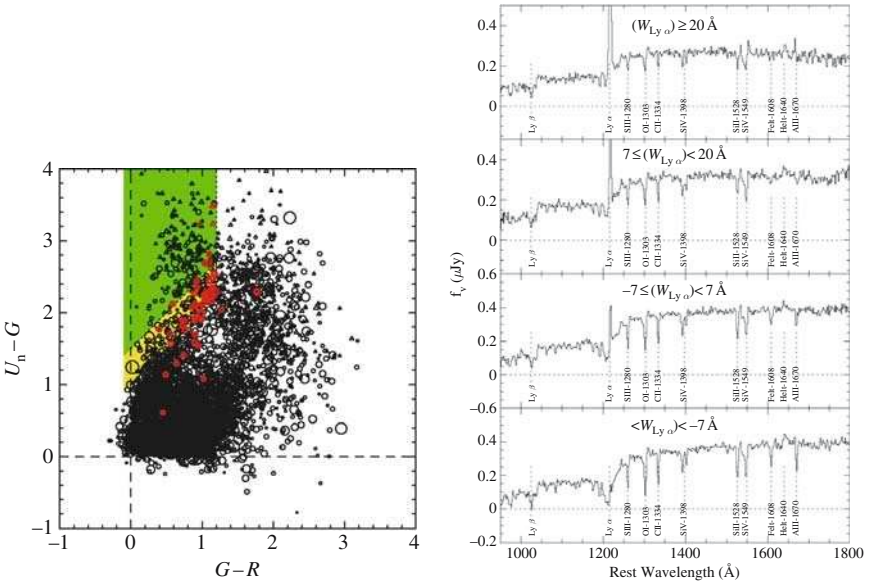

Fig. 13. Location and verification of the Lyman break population in the redshift range $2.7 < z < 3.4$ ([237]). (**Left**) UGR color-color plane for a single field; green and yellow shading refers to variants on the color selection of high redshift candidates. Contaminating Galactic stars cut across the lower right corner of this selection; those confirmed spectroscopically are marked in red. (**Right**) Coadded rest-frame Keck spectra for samples of typically 200 Lyman break galaxies binned according to the strength of Lyman α emission

other techniques and so understanding their contribution to the integrated stellar mass at, say, $z \simeq 2$, is very important.

The nomenclature here is confusing with intrinsically red sources being termed "extremely red objects (EROs)" or "distant red galaxies (DRGs)" with no agreed selection criteria (see [150] for a review). When star formation is complete, stellar evolution continues in a passive sense with main sequence dimming; the galaxy fades and becomes redder.

Of particular interest are the most distant examples which co-exist alongside the sub-mm and Lyman-break galaxies, i.e. at $z > 2$ selected according to their infrared J–K color ([255], Fig. 15).

2.7 Lecture Summary

We have seen in this brief tour that galaxy formation is a process involving gravitational instability driven by the hierarchical assembly of dark matter halos; this component we understand well. However, additional complexities arise from star formation, dynamical interactions and mergers, environmental

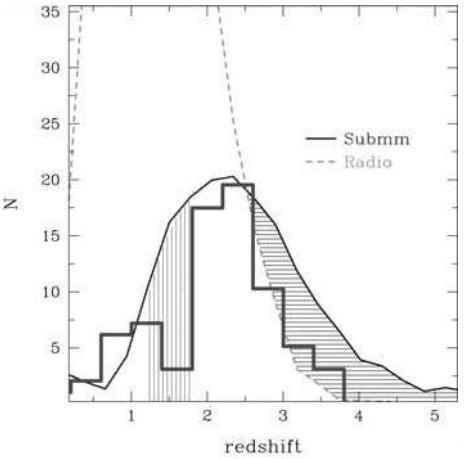

Fig. 14. Redshift distribution of 73 radio-identified SCUBA sources from Chapman et al. (2005) [46]. To illustrate the possible bias arising from the necessary condition of a radio position for a Keck redshift, the solid curve represents a model prediction for the entire >5 mJy sub-mm population. The sub-mm population does not seem to extend significantly beyond z $\simeq$ 4 and has a median redshift of $z = 2.2$

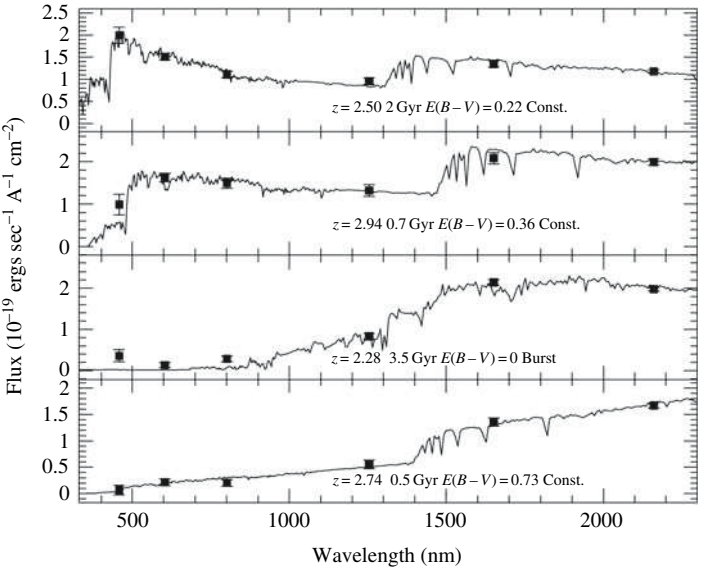

Fig. 15. Observed spectral energy distribution of distant red sources selected with a red *J–K* color superimposed on model spectra (Franx et al. 2003 [89]). Although some sources reveal modest star formation and can be spectroscopically confirmed to lie above $z \simeq 2$, others (such as the lower two examples) appear to be passively-evolving with no active star formation

processes and various forms of feedback which serve to regulate how star formation continues as galaxies grow in mass.

Theorists have attempted to deal with this complexity by augmenting the highly-successful numerical (DM-only) simulations with semi-analytic tools for incorporating these complexities. As the datasets have improved so it is now possible to consider "fine-tuning" these semi-analytical ingredients. Ab initio modeling is never likely to be practical.

I think it fair to say that many observers have philosophical reservations about this "fine-tuning" process in the sense that although it may be possible to reach closure on models and data, we seek a deeper understanding of the physical reality of many of the ingredients. This is particularly the case for feedback processes. Fortunately, high redshift data forces this reality check as it gives us a direct measure of the galaxy assembly history which will be the next topics we discuss.

As a way of illustrating the importance of high redshift data, I have introduced three very different populations of galaxies each largely lying in the redshift range $2 < z < 4$. When these were independently discovered, it was (quite reasonably) claimed by their discoverers that their category represented a major, if not the most significant, component of the distant galaxy population. We now realize that UV-selected, sub-mm selected and non-star forming galaxies each provide a complementary view of the complex history of galaxy assembly and the challenge is to complete the "jig-saw" from these populations.

3 Cosmic Star Formation Histories

3.1 When Did Galaxies Form? Searches for Primeval Galaxies

The question of the appearance of an early forming galaxy goes back to the 1960s. Reference [181] imagined the free-fall collapse of a $700\,L^*$ system at $z \simeq 10$ and predicted a diffuse large object with possible Lyman α emission. Reference [167] considered primeval galaxies might be compact and intense emitters such as quasars.

In the late 1970s and 1980s when the (then) new generation of 4 m telescopes arrived, astronomers sought to discover the distinct era when galaxies formed. Stellar synthesis models [36, 244] suggested present-day passive systems (E/S0s) could have formed via a high redshift luminous initial burst. Placed at $z \simeq 2$–3, sources of the same stellar mass would be readily detectable at quite modest magnitudes, $B \simeq 22$–23, and provide an excess population of blue galaxies.

In reality, the (now well-studied) excess of faint blue galaxies over locally-based predictions is understood to be primarily a phenomenon associated with a gradual increase in star formation over $0 < z < 1$ rather than one due to a distinct new population of intensely luminous sources at high redshift

[75, 140]. Moreover, dedicated searches for suitably intense Lyman α emitters were largely unsuccessful. Reference [188] comprehensively reviews a decade of searching.

Our thinking about primeval galaxies changed in two respects in the late 1980's. Foremost, synthesis models such as those developed by Tinsley and Bruzual assumed isolated systems; dark matter-based models emphasized the gradual assembly of massive galaxies. This change meant that, at $z \simeq 2$–3, the abundance of massive galaxies should be much reduced. Secondly, the flux limits searched for primeval galaxies were optimistically bright; we slowly realized the more formidable challenge of finding these enigmatic sources.

3.2 Local Inventory of Stars

An important constraint on the past star formation history is the present-day stellar density. The former must, when integrated, yield the latter. References [94] and [93] have considered this important problem based on local survey data provided by the SDSS ([126]) and 2dF [51] redshift samples.

The derivation of the integrated density of stars involves many assumptions and steps but is based primarily on the local infrared (K-band) luminosity function of galaxies. The rest-frame K luminosity of a galaxy is a much more reliable proxy for its stellar mass than that at a shorter (e.g. optical) wavelength because its value is largely irrespective of the past star formation history – a point illustrated by [122] (Fig. 16). Another way to phrase this is to say that the infrared mass/light ratio (M/L_K) is fairly independent of the star formation history, so that the stellar mass can be derived from the observed K-band luminosity by a multiplicative factor.

In practice the mass/light ratio depends on the assumed distribution of stellar masses in a stellar population. The zero age or initial mass function is usually assumed to be some form of power law which can only be determined reliable for Galactic stellar populations, although constraints are possible for extragalactic populations from colors and nebular line emission (see reviews by [44, 134, 204]).

In its most frequently-used form the IMF is quoted in mass fraction per logarithmic mass bin: viz:

$$\xi(\log\, m) = \frac{\mathrm{d}n}{\mathrm{d}\, \log\, m} \propto m^{-x} \tag{17}$$

or, occasionally,

$$\xi(m) = \frac{\mathrm{d}n}{\mathrm{d}m} = \frac{1}{m(\ln\, 10)} \xi(\log m) \propto m^{-\alpha} \tag{18}$$

where $x = \alpha - 1$.

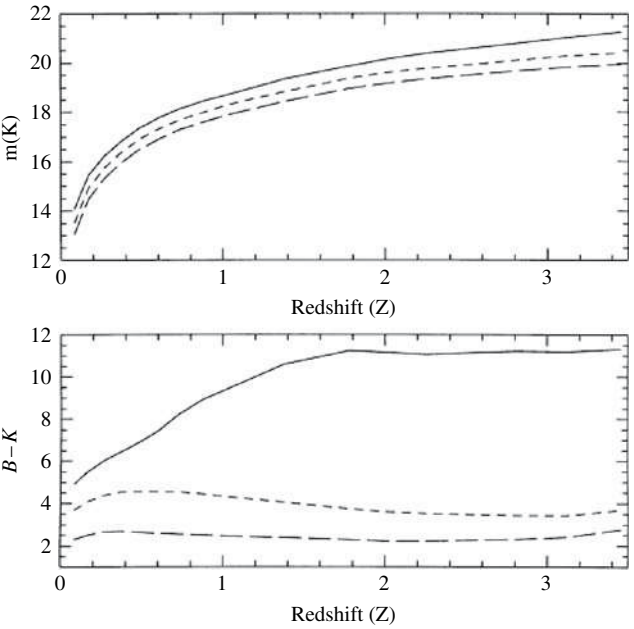

Fig. 16. The robustness of the K-band luminosity of a galaxy as a proxy for stellar mass (Kauffmann & Charlot 1998 [122]). The upper panel shows the K-band apparent magnitude of a galaxy defined to have a fixed stellar mass of $10^{11}\,\mathrm{M}_{\odot}$ when placed at various redshifts. The different curves represent extreme variations in the way the stellar population was created. Whereas the B–K color is strongly dependent upon the star formation history, the K-band luminosity is largely independent of it

In his classic derivation of the IMF, [197] determined a pure power law with $x = 1.35$. More recently adopted IMFs are compared in Fig. 17. They differ primarily in how to restrict the low mass contribution, but there is also some dispute on the high mass slope (although the Salpeter value is supported by various observations of galaxy colors and Hα distributions, [134]).

The IMF has a direct influence on the assumed M/L_{K} (as discussed by [6, 44, 93]) in a manner which depends on the age, composition and past star formation history. The adopted mass/light ratio is then a crucial ingredient for computing both stellar masses (Lecture 4) and galaxy colors.

Baldry[3] has undertaken a very useful comparative study of the impact of various IMF assumptions using the PEGASE 2.0 stellar synthesis code for a population 10 Gyr old with solar metallicity, integrating between stellar masses of 0.1 and 120 $\mathrm{M}_{\odot}$ (Table 2). Stellar masses have been defined in various ways as represented by the 3 columns in Table 2. Typically we are interested in the *observable stellar mass* at a given time (i.e. main sequence

[3] http://www.astro.livjm.ac.uk/~ikb/research/imf-use-in-cosmology.html.

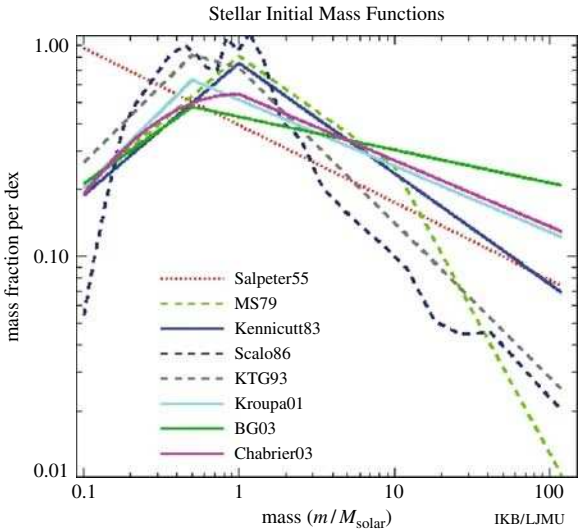

Fig. 17. A comparison of popular stellar initial mass functions (Courtesy of Ivan Baldry).

and giant branch stars), but it is interesting to also compute the total mass which is not in the interstellar medium, which includes that locked in evolved degenerate objects (white dwarfs and black holes). The most inclusive definition of stellar mass (total) is the integral of the past star formation history. Depending on the definition, and chosen IMF, the uncertainties range almost over a factor of 4 for the most popularly-used functions, quite apart from the unsettling question of whether the *form* of the IMF might vary with epoch or type of object.

Although the stellar mass function for a galaxy survey can be derived assuming a fixed mass/light ratio, the useful stellar density is that corrected

Table 2. K-band Stellar Mass/Light Ratios

Source	Stars	Stars + WDs/BHs	Total (Past SFR)
Salpeter (1955)	1.15	1.30	1.86
Miller Scalo (1979) [169]	0.46	0.60	0.99
Kennicutt (1983) [133]	0.46	0.60	1.06
Scalo (1986)	0.52	0.61	0.84
Kroupa et al. (1993) [143]	0.65	0.76	1.09
Kroupa (2001) [144]	0.67	0.83	1.48
Baldry & Glazebrook (2003)	0.67	0.86	1.76
Chabrier (2003)	0.59	0.75	1.42

for the fractional loss, R, of stellar material due to winds and supernovae. Only with this correction ($R = 0.28$ for a Salpeter IMF), does the present-day value represent the integral of the past star formation.

Figure 18 shows the K-band luminosity and derived stellar mass function for galaxies in the 2dF redshift survey from the analysis of [51]. K-band measures were obtained by correlation with the $K < 13.0$ catalog obtained by the 2MASS survey.

The integrated stellar density, corrected for stellar mass loss, is [51]:

$$\Omega_{\rm stars}\, h = 0.0027 \pm 0.00027 \tag{19}$$

for a Salpeter IMF, a value very similar to that derived independently by [93]. By comparison the local mass fraction in neutral HI + He I gas is:

$$\Omega_{\rm gas}\, h = 0.00078 \pm 0.00016 \tag{20}$$

Thus only 5% of all baryons are in stars with the bulk in ionized gas.

3.3 Diagnostics of Star Formation in Galaxies

When significant redshift surveys became possible at intermediate and high redshift through the advent of multi-object spectrographs, so it became possible to consider various probes of the star formation rate (SFR) at different epochs. As in the formalism for calculating the integrated luminosity density, $\rho_{\rm L}$, per comoving Mpc^3, so for a given population various diagnostics of ongoing star formation can yield an equivalent global star formation rate $\rho_{\rm SFR}$ in units of $\mathrm{M_\odot\,yr^{-1}\,Mpc^{-3}}$.

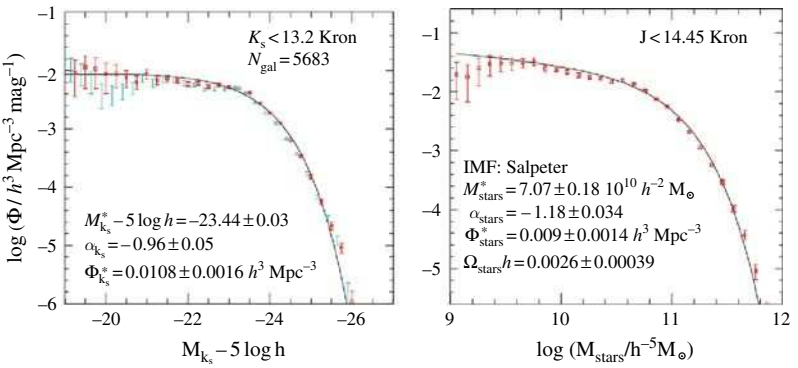

Fig. 18. (**Left**) Rest-frame K-band luminosity function derived from the combination of redshifts from the 2dF survey with photometry from 2MASS Cole et al. (2001) [51]; Schechter parameter fits are shown. (**Right**) Derived stellar mass function assuming a Salpeter IMF corrected for lost material assuming $R = 0.28$ (see text)

Such integrated measures average over a whole host of important details, such as differences in evolutionary behavior between luminous and subluminous galaxies and, of course, morphology. Moreover, in any survey at high redshift, only a portion of the population is rendered visible so uncertain corrections must be made to compare results at different epochs. The importance of the cosmic star formation history, i.e. $\rho_{\rm SFR}(z)$, is that it displays, in a simple manner, the epoch and duration of galaxy growth. By integrating the function, one should recover the present stellar density (Sect. 3.5).

There are various probes of star formation in galaxies, each with its advantages and drawbacks. Not only is there no single "best" method to gauge the current star formation rate of a chosen galaxy, but as each probe samples the effect of young stars in different initial mass ranges, each averages the star formation rate over a different time interval. If, as is often the case in the most energetic sources, the star formation is erratic or burst-like, one would not expect different diagnostics to give the same measure of the instantaneous SFR even for the same galaxies.

Four diagnostics are in common use (see review by [134]).

- The rest-frame ultraviolet continuum ($\lambda\lambda \simeq 1250$–$1500$ Å) has the advantage of being directly connected to well-understood high mass ($> 5\,{\rm M}_\odot$) main sequence stars. Large datasets are available for high redshift star-forming galaxies, including some to $z \simeq 6$. Via the GALEX satellite and earlier balloon-borne experiments, local data is also available. The disadvantage of this diagnostic lies in the uncertain (and significant) corrections necessary for dust extinction and a modest sensitivity to the assumed initial mass function. Obscured populations are completely missed in UV samples. Kennicutt suggests the following calibration for the UV luminosity:

$$SFR({\rm M}_\odot\ {\rm yr}^{-1}) = 1.4\,10^{-28} L_{\rm v}({\rm ergs\,s^{-1}\,Hz^{-1}}) \tag{21}$$

- Nebular emission lines such as Hα and [O II] are also available for a range of redshifts ($z < 2.5$), for example as a natural by-product of faint redshift surveys. Gas clouds are photo-ionized by very massive ($> 10\,{\rm M}_\odot$) stars. Dust extinction can often be evaluated from higher order Balmer lines under various radiative assumptions depending on the escape fraction of ionizing photons. The sensitivity to the initial mass function is strong.

$$SFR({\rm M}_\odot{\rm yr}^{-1}) = 7.9\,10^{-42} L(H\alpha)({\rm ergs\,s^{-1}}) \tag{22}$$

$$SFR({\rm M}_\odot{\rm yr}^{-1}) = 1.4 \pm 0.410^{-41} L({\rm OII})({\rm ergs\,s^{-1}}) \tag{23}$$

- Far infrared emission (10–300 μm) arises from dust heated by young stars. It is clearly only a tracer in the most dusty systems and thus acts as a valuable complementary probe to the UV continuum. As we have seen in Sect. 2, luminous far infrared galaxies are also seen to high redshift.

However, not all dust heating is due to young stars and the bolometric far infrared flux, $L_{\rm FIR}$, is needed for an accurate measurement.

$$SFR(\mathrm{M}_\odot \mathrm{yr}^{-1}) = 4.5\,10^{-44} L_{\rm FIR}(\mathrm{ergs\,s}^{-1}) \tag{24}$$

- Radio emission, e.g. at 1.4 GHz, is thought to arise from synchrotron emission generated by relativistic electrons accelerated by supernova remnants following the rapid evolution of the most massive stars. Its great advantage is that it offers a dust-free measure of the recent SFR. Current radio surveys do not have the sensitivity to see emission beyond $z \simeq 1$, so its promise has yet to be fully explored. This process is also the least well-understood and calibrated. Reference [241] discuss this point in some detail and conclude:

$$SFR(\mathrm{M}_\odot \mathrm{yr}^{-1}) = 1.1\,10^{-28} L_{1.4}(\mathrm{ergs\,s}^{-1}) \tag{25}$$

for bursts of duration $> 100\,\mathrm{Myr}$.

The question of the time-dependent nature of the SFR is an important point ([240, 241]). For an instantaneous burst of star formation, Fig. 19(left) shows the "response" of the various diagnostics. Clearly if the SF is erratic on 0.01–0.1 Gyr timescales, each will provide a different sensitivity. Reference [240] compared UV and Hα diagnostics for a large sample of nearby galaxies and found a scatter beyond that expected from the effects of dust extinction or observational error, presumably from this effect (Fig. 19(right)).

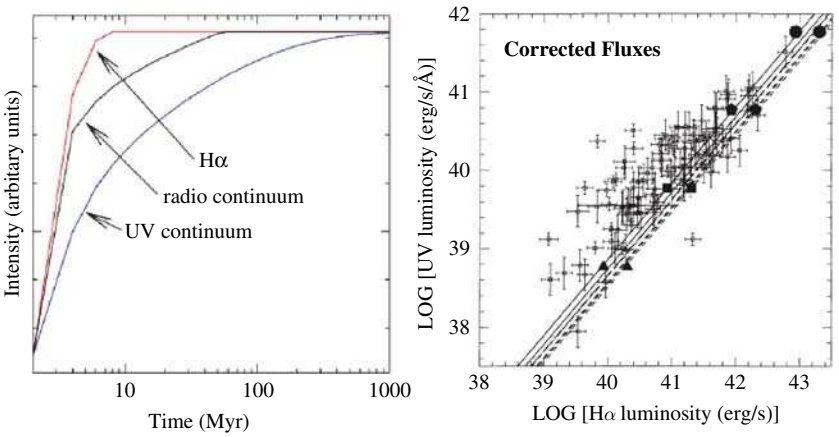

Fig. 19. Time dependence of various diagnostics of star formation in galaxies. (**Left**) sensitivities for a single burst of star formation. (**Right**) scatter in the UV and Hα galaxy luminosities in the local survey of Sullivan et al. (2000) [240]; *lines* represent model predictions for various (*constant*) star formation rates and metallicities. It is claimed that some fraction of local galaxies must undergo erratic periods of star formation in order to account for the offset and scatter

In addition to the initial mass function (already discussed), a key uncertainty affecting the UV diagnostic is the selective dust extinction law. Over the wavelength range $0.3 < \lambda < 1\,\mu$m, differences between laws deduced for the Milky Way, the Magellan clouds and local starburst galaxies ([42]) are quite modest. Significant differences occur around the 2200 Å feature (dominant in the Milky Way but absence in Calzetti's formula) and shortward of 2000 Å where the various formulae differ by ± 2 mags in $A(\lambda)/E(B-V)$.

3.4 Cosmic Star Formation – Observations

Early compilations of the cosmic star formation history followed the field redshift surveys of [76, 148] and the abundance of U-band drop outs in the early deep HST data ([155]). The pioneering papers in this regard include [82, 148, 155, 156].

References [110] and [111] have undertaken a valuable recent compilation, standardizing all measures to the same initial mass function, cosmology and extinction law. They have also integrated the various luminosity functions for each diagnostic in a self-consistent manner (except at very high redshift). Accordingly, their articles give us a valuable summary of the state of the art.

Figure 20 summarizes their findings. Although at first sight somewhat confusing, some clear trends are evident including a systematic increase in star formation rate per unit volume out to $z \simeq 1$ which is close to ([110]):

$$\rho_{\rm SFR}(z) \propto (1+z)^{3.1} \tag{26}$$

A more elaborate formulate is fitted in [111].

There is a broad peak somewhere in the region $2 < z < 4$ where the UV data is consistently an underestimate and the growing samples of sub-mm galaxies are valuable. The dispersion here is only a factor of ± 2 or so, which is a considerable improvement on earlier work. We will return to the question of a possible decline in the cosmic SFR beyond $z \simeq 3$–4 in later sections.

In their recent update, [111] also parametrically fit the resulting $\rho_{\rm SFR}(z)$ in two further redshift sections, beyond $z \simeq 1$, and they use this to predict the growth of the absolute stellar mass density, ρ_*, via integration (Fig. 21). Concentrating, for now, on the reproduction of the *present day* mass density [51], the agreement is remarkably good.

Although in detail the result depends on an assumed initial mass function and the vexing question of whether extinction might be luminosity-dependent, this is an important result in two respects: firstly, as an absolute comparison it confirms that most of the star formation necessary to explain the presently-observed stellar mass has already been detected through various complementary surveys. Secondly, the study allows us to predict fairly precisely the epoch by which time half the present stellar mass was in place; this is $z_{\frac{1}{2}} = 2.0 \pm 0.2$. In Sect. 4 we will discuss this conclusion further attempting to verify it by measuring stellar masses of distant galaxies directly.

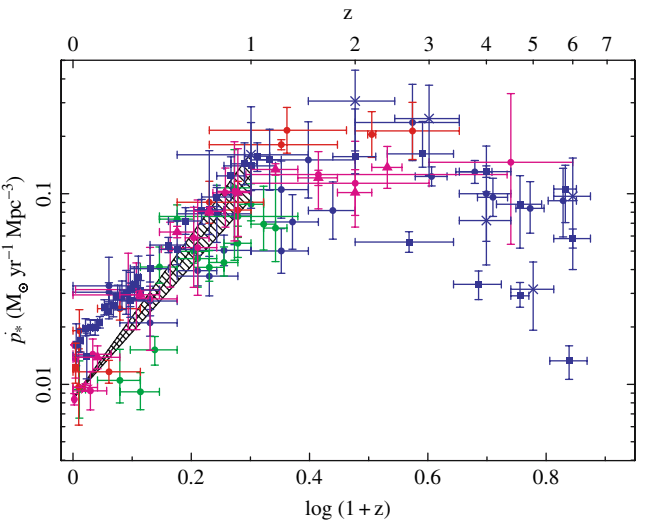

Fig. 20. Recent compilations of the cosmic star formation history. *Circles* are data from Hopkins (2004) [110] color-coded by method: blue: UV, green: [O II], red: Hα/β, magenta: non-optical including sub-mm and radio. New data from Hopkins & Beacom (2006) [111], represented by *various triangles*, stars and squares, include Spitzer FIR measures (*magenta triangles*). The *solid lines* represent a range of the best fitting parametric form for $z < 1$

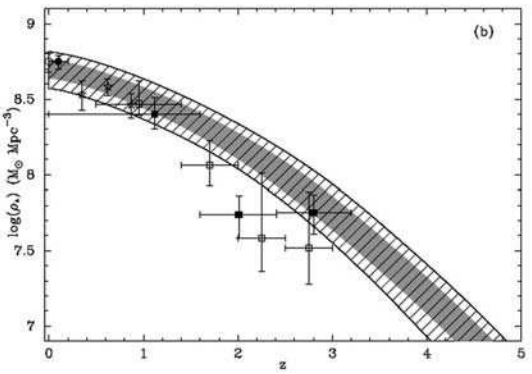

Fig. 21. Growth of stellar mass density, ρ_*, with redshift obtained by direct integration of a parametric fit to the cosmic star formation history deduced by Hopkins & Beacom (2006) [111] see Fig. 20). The integration accurately reproduces the local stellar mass density observed by Cole et al. (2001) [51] and suggests half the present density was in place at $z = 2.0 \pm 0.2$

3.5 Cosmic Star Formation – Theory

As we have discussed, semi-analytical models have had a hard time reproducing and predicting the cosmic star formation history. Amusingly, as the data has improved, the models have largely done a "catch-up" job [9, 10]. To their credit, while many observers were still convinced galaxies formed the bulk of their stars in a narrow time interval (the "primeval galaxy" hypothesis), CDM theorists were the first to suggest the extended star formation histories now seen in Fig. 20.

A particular challenge seems to be that of reproducing the abundance of energetic sub-mm sources whose star formation rates exceed 100–200 $\mathrm{M}_{\odot}\,\mathrm{yr}^{-1}$. Reference [11] have suggested it may require a combination of quiescent and burst modes of star formation, the former involving an initial mass function steepened towards high mass stars. Although there is much freedom in the semi-analytical models, recent models suggest $z_{\frac{1}{2}} \simeq 1.3$. By contrast, for the same cosmological models, hydrodynamical simulations ([172]) predict much earlier star formation, consistent with $z_{\frac{1}{2}} \simeq 2.0$–$2.5$.

The flexibility of these models is considerable so my personal view is that not much can be learned from these comparisons either way. It is more instructive to compare galaxy masses at various epochs with theoretical predictions. Although we are still some ways from doing this in a manner that includes both baryonic and dark components, progress is already promising and will be reviewed in Sect. 4.

3.6 Unifying the Various High Redshift Populations

Integrating the various star-forming populations at high redshift to produce Fig. 20 avoids the important question of the physical relevance and roles of the seemingly-diverse categories of high redshift galaxies. In the previous lecture (Sect. 2), I introduced three broad categories: the Lyman break (LBG), sub-mm and passively-evolving sources (DRGs) which co-exist over $1 < z < 3$. What is the relationship between these objects?

As the datasets on each has improved, we have secured important physical variables including masses, star formation rates and ages. We can thus begin to understand not only their relative contributions to the SFR at a given epoch, but the degree of overlap among the various populations. Several recent articles have begun to evaluate the connection between these various categories [180, 189].

A particularly valuable measure is the clustering scale, r_0, for each population, as defined in Sect. 1.3. This is closely linked to the halo mass according to CDM and thus sets a marker for connecting populations observed at different epochs. Reference [3] demonstrated the strong clustering, $r_0 \simeq 3.8\,\mathrm{Mpc}$, of luminous LBGs at $z \simeq 3$. Reference [9] claimed this was consistent with the progenitor halos of present-day massive ellipticals. The key to the physical nature of LBGs depends the origin of their intense star formation. At $z \simeq 3$, the

bright end of the UV luminosity function is $\simeq 1.5$ mags brighter than its local equivalent; the mean SFR is $45\,M_{\odot}\,yr^{-1}$. Is this due to prolonged activity, consistent with the build up of the bulk of stars which reside in present-day massive ellipticals, or is it a temporary phase due to merger-induced starbursts ([219])?

References [209] and [210] investigated the stellar population and stacked spectra of a large sample of $z \simeq 3$ LBGs and find younger systems with intense SFRs are dustier with weaker Lyα emission while outflows (or "superwinds") are present in virtually all (Fig. 22 (left)). For the young LBGs, a brief period of elevated star formation seems to coincide with a large dust opacity hinting at a possible overlap with the sub-mm sources. During this rapid phase, gas and dust is depleted by outflows leading to eventually to a longer, more quiescent phase during which time the bulk of the stellar mass is assembled.

If young dusty LBGs with SFRs $\simeq 300\,M_{\odot}\,yr^{-1}$ represent a transient phase, we might expect sub-mm sources to simply be a yet rarer, more extreme version of the same phenomenon. The key to testing this connection lies in the relative clustering scales of the two populations (Fig. 22 (right)). Reference [24] find sub-mm galaxies are indeed more strongly clustered than the average LBGs, albeit with some uncertainty given the much smaller sample size.

Turning to the passively-evolving sources, although [150] provides a valuable review of the territory, the observational situation is rapidly changing. For many years, CDM theorists predicted a fast decline with redshift in the abundance of red, quiescent sources. Using a large sample of

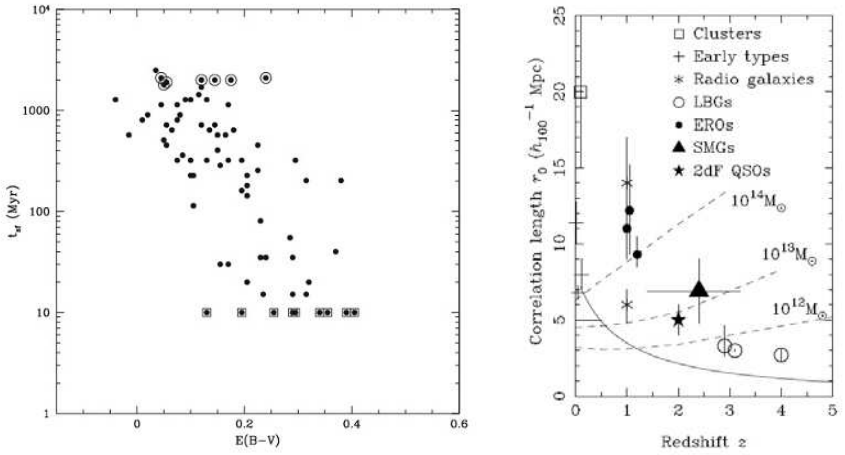

Fig. 22. Connecting Lyman break and sub-mm sources. (**Left**) Correlation between the mean age (for a constant SFR) and reddening for a sample of $z \simeq 3$ LBGs from the analysis of Shapley et al. (2001) [209]. The youngest LGBs seem to occupy a brief dusty phase limited eventually by the effect of powerful galaxy-scale outflows. (**Right**) The LBG-submm connection can be tested through reliable measures of their relative clustering scales, r_0 Blain et al. (2004) [24]

photometrically-selected sources in the COMBO-17 survey, [15] claimed to see this decline in abundance by witnessing a near-constant luminosity density in red sources to $z \simeq 1$ (Fig. 23 (left)). The key point to understand here is that a passively-evolving galaxy fades in luminosity so that the red luminosity density should *increase* with redshift unless the population is growing. Bell et al. surmises the abundance of red galaxies was 3 times less at $z \simeq 1$ as predicted in early semi-analytical models ([124]).

By contrast, the Gemini Deep Deep Survey ([101]) finds numerous examples of massive red galaxies with $z > 1$ in seeming contradiction with the decline predicted by CDM supported by [15]. Of particular significance is the detailed spectroscopic analysis of 20 red galaxies with $z \simeq 1.5$ ([151], Fig. 23 (right)) whose inferred ages are 1.2–2.3 Gyr implying most massive red galaxies formed at least as early as $z \simeq 2.5$–3 with SFRs of order 300–500 $M_\odot\,yr^{-1}$. Could the most massive red galaxies at $z \simeq 1.5$ then be the descendants of the sub-mm population? One caveat is that not all the stars whose ages have been determined by McCarthy et al. need necessarily have resided in single galaxies at earlier times. The key question relates to the reliability of the abundance of early massive red systems. Using a new color-selection technique, [139] suggest the space density of quiescent systems with stellar mass $> 10^{11}\,M_\odot$ at $z \simeq 1.5$–2 is only 20% of its present value.

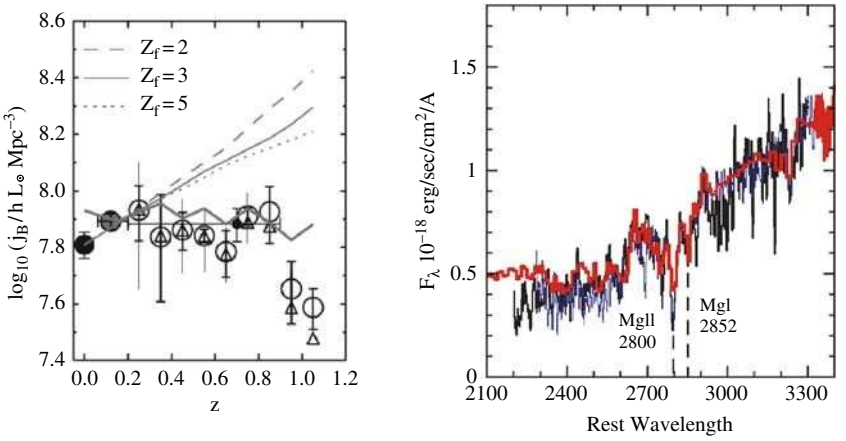

Fig. 23. Left: The rest-frame blue luminosity density of "red sequence" galaxies as a function of redshift from the COMBO-17 analysis of Bell et al. (2004) [15]. Since such systems should brighten in the past the near-constancy of this density implies 3 times fewer red systems exist at $z \simeq 1$ than locally – as expected by standard CDM models. **Right**: Composite spectra of red galaxies with $1.3 < z < 1.4$ (**blue**), $1.6 < z < 1.9$ (**black**) from the Gemini Deep Deep Survey (McCarthy et al. 2004 [151]). A stellar synthesis model with age 2 Gyr is overlaid in red. The analysis suggests the most massive red galaxies with $z \simeq 5$ formed with spectacular SFRs at redshifts $z \simeq 2.4$–3.3

As we will see in Sect. 4, the key to resolving the apparent discrepancy between the declining red luminosity density of Bell et al. and the presence of massive red galaxies at $z \simeq 1$–1.5, lies in the mass-dependence of stellar assembly ([248]).

Finally, a new color-selection has been proposed to uniformly select all galaxies lying in the strategically-interesting redshift range $1.4 < z < 2.5$. Reference [56] have proposed the "BzK" technique, combining $(z - K)$ and $(B - z)$ to locate both star-forming and passive galaxies with $z > 1.4$; such systems are termed "sBzK" and "pBzK" galaxies respectively. Reference [189] claim there is little distinction between the star-forming sBzK and Lyman break galaxies – both contribute similarly to the star formation density over $1.4 < z < 2.6$ and the overlap fractions are at least 60–80%.

More interestingly, both [189] and [139] suggest significant overlap between the passive and actively star-forming populations. Kong et al. find the angular clustering is similar and Reddy et al. find the stellar mass distributions overlap.

3.7 Lecture Summary

Clearly multi-wavelength data is leading to a revolution in tracking the history of star formation in the Universe. Because of the vagaries of the stellar initial mass function, dust extinction and selection biases, we need multiple probes of star formation in galaxies.

The result of the labors of many groups is a good understanding of the comoving density of star formation since a redshift $z \simeq 3$. Surprisingly, the trends observed can account with reasonable precision for the stellar mass density observed today. The implication of this result is that half the stars we see today were in place by a redshift $z \simeq 2$.

What, then, are we to make of the diversity of galaxies we observe during the redshift range ($z \simeq 2$–3) of maximum growth? Through detailed studies some connections are now being made between both UV-emitting Lyman break galaxies and dust-ridden sub-mm sources.

More confusion reigns in understanding the role and decline with redshift in the contribution of passively-evolving red galaxies. Some observers claim a dramatic decline in their abundance whereas others demonstrate clear evidence for the presence of a significant population of old, massive galaxies at $z \simeq 1.5$. We will return to this enigma in Sect. 4.

4 Stellar Mass Assembly

4.1 Motivation

Although we may be able to account for the present stellar mass density by integrating the comoving star formation history (Sect. 3), this represents only

a small step towards understanding the history of galaxy assembly. A major limitation is the fact that the star formation density averages over a range of physical situations and luminosities; we are missing a whole load of important physics. As we have seen, a single value for the stellar mass density, e.g. $\rho_*(z = 2)$, is useful when considered as a global quantity (e.g. compared to an equivalent estimate at $z = 0$), but it does not describe whether the observed star formation is steady or burst-like in nature, or even whether the bulk of activity within a given volume arises from a large number of feeble sources or a small number of intense objects. Such details matter if we are trying to construct a clear picture of how galaxies assemble.

Of course we could extend our study of time-dependent star formation to determine the *distribution functions* of star formation at various epochs (e.g. the UV continuum, Hα or sub-mm luminosity functions), but making the integration check only at $z = 0$ is second-best to measuring the *assembled mass* and its distribution function at various redshifts. This would allow us to directly witness the growth rate of galaxies of various masses at various times and, in some sense, is a more profound measurement, closer to theoretical predictions.

Ideally we would like to measure both baryonic and non-baryonic masses for large numbers of galaxies but, at present at least, we can only make dynamical or lensing-based total mass estimates for specific types of distant galaxy and crude estimates for the gaseous component. The bulk of the progress made in the last few years has followed attempts to measure *stellar masses* for large populations of galaxies. We will review their achievements in this lecture.

4.2 Methods for Estimating Galaxy Masses

What are the options available for estimating galaxy masses of any kind at intermediate to high redshift? Basically, we can think of three useful methods.

- Dynamical methods based on resolved rotation curves for recognisable disk systems ([257, 258]) or stellar velocity dispersions for pressure-supported spheroidals ([247, 254]). These methods only apply for systems known or assumed a priori to have a particular form of velocity field. Interesting constraints are now available for a few hundred galaxies in the field and in clusters out to redshifts of $z \simeq 1.3$. Key issues relate to biases associated with preferential selection of systems with "regular" appearance and how to interpret mass dynamically-derived over a limited physical scale (c.f. [47]). In the absence of resolved rotation curves, sometimes emission linewidths are considered a satisfactory proxy ([174]).
- Gravitational lensing offers the cleanest probe of the total mass distribution but, as a geometric method, is restricted in its application to compact, dense lenses (basically, spheroidals) occupying cosmic volumes typically half way to those probed by faint star-forming background field galaxies. In practice this means $z_{\rm lens} < 1$. Even so, by sifting through spectra of

luminous red galaxies in the SDSS survey and locating cases where an emission line from a background lensed galaxy enters the spectroscopic fiber, [28] have identified a new and large sample of Einstein rings enabling us to gain valuable insight into the relative distribution of dark and visible mass over $0 < z < 1$.

- Stellar masses derived from near-infrared photometry represents the most popular technique in use at the current time. The idea has its origins in the recognition [34, 122] that the rest-frame K-band luminosity of a galaxy is less affected by recent star formation than its optical equivalent (Fig. 16), and thus can act as a closer proxy to the well-established stellar population. A procedure for fitting the rest-frame optical-infrared spectral energy distribution of a distant galaxy, deriving a stellar mass/light ratio ($M/L_{\rm K}$) and hence the stellar mass if the redshift is known, was introduced by [33]. The popularity of the technique follows from the fact it can easily be applied to large catalogs of galaxies in panoramic imaging surveys and extended to very high redshift if IRAC photometry is available. The main difficulties relate to the poor precision of the method, particularly if the same photometric data is being used to estimate the redshift [39], plus degeneracies arising from poor knowledge of the past star formation history ([211, 252]).

For many galaxies, an important and usually neglected component is the mass locked up in both ionized and cool gas. In nearby systems amenable to study of hot ionized gas (from nebular emission lines) and its usually dominant cooler neutral component (probed by 21 cm studies), as much as 20% of the baryonic mass of a luminous star-forming galaxy can be found in this form ([266]). At present, it is not possible to routinely use radio techniques to reliably estimate gaseous masses of distant galaxies although approximate gas masses have been derived assuming the projected surface density of nebular emission correlates with the gas mass within some measured physical scale [74].

4.3 Results: Regular Galaxies $0 < z < 1.5$

Because of the simplicity of their stellar populations, velocity fields and the lack of confusing gaseous components, rather more is known about the mass assembly history of ellipticals than for spirals. Concerning ellipticals, one of the key challenges is separating the *age of the stars* from the *age of the assembled mass*.

The Fundamental Plane [17, 61, 65, 118] represents an empirical correlation between the dynamical mass (via the central stellar velocity dispersion σ_0), the effective radius ($R_{\rm E}$) and light distribution (via the enclosed surface brightness $\mu_{\rm E}$) for ellipticals, viz:

$$\log R_{\rm E} = a \log \sigma_0 + b\,\mu_{\rm E} + c \tag{27}$$

For example, with σ in km sec^{-1}, $R_{\rm E}$ in kpc and $\mu_{\rm E}$ in mags arcsec^{-2} in the B band, $a = 1.25$, $b = 0.32$, $c = -8.970$ for $h = 0.65$.

These observables define an effective dynamical mass $M_{\rm E} \propto \sigma^2 R_{\rm E}/G$ which correlates closely with that deduced from lensing ([249]). Deviations from the local FP at a given redshift z can be used to deduce the change in mass/light ratio $\Delta \log(M/L)$.

The most comprehensive studies of the evolving FP come from two independent and consistent studies of field spheroidals to $z \simeq 1$ ([248, 251]). The evolution in mass/light ratio $\Delta \log(M/L)$ deduced from the GOODS-N survey of [248] is shown in Fig. 24. These authors find as little as 1–3% by stellar mass of the present-day population in massive ($> 10^{11.5}\,{\rm M}_\odot$) galaxies formed since $z = 1.2$, whereas for low mass systems ($< 10^{11}\,{\rm M}_\odot$) the growth fraction is 20–40%. This result, confirmed independently by [251], is an important illustration of the mass-dependent growth in galaxies with the most massive systems shutting off earliest.

Of course, one should not confuse the *age of stars*, as probed by the FP, with the *age of the assembled mass*. Reference [253] has argued that if spheroidals preferentially merge with similar gas-poor systems (a process called "dry mergers") the FP analyses could well indicate early eras of major star formation even though the bulk of the assembled mass in individual

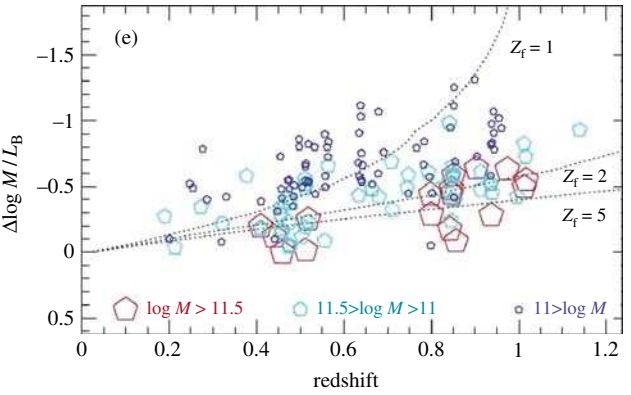

Fig. 24. Change in B-band mass/light ratio with redshift deduced from the Keck dynamical survey of over 150 field spheroidals in GOODS-N (Treu et al. 2005 [248]). The *plot* shows the change in mass/light ratio for galaxies of different effective dynamical masses (*red: massive, cyan: intermediate, blue: low mass*). Curves illustrate the change in mass/light ratio expected as a function of redshift for galaxies assembled monolithically with simply evolving stellar populations since a redshift of formation z_f. Clearly the stars in the most massive galaxies largely formed at high redshift whereas assembly continued apace in lower mass systems

systems occurred at $z < 1$[4]. References [16] and [246] have cataloged individual cases of dry mergers in both field and cluster samples, respectively. Their occurrence is not in dispute; however, only via morphological or other measures of the global mass assembly can the role of dry mergers as a major feature of galaxy assembly be addressed.

Related insight into this problem arises from the relative distributions of baryonic and dark matter deduced from the combination of lensing and stellar dynamics for the recently-located SDSS-selected Einstein rings ([249], Fig. 25). Although it might be thought that lensing preferentially selects the most massive and compact sources, Treu et al. compare the FP of such lenses with those in the larger field sample (Fig. 25) and deduce otherwise. An important result from the study of the first set of such remarkable lenses is how well the total mass profile can be represented by an isothermal form with mass tracing light, viz:

$$\rho_{\rm tot} = \rho \left(\frac{r}{r_0} \right)^{-\gamma} \tag{28}$$

Even over $0 < z < 1$, the mass slope γ is constant at 2.0 to within 2% precision indicating rather precise collisional coupling of dark matter and gas.

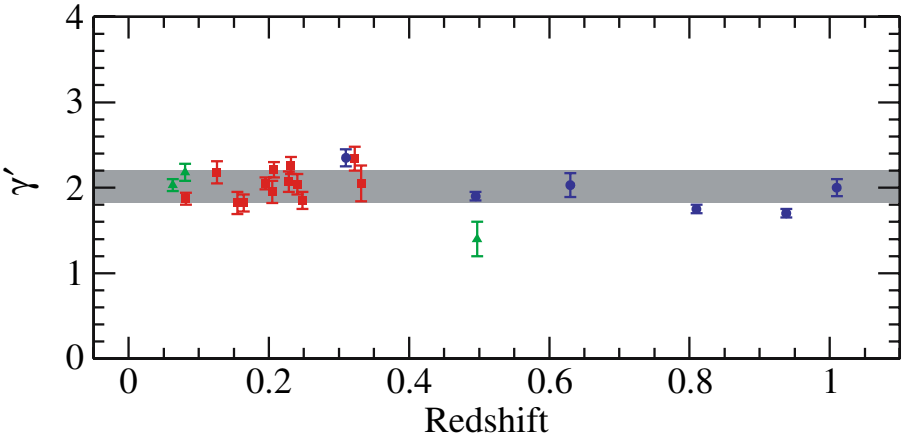

Fig. 25. Distribution of the gradient *gamma* in the total mass profile with redshift as deduced from the combination of stellar dynamics and lensing for ellipticals in the SLAC survey (Treu et al., in prep). The remarkable constancy of the profile slope indicates massive relaxed systems were in place at $z \simeq 1$ and that "dry merging" cannot be a prominent feature of the assembly history of large galaxies

[4] The reason observers have gone somewhat out of their way to consider such complicated scenarios is because late assembly of massive spheroidals was, until recently, a fundamental tenet of the CDM hierarchy to be salvaged at all costs ([124]).

Sadly, far less is known about the mass assembly history of regular spirals. The disk scaling law equivalent to the FP for ellipticals, the Tully-Fisher relation ([250]) which links rotational velocity to luminosity gives ambiguous information without additional input. Modest evolution in the TF relationship was deduced from the pioneering Keck study of $z \simeq 1$ spirals by [258] but this could amount to $\simeq 0.6$ mag of B-band luminosity brightening in sources of a fixed rotational velocity to $z \simeq 1$, or more enhancement if masses were reduced.

Additional variables capable of breaking the degeneracy between dynamical mass and luminosity include physical size and stellar mass. Reference [149] examined the *size-luminosity* relation for several hundred disks to $z \simeq 1$ in a HST-classified redshift survey sample (see [202] for an update) and found no significant growth for the largest systems. Reference [47] correlated stellar and dynamical masses for $\simeq 100$ spirals with resolved dynamics in the context of a simple halo formulation. Although their deduced halo masses must be highly uncertain, they likewise deduced that growth must be modest since $z \simeq 1$, occurring in a self-similar fashion for the baryonic and dark components.

4.4 Stellar Masses from Multi-Color Photometry

References [38] and [39] give a good summary and critical analysis of the now well-established practice of estimating stellar masses from multi-color optical-infrared photometry. Figure 26 gives a practical illustration of the technique where it can be seen that even for low z galaxies with good photometry, the precision in mass is only $\simeq \pm 0.1$–0.2 dex. In most cases, even random uncertainties are at the ± 0.2–0.3 dex level and systematic errors are likely to be much higher.

Since analysing stellar mass functions is now a major industry in the community, it is worth spending some time considering the possible pitfalls. A significant fraction of the large datasets being used are purely photometric, with both redshifts and stellar masses being simultaneously deduced from multi-color photometry [15, 68, 88]. Few large surveys have extensive spectroscopy from which to check that this procedure works.

Reference [39] address this important question in the context of their extensive DEEP2 spectroscopic sample by artificially "switching off" the knowledge of the spectroscopic z: how does the stellar mass deduced when simultaneously deriving the photometric redshift from the same data compare with that derived when the spectroscopic value is externally input? Figure 27 illustrates that a significant component of error is one beyond that expected solely from the error in the derived redshift (as measured from the $z_{\rm spec}$ vs. $z_{\rm photo}$ comparison).

A second restriction in many stellar mass determinations is the absence of any near-infrared photometry. This may seem surprising given the classic papers [33, 122] stressed its key role. However, panoramic near-infrared imaging is much more expensive in telescope time as few observatories, until recently,

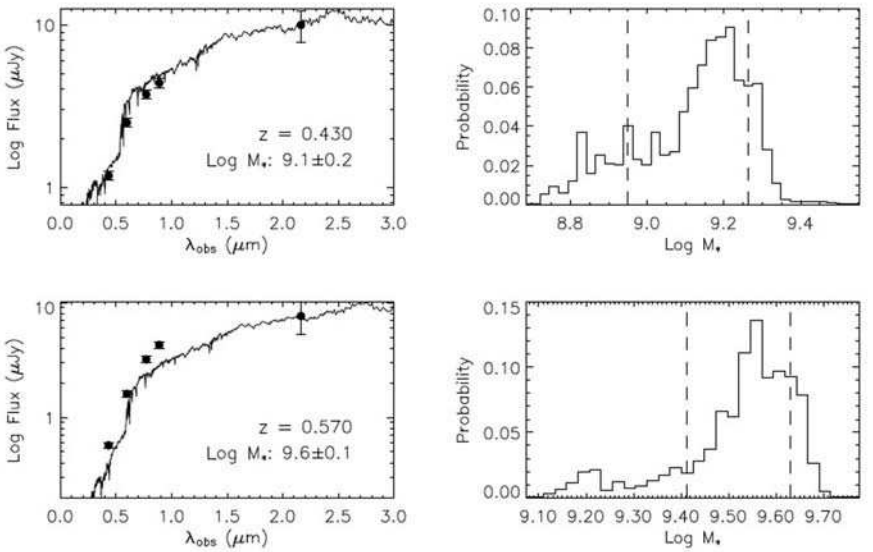

Fig. 26. Deriving stellar masses for galaxies of known redshift via multi-color optical photometry (after Bundy et al. 2006 [39]). (**Left**) Rest-frame spectral energy distribution of two DEEP2 galaxies of known redshift from broad-band $BRIK$ photometry with a fit deduced by fitting from a stellar synthesis library. The fit yields the most likely mass/light ratio of the observed population. (**Right**) Likelihood distribution of the stellar mass for the same two galaxies derived by Bayesian analysis. *Dashed lines* indicate the eventual range of permitted solutions

had large format infrared cameras. Inevitably, some groups have attempted to get by without the near-infrared data.

Figure 28 illustrates how the precision degrades when the K band data is dropped in the stellar mass fit [39][5]. Although the systematic error is contained, the noise increases significantly for redshifts $z > 0.7$ where the optical photometry fails to adequately sample the older, lower mass stars.

Reference [211] explore a further uncertainty, which is particularly germane to the analysis of their $z \simeq 2$ sources. Using IRAC to represent the rest-frame infrared at these redshifts, they consider the role that recent bursts of star formation might have on the deduced stellar masses. Although bursts predominantly affect short-wavelength luminosities, one might imagine little effect at the longer wavelengths (c.f. [122]). Shapley et al. consider a wider range of star formation histories and show that the derived stellar mass depends *less strongly* on the long wavelength luminosity L (4.5 μm) than expected. The scatter is consistent with variations in mass/light ratios of ×15 (Fig. 29).

[5] A similar analysis was conducted by [211] at $z \simeq 2$, excluding the relevant IRAC data.

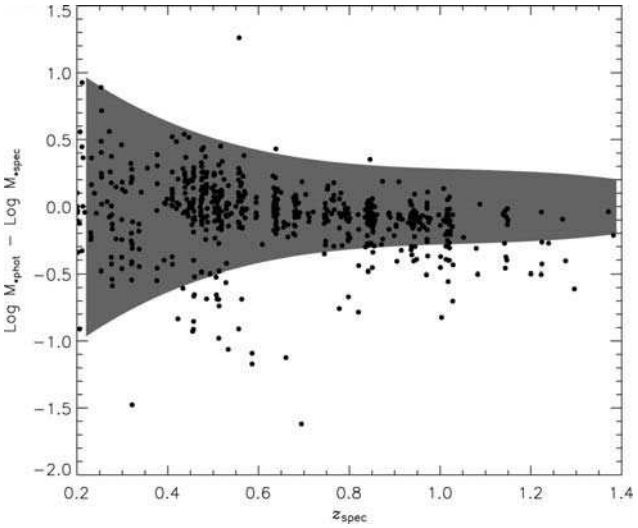

Fig. 27. Error in log stellar mass when the known spectroscopic redshift is ignored (Bundy et al. 2005 [38]). The shaded region defines the expected scatter in mass arising solely from that in the photometric-spectrosopic redshift comparison. Clearly in addition to this component, some catastrophic failures are evident

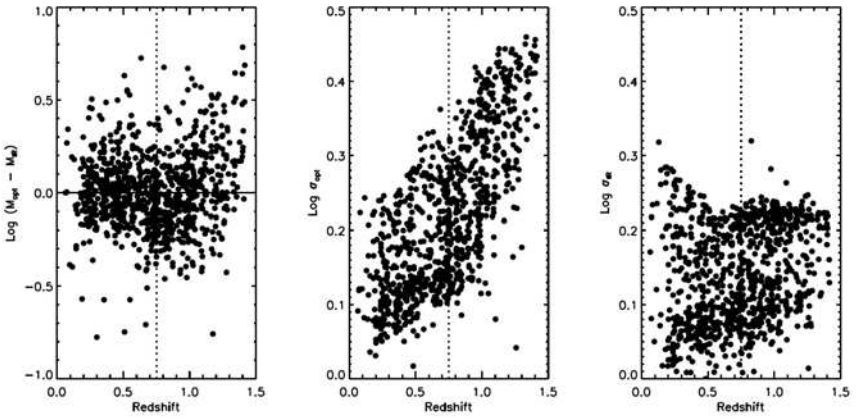

Fig. 28. Effect of deriving stellar masses from optical photometry alone Becuer et al. (2006) [12]. (**Left**) Difference in masses derived using $BRIK$ and BRI photometry. (**Center**) Distribution of uncertainties deduced from Bayesian fitting for the BRI fits. (**Right**) as center for the $BRIK$ fitting. Although the systematic offset is small, the uncertainty in deduced stellar mass increases dramatically for galaxies with $z > 0.7$

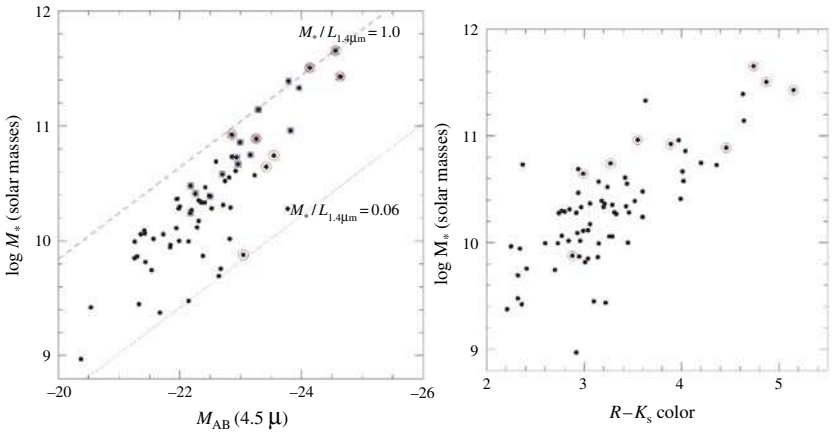

Fig. 29. Left: Stellar masses versus observed IRAC 4.5 μm luminosities for the sample of $z \simeq 2$ galaxies studied by Shapley et al. (2005) [211]. Allowing a wide range of star formation histories illustrates a weaker dependence of mass on long wavelength luminosity than seen in earlier tests at lower redshift based on simpler star formation histories. The range in rest-frame 1.4 μm mass/light ratio is $\simeq \times 15$. (**Right**) The existence of a fairly tight correlation between mass and $R - K$ color demonstrates the importance of considering secondary bursts of star formation in deriving stellar masses

The strong correlation between mass and optical-infrared color emphasizes the importance of secondary activity.

4.5 Results: Stellar Mass Functions $0 < z < 1.5$

Noting the above uncertainties, we now turn to results compiled from the stellar mass distributions of galaxies undertaken in both spectroscopic and photometric surveys. A recent discussion of the various results can be found in [39].

Reference [33] were the first to address the global evolution of stellar mass in this redshift range using a morphological sample of 350 galaxies with spectroscopic redshifts and infrared photometry. The small sample did not permit consideration of detailed stellar mass functions, but the integrated mass density was partitioned in 3 redshift bins for spheroidals, spirals and irregulars. Surprisingly little growth was seen in the overall mass density from $z \simeq 1$ to today; the strongest evolutionary signal seen is a *redistribution* of mass amongst the morphologies dominated by a declining mass density in irregulars with time.

The ratio of the galactic stellar mass $M_\star$ to the current star formation rate SFR, e.g. as deduced spectroscopically or from the rest-frame optical colors, is sometimes termed the *specific star formation rate*, R. This quantity allows us to address the question of whether galaxies have been forming stars

for a significant fraction of the Hubble time, at a rate commensurate with explaining their assembled mass. A low value for R implies a quiescent object whose growth has largely ended; the mean stellar age is quite large. A high value of R implies an active object which has assembled recently. A frequently-used alternative is the "doubling time" – that period over which, at the current SFR, the observed stellar mass would double. This time would be quite short for active objects.

References [49] and [33] and most recently [120], found a surprising trend whereby most massive galaxies over $z \simeq 0.5$–1.5 are quiescent, having presumably formed their stars well before $z \simeq 2$, whereas low mass galaxies remain surprisingly active. The term "downsizing" – a signature of continued growth in lower mass galaxies after that in the high mass galaxies has been completed – was first coined by [49] and has been used rather loosely in the recent literature to imply any signature of anti-hierarchical activity. In particular, it is important to distinguish between *downsizing in star formation activity*, which presumably represents some physical process that permits continued star formation in lower mass systems when that in massive galaxies has concluded, from *downsizing in mass assembly*, a truly "anti-hierarchical" process whereby new mass is being added to lower mass galaxies at later times (see discussion in [39]).

Before trying to understanding in more detail what causes this mass-dependent star formation, it is worth returning to the issue of *dry mergers* raised in Sect. 4.3. A "downsizing" signature was also seen in the growth of spheroidal galaxies as analysed by their location on the evolving Fundamental Plane ([248, 251]). However, [253] argued that while this may reflect older stars in the more massive galaxies, the *assembled mass* may still be young if merging preferentially occurs between quiescent objects. The only way to test this hypothesis is to directly measure the growth rate in spheroidal systems.

Using the deeper and more extensive sample of galaxies available in the GOODS fields, [38] produced type-dependent *stellar mass functions* (Fig. 30). Here, for the first time, one can see the morphological evolution detected by Brinchmann & Ellis largely arises via the transformation of intermediate mass ($\simeq 5 \times 10^{10} - 2 \times 10^{11}\,\mathrm{M}_{\odot}$) irregulars and spirals into spheroidals. If merging is responsible for this transformation, it is predominantly occurring between gas-rich and active systems even at the very highest masses. The bulk of the evolution below $z \simeq 1$ is simply a redistribution of star formation activity, perhaps as a result of mergers or feedback processes.

Interestingly, in Fig. 30 there is almost no change in the total mass function with time above $5 \times 10^{10}\,\mathrm{M}_{\odot}$, suggesting little growth in the mass spheroidals that dominate the high mass end. However, it worth remembering that the two GOODS fields are limited in size ($0.1\,\mathrm{deg}^2$) and suffer from cosmic variance at the 20% level at high redshift increasing to 60% at low redshift [38].

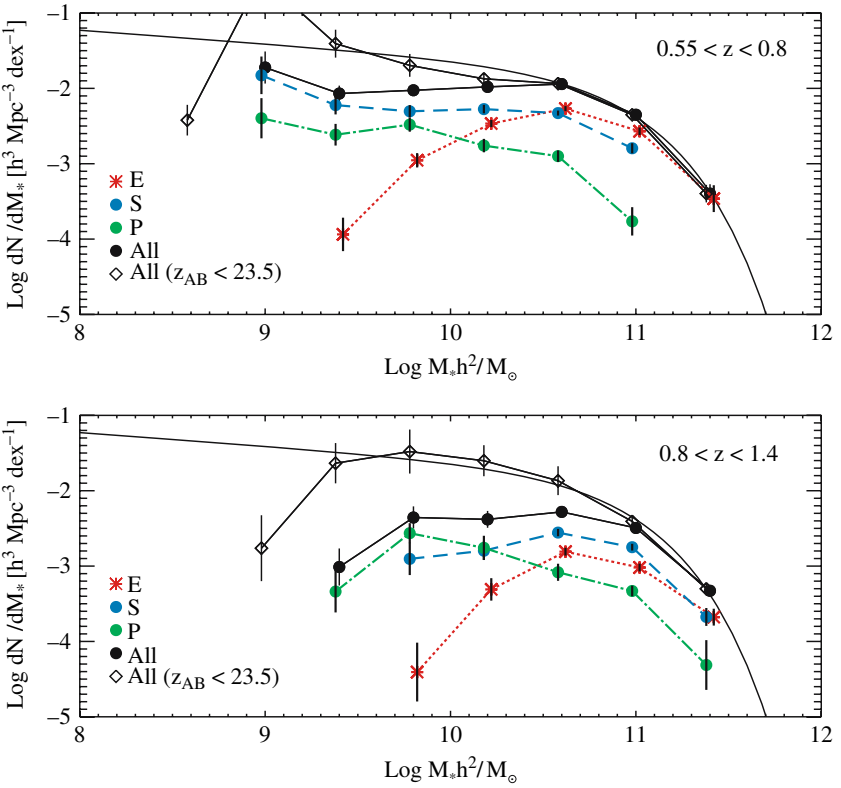

Fig. 30. Stellar mass function arranged by morphology in two redshift bins from the analysis of Bundy et al. (2005) [38] for $h = 1$. The *solid curve* in both plots represents the present day mass function from the 2dF survey (Cole et al. 2001 [51]). Type-dependent mass functions are color-coded with black representing the total. The *solid line* connecting the filled black circles represents the sample with spectroscopic redshifts, the *dotted line* connecting the *open black diamonds* includes masses derived from sources with photometric redshifts

Stellar mass functions for a much larger sample of field galaxies of known redshift have been analyzed by [39] utilizing the combination of extensive spectroscopy and Palomar K-band imaging in four DEEP2 fields (totalling $1.5\,\mathrm{deg}^2$). This sample has the benefit of being much less affected by cosmic variance although, as there is no complete coverage with HST, morphological classifications are not possible. Color bimodality has been analyzed in the DEEP2 sample ([259]) and Bundy et al. use the rest-frame $U - B$ color and spectroscopic [O II] equivalent width to separate quiescent and active galaxies.

Figure 31 shows a direct comparison of the integrated stellar mass functions from this large survey alongside other, less extensive surveys, most of

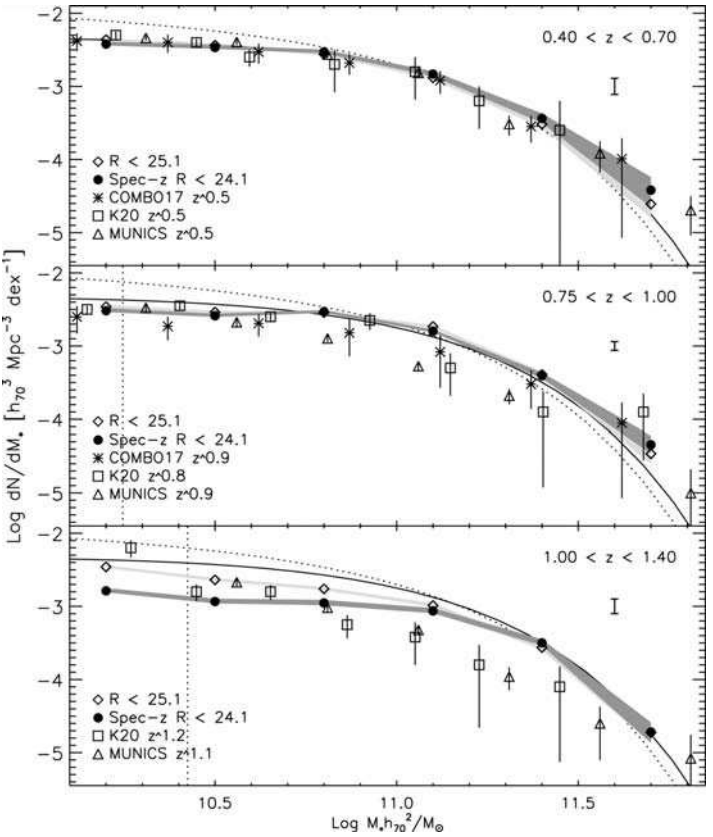

Fig. 31. The evolving stellar mass function from the comparison of Bundy et al. (2006) [39]. The *vertical dashed lines* represent completeness limits for all types

which are based only on photometric redshifts. Although each is variance limited in different ways, one is struck again by the quite modest changes in the abundance of massive galaxies since $z \simeq 1$.

Next we consider (Fig. 32) the stellar mass functions for the quiescent and active star-forming galaxies independently, partitioned according to the rest-frame $U-B$ color. The surprising result here is the existence of a threshold or *quenching mass*, M_Q above which there are no active systems. This is implied independently in Fig. 30 where, within the redshift bin $0.55 < z < 0.8$, there are no spirals or irregulars with mass $> 10^{11}\,\mathrm{M}_\odot$. Interestingly, there is also a modest downward transition in M_Q with time.

A remarkably consistent picture emerges from these studies. Over the redshift range $0 < z < 1.5$, the stellar mass function has changed very little at the high mass end. Substantial growth, in terms of a mass doubling on $\simeq$ 5–8 Gyr timescale, is only possible for galaxies with stellar masses below $10^{10}\,\mathrm{M}_\odot$. Above this mass, the basic evolutionary signal is a *quenching* of star

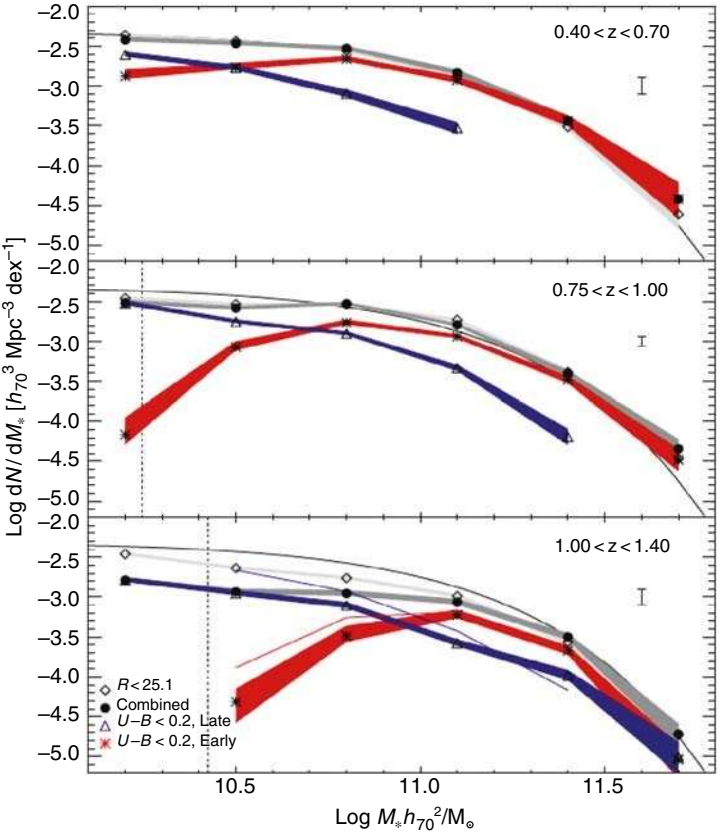

Fig. 32. The evolving stellar mass function split into quiescent (red) and active (blue) sources from the survey of Bundy et al. (2006) [39]. Uncertainties arising from counting statistics and errors in the stellar masses are indicated by the associated shading. The *thin solid line* represents the local 2dF function (Cole et al. 2001 [51]) and the *dark and light grey regions* represent total mass functions using only spectroscopic and including photometric redshifts respectively

formation in well-established systems. This quenching progresses toward lower mass systems at later times, consistent with the mass-dependent trends seen in the ages of stellar populations in the quiescent spheroidals.

The physical origin of the quenching of star formation, the fundamental origin of the various downsizing signals, is unclear. Although the merger-induced production of active galactic nuclei may lead to the temporary expulsion and heating of the gaseous halos that surround galaxies [55, 226], key tests of this hypothesis include sustaining the quenching, the weak environmental dependence of the observational trends and the surprisingly clear redshift dependence of the effect (Fig. 32).

4.6 Results: Stellar Mass Functions $z > 1.5$

With current facilities, the stellar mass data beyond $z \simeq 1.5$ generally probes only the higher mass end of the distribution and relies on photometric rather than spectroscopic data. Nonetheless, the results emerging have received as much attention as those at lower redshift. Unlike the complexities of understanding downsizing and the redistribution of mass and morphology in the $z < 1$ data, the basic question at stake here is simply whether the *abundance* of massive $z \simeq 2$ galaxies is larger than expected in the standard model.

Testing the decline with redshift in the comoving abundance of, say, systems with stellar mass greater than $10^{11}\,\mathrm{M}_{\odot}$, expected in ΛCDM has been a frustrating story for the observers for two reasons. Firstly, the most massive systems are rare and clustered, and so determining reliable density estimates beyond $z \simeq 1.5$ has required panoramic deep infrared data which has only recently arrived. For a review of the early observational efforts see [19].

Secondly, there has been considerable confusion in the theoretical literature on the expected rate of decline in $> 10^{11}\,\mathrm{M}_{\odot}$ systems. Early predictions ([124]) claimed a 3-fold decline in the comoving abundance to $z \simeq 1$ in apparent agreement with the large photometric sample analyzed by [15]. However, careful comparisons of independent semi-analytical predictions [50, 125] reveal substantial differences (by an order of magnitude for the same world model) in the rate of decline even to $z \simeq 1$ [19]. In reality, the predictions depend on many parameters where differences can have a large effect for the region of the mass function where the slope $\mathrm{d}N\,/\,\mathrm{d}M$ is steep.

Reference [59] undertook a pioneering study to evaluate the growth of stellar mass with photometric redshift using deep $H < 26.5$ NICMOS data in the Hubble Deep Field North. Although a tiny field ($5\,\mathrm{arcmin}^2$), the work was the first to demonstrate the existence of $10^{11}\,\mathrm{M}_{\odot}$ galaxies at $z \simeq 1.5$–2 as well as the challenges of estimating reliable abundances. A similar HDF-S analysis was undertaken by [196].

A more representative area of 4×30 arcmin was probed spectroscopically in the Gemini Deep Deep Survey ([101]), albeit to a much shallower depth. These authors claimed a "surprising" abundance of $10^{11}\,\mathrm{M}_{\odot}$ systems to $z \simeq 2$. Although an *absolute* comparison of the abundances with semi-analytical predictions is not likely to be illuminating for the reasons mentioned above, the redshift-dependent *mass growth rate* over $1 < z < 2$, derived empirically from both the stellar masses and the integrated star formation rate, seems much slower than expected in the semi-analytical models ([103]). Of particular interest is the fact that the mass growth rate seems not depend strongly on the mass range in question in apparent disagreement with the model predictions (Fig. 33).

Although many of the observers associated with the Gemini Deep Deep Survey have claimed the abundance and slow growth rate of massive galaxies

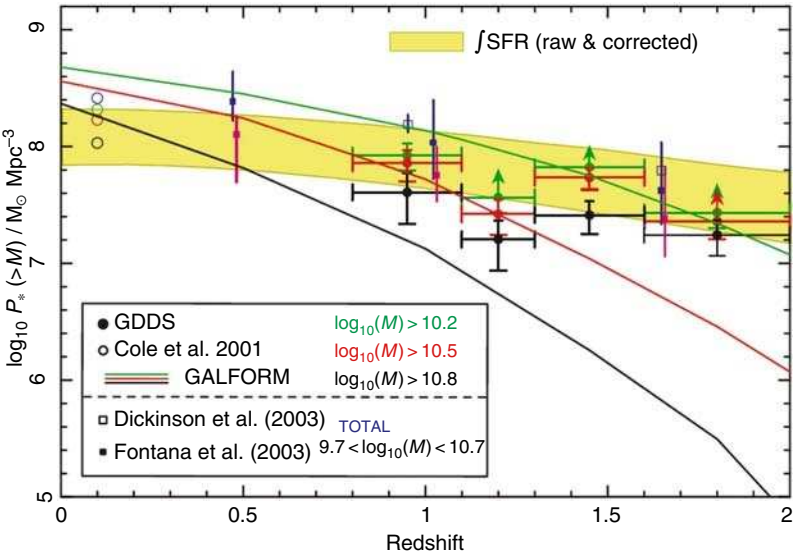

Fig. 33. The mass assembly history derived from stellar masses in the Gemini Deep Deep spectroscopic survey (Glazebrook et al. 2004 [101]). *Colored data points refer* to observed cumulative densities $\rho(>M)$ for various mass ranges and the *green shaded region* to estimates derived from the integrated star formation history. The *colored lines* refer to predicted growth rates according to the GALFORM semi-analytic models

over $1.5 < z < 2$ poses a crisis for the standard model, alongside the puzzling "anti-hierarchical" behavior observed for $z < 1.5$, the plain fact is that there is considerable uncertainty in the semi-analytical predictions.

4.7 Quiescent Galaxies with $2 < z < 3$

Finally, it is illustrative to consider the dramatic effect that infrared data from panoramic ground-based cameras and the Spitzer Space Telescope is having, not only on our knowledge of the distribution of stellar masses at high redshift, but also how stellar mass is distributed among quiescent and star-forming populations. Until recently, there was widespread belief that the bulk of the star formation in this era, and probably a significant fraction of the stellar mass, lay in the Lyman break population. By contrast, [256] have examined the rest-frame $U - V$ colors of a sample of $300 > 10^{11}\,\mathrm{M}_{\odot}$ galaxies with $2 < z_{\mathrm{photo}} < 3$ (Fig. 34) and claim almost 80% are quiescent. The definition of "quiescent" here is somewhat important to get correct, given it now emerges that many of the originally-selected distant red galaxies with $J - K > 2.3$ (Sect. 2.6) turn out to have quite respectable star formation rates. Infrared spectroscopy sensitive to Hα emission and the 4000 Å break is making great strides in clarifying this question ([142]).

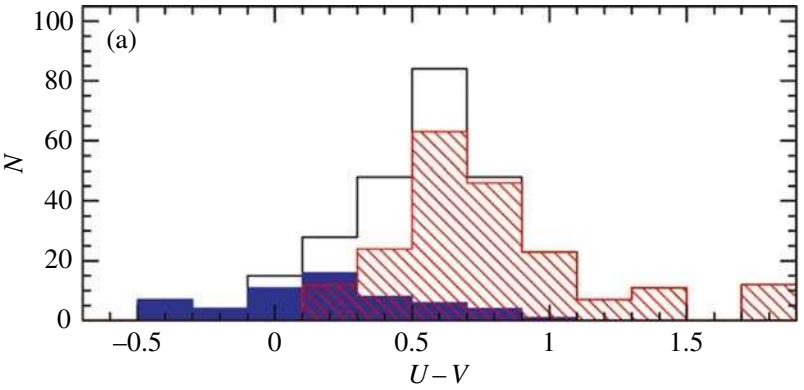

Fig. 34. The distribution of rest-frame $U-V$ color for a Spitzer-selected sample of $2 < z < 3$ galaxies with stellar masses $> 10^{11}\,\mathrm{M}_{\odot}$ from the analysis of van Dokkum et al. (2006) [256]. Almost 80% of this sample in a quiescent state consistent with having completed the bulk of their assembly

4.8 Lecture Summary

In this lecture we have demonstrated important new techniques for estimating the stellar masses of all types of galaxies to impressive high redshift ($z \simeq 3$). The reliability of these techniques is improved greatly by having spectroscopic, rather than photometric, redshift and, for $z > 0.7$, by the addition of infrared photometry. These techniques augment and extend more precise measures available for restricted classes of galaxies – such as the Fundamental Plane for pressure-supported spheroidals, and the Tully-Fisher relationship for rotationally-supported disks.

These various probes point to a self-consistent, but puzzling, description of mass assembly since $z \simeq 1.5$. During this era, massive galaxies have hardly grown at all – most of the star formation and rapid growth is occurring in systems whose masses are less than $10^{10}\,\mathrm{M}_{\odot}$. Remarkably, the star formation appears to be quenched above a certain threshold mass whose value, in turn, is declining with time.

Energetic sources such as supernovae or active nuclei may be responsible for this "downsizing" signature but further work is needed to verify both the weak environmental trends seen in the observations and the redshift-dependent trends in the threshold mass.

Beyond $z \simeq 1.5$, the number and distribution of massive ($> 10^{11}\,\mathrm{M}_{\odot}$ galaxies has led to some surprises. Although the sheer abundance may not be a problem for contemporary models, the fact that so many have apparently completed their star formation is more challenging and consistent with the slow growth rate observed at later times. The observational situation is rapidly developing but consistent with the presence of a surprisingly abundant and mature population of massive galaxies by $z \simeq 3$.

5 Witnessing the End of Cosmic Reionization

5.1 Introduction – Some Weighty Questions

We now turn to the exciting and rapidly developing area of understanding cosmic reionization. This event, which marks the end of the so-called "Dark Ages" when the intergalactic medium became transparent to ultraviolet photons, was a landmark in cosmic history. In some ways the event might be considered as important as the epoch of recombination which isolates the formation of the hydrogen atom, or the asssociated surface of last scattering when photons and baryonic matter decouple. In the case of the era of cosmic reionization, although we cannot yet be sure, many believe we are isolating that period when the first sources had sufficient output to contribute to the energy balance of the intergalactic medium. Even though some early luminous forerunners might be present, the epoch of reionization can be directly connected with cosmic dawn for starlight.

It seems an impossible task to give an authoritative observational account of how to probe this era. So many issues are complete imponderables! When did reionization occur? Was it a gradual event made possible by a complex time sequence of sources, or was there a spectacular synchronized moment? Can we conceive of an initial event, followed by recombination and a second phase?

What were the sources responsible? History has shown the naivety of astronomers in assuming a single population to be responsible for various phenomena – usually some complex combination is the answer. And, perhaps most ambitiously, what is the precise process by which photons escape the sources and create ionized regions?

Four independent observational methods are helpful in constraining the redshift range where we might search for answers to the above questions. In this lecture, we will explore how these work and the current (and rapidly changing) constraints they offer. These are:

- Evolution in the optical depth of Lyman series absorption lines in high redshift quasars ([84, 85]). Although QSOs are only found to redshifts of $z \simeq 6.5$, high quality spectroscopic data gives us a glimpse of a potential change in the degree of neutrality of the IGM beyond $z \simeq 5.5$ such as might be expected if reionization was just ending at that epoch.
- The ubiquity of metal absorption lines in the IGM as probed again by various spectra of the highest redshift QSOs. Carbon, in particular, is only produced in stellar nuclei and its presence in all sightlines to $z \simeq 5$–6 ([221]) is highly indicative of a widespread process of earlier enrichment from the first generation of supernovae.
- The large angular scale power in the temperature - polarization cross-correlation function seen in the microwave background ([138, 225]). This signal is produced by electron scattering foreground to the CMB. Depending

on models of large scale structure, the optical depth of scattering gives some indication of the redshift range where the ionized particles lie.

- The stellar mass density in assembled sources at redshifts $z \simeq 5$–6, about 1 billion years after the Big Bang, as probed by the remarkable combination of HST and Spitzer ([80, 229, 230, 263]). A census of the mass in stars must be the integral of past activity which can be compared with that necessary to reionize the Universe.

5.2 The Gunn-Peterson Test and SDSS QSOs

In a remarkable paper, long before QSOs were located at redshifts beyond 2.5, James Gunn and Bruce Peterson realized ([104]) that the absence of broad troughs of hydrogen absorption in the spectra of QSOs must indicate intergalactic hydrogen is ionized. They postulated a future test whereby the spectra of QSOs of successively higher redshift would be scrutinized to locate that epoch when the IGM was neutral.

For an optical spectrum, redshifted to reveal that portion of the rest-frame UV shortward of the Lyman α emission line of the QSO itself ($\lambda = 1216\,\text{Å}$), the relative transmission T is defined as

$$T = f(\lambda) \,/\, f_{\rm cont} \tag{29}$$

where $f_{\rm cont}$ represents the continuum radiation from the QSOs. The transmission is reduced by Lyman α absorption in any foreground (lower redshift) clouds of neutral hydrogen whose *Gunn-Peterson opacity* is then

$$\tau_{\rm GP} = -\ln T \tag{30}$$

The first "complete troughs" in the absorption line spectra of distant QSO were presented by [13] and [62] and a more comprehensive sampling of 11 SDSS QSOs was presented by [83]. Recently, an analysis of $195.74 < z < 6.42$ QSOs was presented by [85].

Figure 35 illustrates how absorption structures along the same line of sight can be independently probed using Lyman α and higher order lines such as Lyman β.

To understand how this is effective, in a uniform medium $\tau_{\rm GP}$ is related to the abundance of absorbing neutral hydrogen atoms by the following expression:

$$\tau_{\rm GP} = \frac{\pi e^2}{m_{\rm e} c}\, f_\alpha\, \lambda_\alpha\, H^{-1}(z)\, n_{\rm HI} \tag{31}$$

where f is the oscillator strength of the Lyman α line, H is the redshift-dependent Hubble parameter and $n_{\rm HI}$ is the neutral number density.

Numerically, this becomes

$$\tau_{\rm GP} = 1.8\,10^5\, h^{-1}\, \Omega_m^{-\frac{1}{2}}\, \frac{\Omega_{\rm b} h^2}{0.02} \left(\frac{1+z}{7}\right)^{3/2} \frac{n_{\rm HI}}{n_{\rm H}} \tag{32}$$

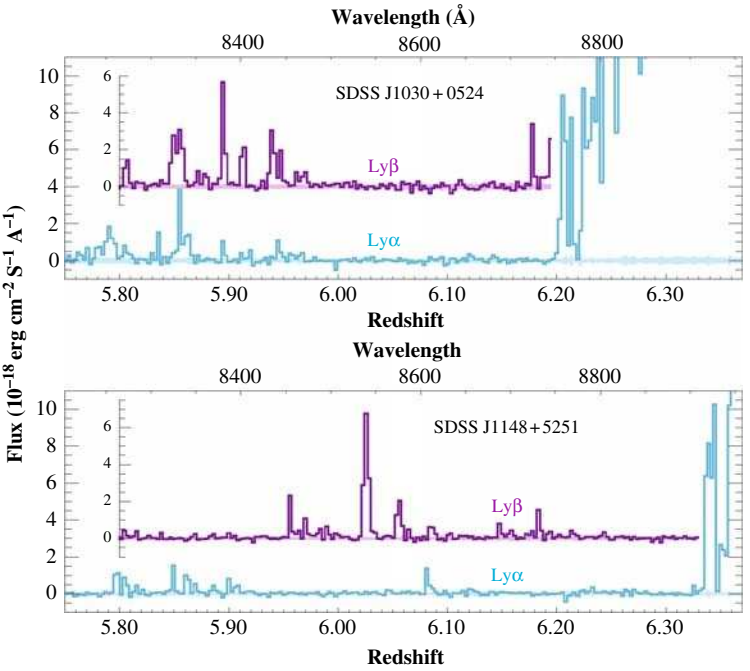

Fig. 35. The absorption line spectrum of two SDSS QSOs from the recent study of Fan et al. (2006a) [84]. Here the structures in the Lyman series absorption lines α and β have been aligned in redshift space thereby improving the signal to noise along a given sightline. The effects of cosmic variance can clearly be seen by comparing the structures seen along the two sightlines

where $x_{\rm HI} = (\frac{1+z}{7})^{3/2}\,\frac{n_{\rm HI}}{n_{\rm H}}$ is then the neutral fraction.

Inspection of this equation is quite revealing. Firstly, even a tiny neutral fraction $x_{\rm HI} \sim 10^{-4}$, would give a very deep, seemingly complete GP trough; for reference $x_{\rm HI} \simeq 10^{-5}$ today. So clearly the test is not a very sensitive one in absolute terms.

Secondly, since $\tau_{\rm GP} \propto f\,\lambda$, for the same $n_{\rm H}$, the optical depth in the higher order Lyman β and γ lines would be $\simeq$6 and 18 times smaller respectively.

In practice, the above relations are greatly complicated by any clumpiness in the medium. This affects our ability to make direct inferences on $x_{\rm HI}$ as well as to *combine* the various Lyman lines into a single test.

Instead, workers have examine the *relative* distribution of $\tau_{\rm GP}$ with redshift independently from the various Lyman series absorption statistics. An increase in $\tau_{\rm GP}$ with redshift could just be a natural thickening of the Lyman absorption forest and, given its weak connection with $x_{\rm HI}$, not imply anything profound about cosmic reionization. However, if it can be shown empirically that the various diagnostics show a *discontinuity* in the $x_{\rm HI}$-redshift trends, conceivably we are approaching the neutral era.

Figure 36 shows that for $z < 5.5$, $\tau \propto (\frac{1+z}{5})^{4.3}$ for both the Lyα and Lyβ forests to reasonable precision. However, beyond $z \simeq 5.5$, both redshift trends are much steeper, $\propto (1+z)^{11}$. The dispersion around the trends also increases significantly at higher redshift. Taken together, both results suggest a qualitative change in nature of the IGM beyond $z \simeq 5.5$ (but for an alternative explanation see [8]).

References [84] and [85] discuss several further probes of the nature of the IGM at $z \simeq 6$. One relates to the *proximity effect* – the region around each QSO where it is clear from the spectrum that the IGM is being ionized by the QSO itself. Although this region is excluded in the analyses above, the *extent* of this region contains valuable information on the nature of the IGM. It appears that the radius of the region affected, R is less at higher redshift according to $R \propto [(1+z)\, x_{\rm HI}]^{-1/3}$ suggesting that the most distant QSOs in the sample ($z \simeq 6.5$) lie in a IGM whose neutral fraction is $\simeq 14$ times higher than those at $z \simeq 5.7$.

Another valuable measure is how the regions of complete absorption, the so-called "indexdark gaps" in the spectrum where transmission is effectively zero, are distributed. Fan et al. define a "gap" as a contiguous region in redshift space where $\tau > 3.5$. The distribution of gaps contains some information on the topology of reionization. We would expect regions of high transmission to be associated with large HII regions, centered on luminous star-forming sources. The dark gaps increase in extent from $\simeq 10$ to 80 comoving Mpc over the redshift range samples suggesting the IGM is still not neutral at $z \simeq 6.5$. Although the Gunn-Peterson and gap statistics make similar statements about reionization, suggesting the neutral fraction at $z \simeq 6.2$ is 1–4%, Fan et al. consider that beyond $z \simeq 6.5$, gap statistics will become a more powerful probe. This is because the redshift distribution of those few spectroscopic pixels where the transmission is non-zero will become the only effective signal.

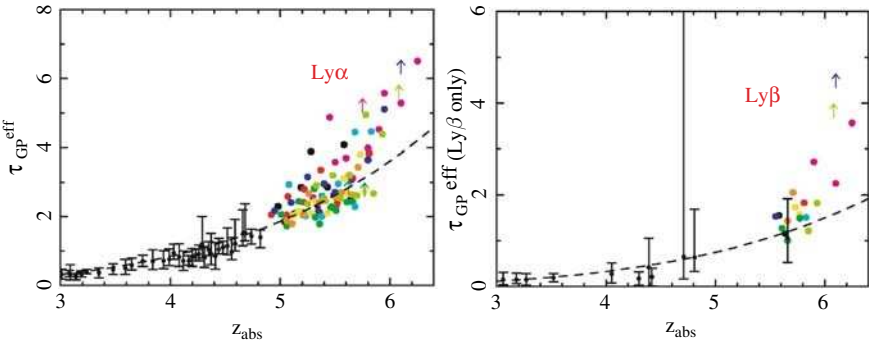

Fig. 36. Evolution in the Gunn-Peterson optical depth, τ, for both the Lyman α **left** and β (**right**) forests from the distant SDSS QSO analysis of Fan et al (2006a) [84]. The dotted lines represent fits to the data for $z_{\rm abs} < 5.5$, beyond which there is evidence in both species for an upturn in the opacity of the intergalactic medium

5.3 Metallicity of the High Redshift IGM

A second measure of the redshift range of early star formation is contained in the properties of the CIV forest observed in the spectra of high redshift QSOs ([222, 223]). Carbon is only produced in stellar nuclei (it is not produced in the hot Big Bang) and so the ubiquity of CIV along many sightlines to $z \simeq 5$–6 QSOs is a powerful argument for early enrichment.

CIV was seen in the Lyα forest in 1995 ([48]) with N(CIV)/N(HI) $\sim 10^{-2}$–10^{-3}. However, it was subsequently seen in even the weakest Lyα systems (Ellison et al. 2000) [78]. This is a particularly powerful point since it argues that enrichment is not confined to localized regions of high column density but is generic to the intergalactic medium as a whole (Fig. 37)

A quantitative interpretation of the CIV abundance, in terms of how much early star formation occurred earlier than the highest redshift probed, relies on locating a "floor" in the abundance-redshift relation. Unfortunately, the actual observed trend, measured via the contribution of the ion to the mass density $\Omega(CIV)$ from $z \simeq 5$ to 2, does not seem to behave in the manner expected. For example, there is no strong rise in the CIV abundance to lower redshift despite the obvious continued star formation that occurred within these epochs ([223], Fig. 38). This is a major puzzle (cf. [176]).

5.4 Linear Polarization in the WMAP Data

In 2003, the WMAP team ([138]) presented the temperature-polarization cross angular power spectrum from the first year's data and located a 4σ non-zero signal at very low multipoles ($l < 8$) which they interpreted in terms of foreground electron scattering of microwave background photons with an optical depth $\tau_e = 0.17 \pm 0.04$ corresponding to ionized structures at $z_{\rm reion} \simeq 20$ (Fig. 39a). The inferred redshift range depends sensitively on the history of the reionization process. Reference [18] argued that if reionization occurred instantaneously it corresponds to a redshift $z_{\rm reion} \simeq 17 \pm 5$, whereas adopting a more reasonable Press-Schechter formalism and an illustrative cooling and enrichment model, [92] demonstrated that the same signal can be interpreted with a delayed reionization occurring at $z_{\rm reion} \simeq 9$–10.

Just before the Saas-Fee lectures, the long-awaited third year WMAP data was published ([225]). A refined analysis significantly lowered both the normalization of the dark matter power spectrum to $\sigma_8 = 0.74 \pm 0.05$, and the optical depth to electron scattering to $\tau_e = 0.09 \pm 0.03$. The same model of instantaneous reionization reduces the corresponding redshift to $z_{\rm reion} = 11 \pm 3(2\sigma)$ – a significant shift from the 1 year data.

To illustrate the uncertainties, Spergel et al. introduce a more realistic history of reionization via the ionization fraction x_e. Suppose above $z_{\rm reion}$, $x_{\rm e} \equiv 0$ and below $z = 7$, $x_{\rm e} \equiv 1$. Then suppose $z_{\rm reion}$ is defined as that intermediate point when $x_{\rm e} = x_{\rm e}^0$ for $7 < z < z_{\rm reion}$. Fig. 39b illustrates the remarkable *insensitivity* of $z_{\rm reion}$ to the adopted value of $x_{\rm e}^0$ for $x_{\rm e}^0 < 0.5$.

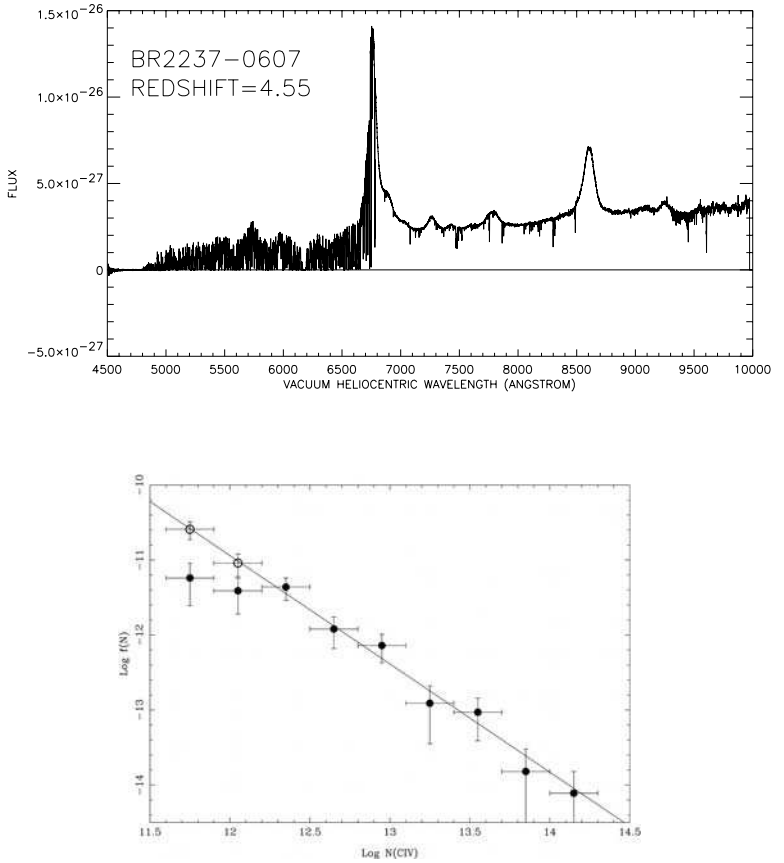

Fig. 37. (**Top**) Keck ESI absorption line spectrum of the z = 4.5 QSO BR2237-0607 from the study of Songaila (2005) [222]. A sparsely populated CIV forest from 7500 to 8500 *Å* accompanies the dense Lyα forest seen below 6800 Å. (**Bottom**) Distribution of column densities of CIV absorbers per unit redshift interval in Q1422+231 from the survey of Ellison et al. (2000) [78]

Despite improved data, the redshift range implied by the WMAP data spans the full range $10 < z < 20$.

5.5 Stellar Mass Density at $z \simeq 5$–6

Neither the Gunn-Peterson test nor the WMAP polarization data necessarily demonstrate that reionization was caused by early star-forming sources; both only provide constraints on when the intergalactic medium was first reionized.

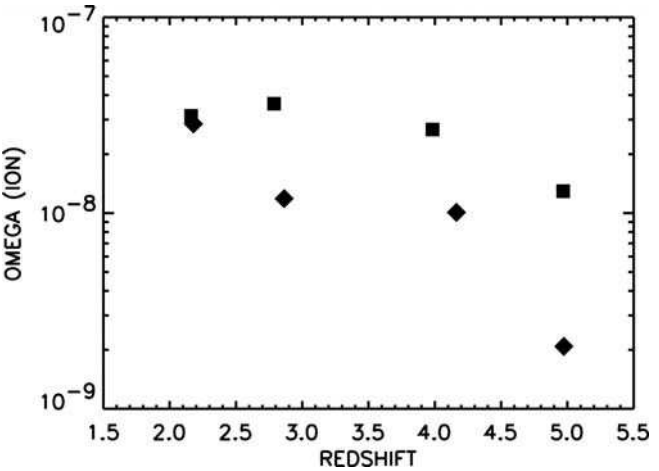

Fig. 38. Modest evolution in the contribution of intergalactic CIV and SIV over $2 < z < 5$ as measured in terms of the ionic contribution to the mass density, Ω (Songaila 2005 [222])

The CIV test is a valuable complement since it provides a measure of early enrichment which can only come from star-forming sources. Unfortunately, powerful though the ubiquitous presence of CIV is in this context, as we have seen, quantitative constraints are hard to derive. We thus seek a further constraint on the amount of early star formation that might have occurred.

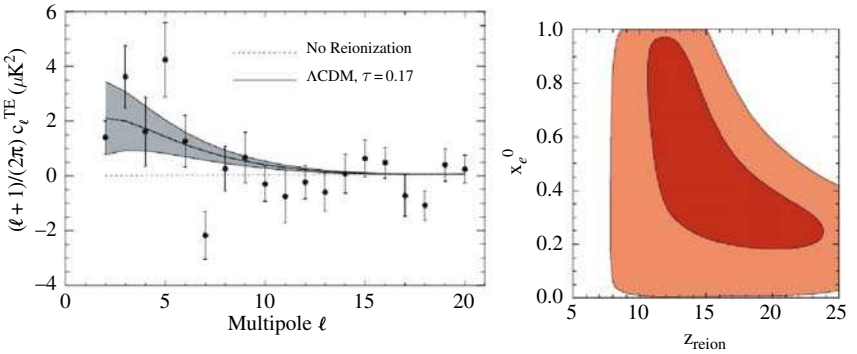

Fig. 39. (**Left**) The 4σ detection of reionization via an excess signal at large scales in the angular cross correlation power spectrum of the temperature and polarization data in the first year WMAP data (Kogut et al. 2003 [138]). (**Right**) Constraints on the redshift of reionization, $z_{\rm reion}$, from the third year WMAP data (Spergel et al. 2006 [225]). The contours illustrate how the $z_{\rm reion}$ inferred from the lowered optical depth depends on the history of the ionized fraction $x_{\rm e}(z)$, see text for details

In Lecture 4 we introduced the techniques astronomers are now using to derive *stellar masses* for distant galaxies. Although the techniques remain approximate, it must follow that the stellar mass density at a given epoch represents the integral over time and volume of the past star formation. Indeed, we already saw a successful application of this in reconciling past star formation with the local stellar mass density observed by 2dF (Fig. 21). Specifically, at a particular redshift, z

$$M_* = \int_{\infty}^{z} \rho_*(z)\, dV(z) \tag{33}$$

Using the techniques described in Lecture 4, stellar mass estimates have become available for some very high redshift galaxies detected by Spitzer [79, 171, 262]. For the most luminous Lyman "dropouts", these estimates are quite substantial, some exceeding $10^{11}\,\mathrm{M}_{\odot}$ implying much earlier activity. Recently, several groups [80, 230, 263] have been motivated to provide the first crude estimates on the volume averaged *stellar mass density* at these early epochs. Part of this motivation is to check whether the massive galaxies seen at such high redshift can be reconciled with hierarchical theory, but as [229] proposed, the established stellar mass can also be used to probe earlier star formation and its likely impact upon cosmic reionization.

A very relevant question is whether the observed mass density at $z \simeq 5$–6 is greater than can be accounted for by the observed previous star formation history. We will review the rather uncertain data on the star formation density $\rho_*(z)$ beyond $z \simeq 6$ in the next Lecture. However, [230] find that even taking a reasonably optimistic measure of $\rho_*(z)$ from recent compilations by [29] and [41], it is hard to account for the stellar mass density at $z \simeq 5$ (Fig. 40).

There are currently two major limitations in this comparison. First, most of the v- and i-drops are located photometrically; even a small degree of contamination from lower redshift galaxies could upward bias the stellar mass density. On the other hand only star-forming galaxies are located by the Lyman break technique so this bias could easily be offset if there are systems in a quiescent state as evidenced by the prominent Balmer breaks seen in many of the Spitzer-detected sources [79, 80]. This limitation will ultimately be overcome with more careful selection methods and deeper spectroscopy. Secondly, and more profoundly, the precision of the stellar masses may not be up to this comparison. Much has to be assumed about the nature of the stellar populations involved which may, quite reasonably, be somewhat different from those studied locally. The discrepancy noted by Stark et al. is only a factor of $\times 2$–3, possibly within the range of uncertainty.

Regardless, if this mismatch is reinforced by better data, the implications are very interesting in the context of reionization. It could mean early star-forming systems are extincted, lie beyond $z \simeq 10$ where current searches end, or perhaps most likely that early star formation is dominated by lower luminosity systems (Fig. 40). By refining this technique and using diagnostics such as the strength of the tell-tale Balmer break, it may ultimately be possible

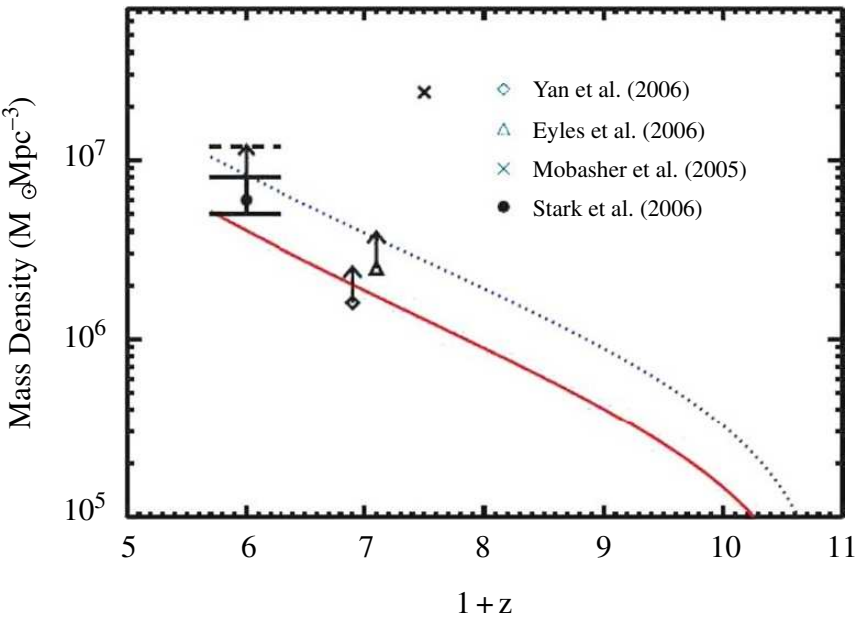

Fig. 40. A comparison of the assembly history of stellar mass inferred from the observed decline in star formation history to $z \simeq 10$ (solid line) with extant data on the stellar mass density at $z \simeq 5$ and 6 (data points from Stark et al. 2006a [230], Yan et al. 2006 [263], Eyles et al. 2006 [80]). Different estimates at a given redshift represent lower limits based on spectroscopically-confirmed and photometric redshift samples. The red line shows the growth in stellar mass expected from the presently-observed luminous star forming galaxies; a shortfall is observed. The blue dotted line shows the improvement possible when a dominant component of high z lower luminosity systems is included

to age-date the earlier activity and compare its efficacy with that required to reionize the Universe.

5.6 Lecture Summary

In this lecture we have introduced four very different and independent probes of cosmic reionization, each of which suggests star formation activity may extend well into the redshift range $6 < z < 20$. Two of these probes rely on a contribution from early star formation (the metallicity of the intergalactic medium and the assembled stellar mass density at $z \simeq 5$–6).

The earliest result was the presence of neutral hydrogen troughs in the spectra of distant QSOs. Although the arguments for reionization ending at $z \simeq 6$ seem compelling at first sight, they ultimately rely on an empirically-deduced transition in the changes in the opacity of the Lyα and Lyβ line below and above $z \simeq 5.5$.

The second result – the ubiquity of carbon in even the weakest absorbing clouds at $z \simeq 5$ is firm evidence for early star formation. However, it seems hard to locate the high redshift "abundance floor" and hence to quantify whether this early activity is sufficient for reionization. Indeed, a major puzzle is the lack of growth in the carbon abundance over the redshift range where galaxies are assembling the bulk of their stars.

The WMAP polarizations results have received the most attention, mainly because the first year data indicated a surprisingly large optical depth and a high redshift for reionization. However, there were some technical limitations in the original analysis and it now seems clear that the constraints on the redshift range when the foreground polarization is produced are not very tight.

Finally, an emerging and very promising technique is simply the census of early star formation activity as probed by the stellar masses (and ages) of the most luminous dropouts at $z \simeq 5$–6. Although significant uncertaintes remain, the prospects for improving these constraints are good and, at this moment, it seems there must have been quite a significant amount of early ($z > 6$) star formation activity, quite possibly in low luminosity precursors.

6 Into the Dark Ages: Lyman Dropouts

6.1 Motivation

Surveys of galaxies at and beyond a redshift $z \simeq 6$ represent the current observational frontier. We are motivated to search to conduct a census of the earliest galaxies seen 1 Gyr after the Big Bang as well as to evaluate the contribution of early star formation to cosmic reionization. Although impressive future facilities such as the next generation of extremely large telescope[6] and the James Webb Space Telescope ([96]) are destined to address these issues in considerable detail, any information we can glean on the abundance, luminosity and characteristics of distant sources will assist in planning their effective use.

In this lecture and the next, we will review the current optical and near-infrared techniques for surveying this largely uncharted region. They include

- Lyman dropouts:
- Lyman dropouts photometric searches based on locating the rest-frame ultraviolet continuum of star-forming sources introduced in Lecture 3. The key issue here is reducing contamination from foreground sources since most sources selected via this technique are too faint for confirmatory spectroscopy.
- Lyman alpha emitters: spectroscopic or narrow-band searches for sources with intense Lyα emission. As the line is resonantly-scattered by neutral

[6] E.g. The US-Canadian Thirty Meter Telescope - http//www.tmt.org

hydrogen its profile and strength gives additional information on the state of the high redshift intergalactic medium.
- Strong gravitational lensing: by coupling both above techniques with the magnification afforded by lensing clusters, it is possible to search for lower luminosity sources at high redshift. Since the magnified areas are small, the technique is only advantageous if the luminosity function has a steep faint end slope.

For the Lyman dropouts discussed here, as introduced in the last lecture, there is an increasingly important role played by the Spitzer Space Telescope in estimating stellar masses and earlier star formation histories.

The key questions we will address in this lecture focus on the (somewhat controversial) conclusions drawn from the analyses of Lyman dropouts thus far, namely:

1. How effective are the various high z selection methods? The characteristic luminosity L^* at $z \simeq 6$ corresponds to $i_{\rm AB} \simeq 26$ where spectroscopic samples are inevitably biased to those with prominent Lyα emission. Accordingly there is great reliance on photometric redshifts and a real danger of substantial contamination by foreground red galaxies and Galactic cool stars.
2. Is there a decline in the UV luminosity density, $\rho_{\rm UV}$, over the range $3 < z < 6$? The early results were in some disagreement. Key issues relate to the degree of foreground contamination and cosmic variance in the very small deep fields being examined.
3. Is the observed UV density, $\rho_{\rm UV}$, at $z \simeq 6$ sufficient to account for reionization? The answer depends on the contribution from the faint end of the luminosity function and whether the UV continuum slope is steeper than predicted for a normal solar-metallicity population.
4. Significant stellar masses have been determined for several $z \simeq 6$ galaxies. Are these in conflict with hierarchical structure formation models?

6.2 Contamination in $z \simeq 6$ Dropout Samples

The traditional dropout technique exploited very effectively at $z \simeq 3$ (Lecture 3) is poorly suited for $z \simeq 6$ samples because the use of a simple $i-z > 1.5$ color cut still permits significant contamination by passive galaxies at $z \simeq 2$ and Galactic stars. The addition of an optical-infrared color allows some measure of discrimination ([228]) since a passive $z \simeq 2$ galaxy will be red over a wide range in wavelength, whereas a star-forming $z \simeq 6$ galaxy should be relatively blue in the optical-infrared color corresponding to its rest-frame ultraviolet (Fig. 41). Application of this two color technique suggests contamination by foreground galaxies is $\simeq 10\%$ at the bright end ($z_{\rm AB} < 25.6$) but negligible at the UDF limit ($z_{\rm AB} < 28.5$)

Unfortunately, the spectral properties of cool Galactic L dwarfs are dominated by prominent molecular bands rather than simply by their effective

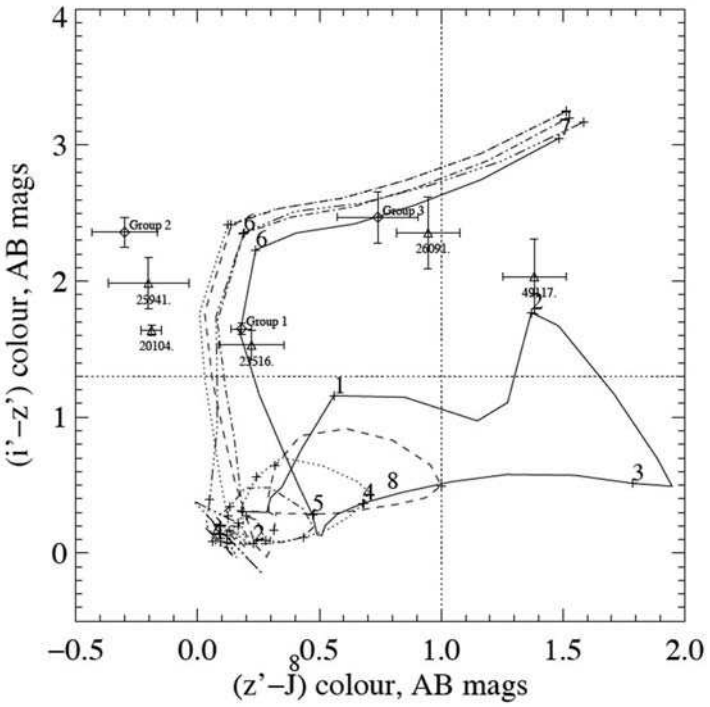

Fig. 41. The combination of a $i - z$ and $z - J$ color cut permits the distinction of $z \simeq 5.7$–6.5 star forming and $z \simeq 2$ passive galaxies. Both may satisfy the $i - z > 1.5$ dropout selector, but the former should lie blueward of the $z - J = 1.0$ divider, whereas $z \simeq 2$ are red in both colors. Crosses represent the location of candidates in the GOODS field and model tracks illlustrate the predicted colors for typical SEDs observed at the respective redshifts (Stanway et al. 2005 [228])

temperature. This means that they cannot be separated from $z \simeq 6$ galaxies in a similar color-color diagram. Indeed, annoyingly, these dwarfs occupy precisely the location of the wanted $z \simeq 6$ galaxies (Fig. 42)! The only practical way to discriminate L dwarfs is either via spectroscopy or their unresolved nature in ACS images.

Reference [227] conducted the first comprehensive spectroscopic and ACS imaging survey of a GOODS i-drop sample limited at $z_{\rm AB} < 25.6$, finding that stellar contamination at the bright end of the luminosity function of a traditional ($i_{\rm AB} - z_{\rm AB} > 1.5$) color cut could be as high as 30–40%. Unfortunately, even with substantial 6–8 hr integrations on the Keck telescope, redshift verification of the distant population was only possible in those dropout candidates with Lyα emission. Reference [228] subsequently analyzed the ACS imaging properties of a fainter subset arguing that stellar contamination decreases with increasing apparent magnitude.

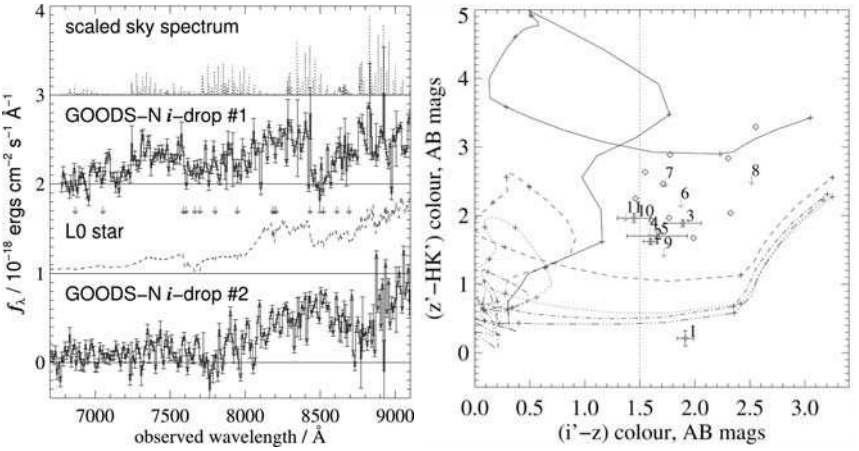

Fig. 42. (**Left**) Keck spectroscopic verification of two contaminating L dwarfs lying within the GOODS $i-z$ dropout sample but pinpointed as likely to be stellar from ACS imaging data. The smoothed spectra represent high signal to noise brighter examples for comparison purposes. Strong molecular bands clearly mimic the Lyman dropout signature. (**Right**) Optical-infrared color diagram with the dropout color selector, $i_{\rm AB} - z_{\rm AB} > 1.5$, shown as the vertical dotted line. Bright L dwarfs (lozenges) frustratingly occupy a similar region of color space as the $z \simeq 6$ candidates (points with error bars) (Stanway et al. 2004 [227])

Further progress has been possible via the use of the ACS grism on board HST ([160]). As the OH background is eliminated in space, despite its low resolution, it is possible even in the fairly low signal/noise data achievable with the modest 2.5 m aperture of HST to separate a Lyman break from a stellar molecular band. It is claimed that of 29 $z_{\rm AB} < 27.5$ candidates with $(i-z) > 0.9$, only 6 are likely to be low redshift interlopers.

Regrettably, as a result of these difficulties, it has become routine to rely entirely on photometric and angular size information without questioning further the degree of contamination. This is likely one reason why there remain significant discrepancies between independent assessments of the abundance of $z \simeq 6$ galaxies [29, 40, 99]. Although there are indications from the tests of [227, 228] and [160] that contamination is significant only at the bright end, the lack of a comprehensive understanding of stellar and foreground contamination remains a major uncertainty.

6.3 Cosmic Variance

The deepest data that has been searched for i-band dropouts includes the two GOODS fields [60] and the Hubble Ultra Deep Field (UDF, [14]). As these represent publicly-available fields they have been analyzed by many groups to various flux limits. The Bunker/Stanway team probed the GOODS fields

to $z_{\rm AB} = 25.6$ (spectroscopically) and 27.0 (photometrically, and the UDF to $z_{\rm AB} = 28.5$. At these limits, it is instructive to consider the comoving cosmic volumes available in each field within the redshift range selected by the typical dropout criteria. For both GOODS-N/S fields, the total volume is $\simeq 5 \times 10^5\,{\rm Mpc}^3$, whereas for the UDF it is only $2.6 \times 10^4\,{\rm Mpc}^3$. These contrast with $10^6\,{\rm Mpc}^3$ for a single deep pointing taken with the SuPrime Camera on the Subaru 8 m telescope.

Reference [220] present a formalism for estimating, for any population, the fractional uncertainty in the inferred number density from a survey of finite volume and angular extent. When the clustering signal is measurable, the cosmic variance can readily be calculated analytically. However, for frontier studies such as the i-dropouts, this is not the case. Here Somerville et al. propose to estimate cosmic variance by appealing to the likely halo abundance for the given observed density using this to predict the clustering according to CDM models. In this way, the uncertainties in the inferred abundance of i-dropouts in the combined GOODS fields could be $\simeq$ 20–25% whereas that in the UDF could be as high as 40–50%.

It seems these estimates of cosmic variance can only be strict lower limits to the actual fluctuations since Somerville et al. make the assumption that halos containing star forming sources are visible at all times. If, for example, there is intermittent activity with some duty cycle whose "on/off" fraction is f, the cosmic variance will be underestimated by that factor f (Stark et al., in prep).

6.4 Evolution in the UV Luminosity Density $3 < z < 10$?

The complementary survey depths means that combined studies of GOODS and UDF have been very effective in probing the shape of the UV luminosity function (LF) at $z \simeq 6$[7]. Even so, there has been a surprising variation in the derived faint end slope α. Reference [40] claim their data (54 i-dropouts) is consistent with the modestly-steep $\alpha = -1.6$ found in the $z \simeq 3$ Lyman break samples ([237]), whereas [261] extend the UDF counts to $z_{\rm AB} = 30.0$ and, based on 108 candidates, find $\alpha = -1.9$, a value close to a divergent function! Issues of sample completeness are central to understanding whether the LF is this steep.

In a comprehensive analysis based on all the extant deep data, [30] have attempted to summarize the decline in rest-frame UV luminosity density over $3 < z < 10$ as a function of luminosity (Fig. 43). They attribute the earlier discrepancies noted between [40, 29, 99] to a mixture of cosmic variance and differences in contamination and photometric selection. Interestingly, they claim a *luminosity-dependent* trend in the sense that the bulk of the decline occurs in the abundance of luminous dropouts, which they attribute to hierarchical growth.

[7] In this section we will only refer to the *observed* (extincted) LFs and luminosity density

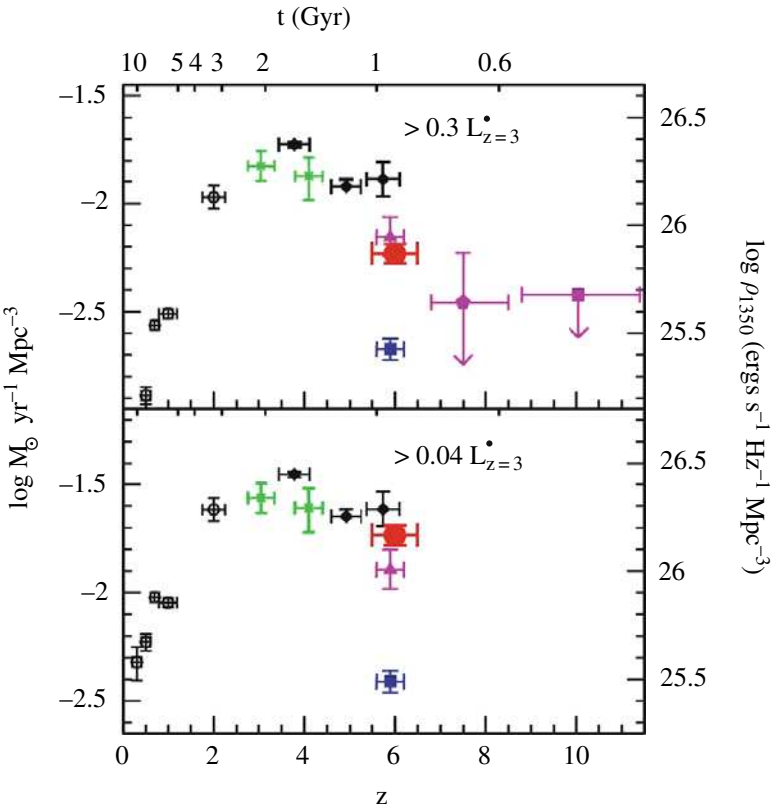

Fig. 43. Evolution in the rest-frame UV (1350 Å) luminosity density (right ordinate) and inferred star formation rate density ignoring extinction (left ordinate) for dropout samples in two luminosity ranges from the compilation by Bouwenes et al. (2006) [30]. A marked decline is seen over $3 < z < 6$ in the contribution of luminous sources

A similar trend is seen in ground-based data obtained with Subaru. Although HST offers superior photometry and resolution which is effective in eliminating stellar contamination, the prime focus imager on Subaru has a much larger field of view so that each deep exposure covers a field twice as large as both GOODS N+S. As they do not have access to ACS data over such wide fields, the Japanese astronomers have approached the question of stellar contamination in an imaginative way. Reference [213] used two intermediate band filters at 709 and 826 nm to estimate stellar contamination in both $z \simeq 5$ and $z \simeq 6$ broad-band dropout samples. By considering the *slope* of the continuum inbetween these two intermediate bands, in addition to a standard $i - z$ criterion, they claim an ability to separate L and T dwards. In a similar, but independent, study, [212] split the z band into two intermediate filters thereby measuring the rest-frame UV slope just redward of the Lyman

discontinuity. These studies confirm both the redshift decline and, to a lesser extent, the luminosity-dependent trends seen in the HST data.

Although it seems there is a $\times 5$ abundance decline in luminous UV emitting galaxies from $z \simeq 3$ to 6, it's worth noting again that the relevant counts refer to sources uncorrected for extinction. This is appropriate in evaluating the contribution of UV sources to the reionization process but not equivalent, necessarily, to a decline in the star formation rate density. Moreover, although the luminosity dependence seems similar in both ground and HST-based samples, it remains controversial (e.g. [14]).

6.5 The Abundance of Star Forming Sources Necessary for Reionization

Have enough UV-emitting sources been found at $z \simeq 6$–10 to account for cosmic reionization? Notwithstanding the observational uncertainties evident in Fig. 43, this has not prevented many teams from addressing this important question. The main difficulty lies in understanding the physical properties of the sources in question. The plain fact is that we cannot predict, sufficiently accurately, the UV luminosity density that is sufficient for reionization!

Some years ago, [157] estimated the star formation rate density based on simple parameterized assumptions concerning the stellar IMF and/or metallicity Z essential for converting a 1350 Å luminosity into the integrated UV output, the fraction $f_{\rm esc}$ of escaping UV photons, the clumpiness of the surrounding intergalactic hydrogen, $C =< \rho^2_{\rm HI} > / < \rho_{\rm HI} >^2$, and the temperature of the intergalactic medium $T_{\rm IGM}$. In general terms, for reionization peaking at a redshift $z_{\rm reion}$, the necessary density of sources goes as:

$$\rho \propto f_{\rm esc}^{-1}\, C\, (1+z)^3\, (\Omega_{\rm B} h^2)^2\, {\rm Mpc}^{-3} \tag{34}$$

For likely ranges in each of these parameters, [233] tabulate the required source surface density which, generally speaking, lie above those observed at $z \simeq 6$ (e.g. [40]).

It is certainly possible to reconcile the end of the reionization at $z \simeq 6$ with this low density of sources (Fig. 43) by appealing to cosmic variance, a low metallicity and/or top-heavy IMF ([233]) or a steep faint end slope of the luminosity function ([261]) but none of these arguments is convincing without further proof. As we will see, the most logical way to proceed is to explore both the extent of earlier star formation from the mass assembled at $z \simeq 5$–6 (Lecture 5) and to directly measure, if possible, the abundance of low luminosity sources at higher redshift.

6.6 The Spitzer Space Telescope Revolution: Stellar Masses at $z \simeq 6$

One of the most remarkable aspects of our search for the most distant and early landmarks in cosmic history is that a modest cooled 85 cm telescope,

the Spitzer Space Telescope, can not only assist but provide crucial diagnostic data! The key instrument is the InfraRed Array Camera (IRAC) which offers four channels at 3.6, 4.5, 5.8 and 8 μm corresponding to the rest-frame optical 0.5–1 μm at redshifts $z \simeq 6$–7. In the space of only a year, the subject has progressed from the determination of stellar masses for a few $z \simeq 5$–7 sources to mass densities and direct constraints on the amount of early activity, as discussed in Lecture 5.

An early demonstration of the promise of IRAC in this area was provided by [79] who detected two spectroscopically-confirmed $z \simeq 5.8$ i-band dropouts at 3.6 μm, demonstrating the presence of a strong Balmer break in their spectral energy distributions (Fig. 44). In these sources, the optical detection of Lyα emission provides an estimate of the current ongoing star formation rate, whereas the flux longward of the Balmer break provides a measure of the past averaged activity. The combination gives a measure of the *luminosity weighted age* of the stellar population. In general terms, a Balmer break appears in stars whose age cannot, even in short burst of activity, be younger than 100 Myr. Eyles et al. showed such systems could well be much older (250–650 Myr) depending on the assumed form of the past activity. As the Universe is only 1 Gyr old at $z \simeq 6$, the IRAC detections gave the first indirect glimpse

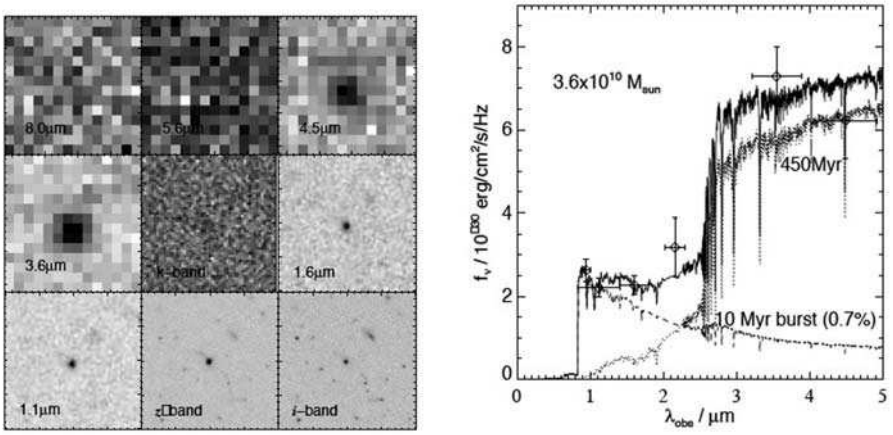

Fig. 44. (**Left**) Detection of a spectroscopically-confirmed i-drop at $z = 5.83$ from the analysis of Eyles et al. (2005) [79]. (**Right**) Spectral energy distribution of the same source. Data points refer to IRAC at 3.6 and 4.5 μm, VLT (K) and HST NICMOS (J,H) overplotted on a synthesised spectrum; note the prominent Balmer break. Synthesis models indicate the Balmer break takes 100 Myr to establish. However, the luminosity-weighted age could be significantly older depending on the assumed past star formation rate. In the example shown, a dominant 450 Myr component ($z_F \sim 10$) is rejuvenated with a more recent secondary burst whose ongoing star formation rate is consistent with the Lyα flux observed in the source

of significant earlier star formation – a glimpse that was elusive with direct searches at the time.

Independent confirmation of both the high stellar masses and prominent Balmer breaks was provided by the analysis of [262] who studied 3 $z \simeq 5.9$ sources. Moreover, Yan et al. also showed several objects had $(z - J)$ colors bluer than the predictions of the Bruzual-Charlot models for all reasonable model choices – a point first noted by [227].

Eyles et al. and Yan et al. proposed the presence of established stellar populations in $z \simeq 6$ i-drops and also to highlight the high stellar masses $(M \simeq (1-4) \times 10^{10}\,\mathrm{M}_{\odot})$ they derived. At first sight, the presence of $z \simeq 6$ sources as massive as the Milky Way seems a surprising result. Yan et al. discuss the question in some detail and conclude the abundance of such massive objects is not inconsistent with hierarchical theory. In actuality it is hard to be sure because cosmic variance permits a huge range in the derived volume density and theory predicts the halo abundance (e.g. [7]) rather than the stellar mass density. To convert one into the other requires a knowledge of the star formation efficiency and its associated duty-cycle.

One early UDF source detected by IRAC has been a particular source of puzzlement. Reference [171] found a J-dropout candidate with a prominent detection in all 4 IRAC bandpasses. Its photometric redshift was claimed to be $z \simeq 6.5$ on the basis of both a Balmer and a Lyman break. However, despite exhaustive efforts, its redshift has not been confirmed spectroscopically. The inferred stellar mass is $(2-7) \times 10^{11}\,\mathrm{M}_{\odot}$, almost an order of magnitude larger than the spectroscopically-confirmed sources studied by Eyles et al. and Yan et al. If this source is truly at $z \simeq 6.5$, finding such a massive galaxy whose star formation likely peaked before $z \simeq 9$ is very surprising in the context of contemporary hierarchical models. Such sources should be extremely rare so finding one in the tiny area of the UDF is all the more puzzling. Reference [70] have proposed the source must be foreground both on account of an ambiguity in the photometric redshift determination and the absence of similarly massive sources in a panoramic survey being conducted at UKIRT ([152]).

This year, the first estimates of the *stellar mass density* at $z \simeq 5$–6 have been derived [80, 230, 263]. Although the independenty-derived results are consistent, both with one another and with lower redshift estimates (Fig. 45) the uncertainties are considerable as discussed briefly in the previous lecture. There are four major challenges to undertaking a census of the star formation at early times.

Foremost, the bulk of the faint sources only have photometric redshifts. Even a small amount of contamination from foreground sources would skew the derived stellar mass density upward. Increasing the spectroscopic coverage would be a big step forward in improving the estimates.

Secondly, IRAC suffers from image confusion given its lower angular resolution than HST (4 arcsec c.f. 0.1 arcsec). Accordingly, the IRAC fluxes cannot be reliably estimates for blended sources. Stark et al. address this by measuring the masses only for those uncontaminated, isolated sources, scaling up

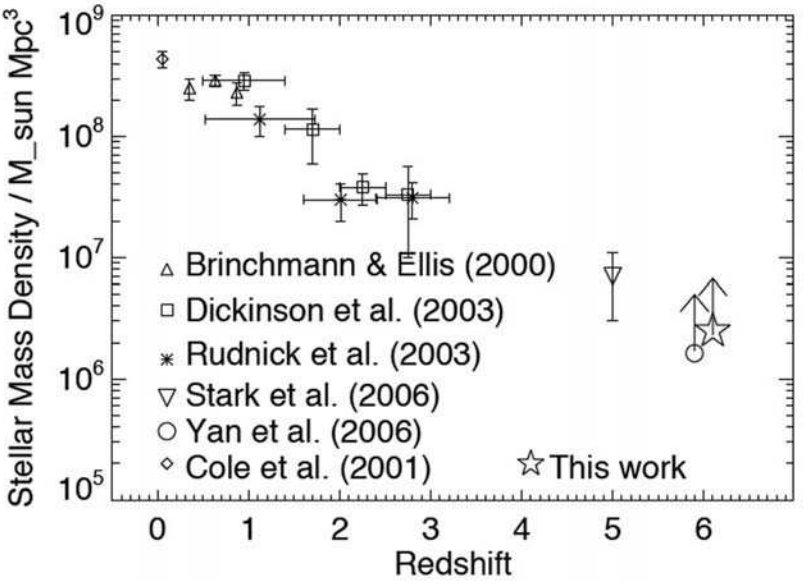

Fig. 45. Evolution in the comoving stellar mass density from the compilation derived by Eyles et al. (2006) [80]. The recent $z \simeq 5$–6 estiimates constitute lower limits given the likelihood of quiescent sources missed by the drop-out selection technique. Results at $z \simeq 6$ are offset slightly in redshift for clarity

their total by the fraction omitted. This assumes confused sources are no more or less likely to be a high redshift.

Thirdly, as only star forming sources are selected using the $v-$ and $i-$ dropout technique, if star formation is episodic, it is very likely that quiescent sources are present and thus the present mass densities represent lower limits. The missing fraction is anyone's guess. As we saw at $z \simeq 2$, the factor could be as high as $\times 2$.

Finally, as with all stellar mass determinations, many assumptions are made about the nature of the stellar populations involved and their star formation histories. Until individual $z \simeq 5$–6 sources can be studied in more detail, perhaps via the location of one or two strongly lensed examples, or via future more powerful facilities, this will regrettably remain the situation. At present, such density estimates are unlikely to be accurate to better than a factor of 2. Even so, they provide good evidence for significant earlier star formation ([230], Lecture 5).

6.7 Lecture Summary

In this lecture we have discussed the great progress made in using v, i, z and J band Lyman dropouts to probe the abundance of star forming galaxies over

$3 < z < 10$. At redshifts $z \simeq 6$ alone, [30] discuss the properties of a catalog of 506 sources to $z_{\rm AB} = 29.5$.

In practice, the good statistics are tempered by uncertain contamination from foreground cool stars and dusty or passively-evolving red $z \simeq 2$ galaxies and the vagaries of cosmic variance in the small fields studied. It may be that we will not overcome these difficulties until we have larger ground-based telescopes.

Nonetheless, from the evidence at hand, it seems that the comoving UV luminosity density declines from $z \simeq 3$ to 10, and that only by appealing to special circumstances can the low abundance of star forming galaxies at $z > 6$ be reconciled with that necessary to reionize the Universe.

One obvious caveat is our poor knowledge of the contribution from lower luminosity systems. Some authors [30, 261] have suggested a steepening of the luminosity function at higher redshift. Testing this assumption with lensed searches is the subject of our next lecture.

Finally, we have seen the successful emergence of the Spitzer Space Telescope as an important tool in confirming the need for star formation at $z > 6$. Large numbers of $z \simeq 5$–6 galaxies have now been detected by IRAC. The prominent Balmer breaks and high stellar masses argue for much earlier activity. Reconciling the present of *mature* galaxies at $z \simeq 6$ with the absence of significant star formation beyond, is one of the most interesting challenges at the present time.

7 Lyman Alpha Emitters and Gravitational Lensing

7.1 Strong Gravitational Lensing – A Primer

Slowly during the twentieth century, gravitational lensing moved from a curiosity associated with the verification of General Relativity [72] to a practical tool of cosmologists and those studying distant galaxies. There are many excellent reviews of both the pedagogical aspects of lensing [25, 168, 190] and a previous Saas-Fee contributor [207].

To explore the distant Universe, we are primarily concerned with *strong* lensing – where the lens has a projected mass density, above a critical value, $\Sigma_{\rm crit}$, so that multiple images and high source magnifications are possible. For a simple thin lens

$$\Sigma_{\rm crit} = \frac{c^2\, D_{\rm OS}}{4\,\pi\, G\, D_{\rm OL}\, D_{\rm LS}} \tag{35}$$

where D represents the angular diameter distance and the subscripts O, L, S refer to the observer, source and lens, respectively. Rather conveniently, for a lens at $z \simeq 0.5$ and a source at $z > 2$, the critical projected density is about $1\,{\rm g\,cm^{-2}}$ – a value readily exceeded by most massive clusters. The merits of exploring the distant Universe by imaging through clusters was sketched in a remarkably prophetic article by [268].

In lensing theory it is convenient to introduce a *source plane*, the true sky, and an *image plane*, the detector at our telescope, where the multiple images are seen. The relationship between the two is then a mapping transformation which depends on the relative distances (above). Crucially, what the observer sees depends on the degree of alignment between the source and lens as illustrated in Fig. 46.

An elliptical lens with $\Sigma > \Sigma_{\rm crit}$ for a given source and lens distance produces a pair of *critical lines* in the image plane where the multiple images

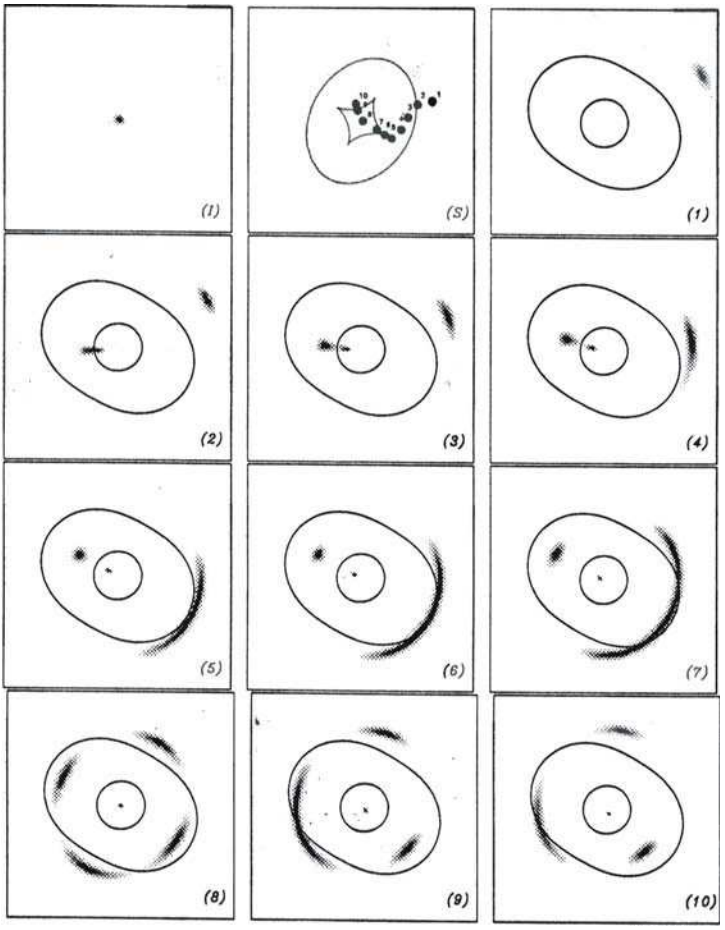

Fig. 46. Configurations in the image plane for an elliptical lens as a function of the degree of alignment between the source and lens (second panel). Lines in the source plane refer to "caustics" which map to "critical lines" in the image plane (see text for details) (Courtesy of Jean-Paul Kneib)

lie. These lines map to *caustics* in the source plane. The outer critical line is equivalent to the Einstein radius $\theta_{\rm E}$

$$\theta_{\rm E} = \sqrt{\frac{4\,G\,M\,D_{\rm OS}}{c^2\,D_{\rm OL}\,D_{\rm LS}}} \tag{36}$$

and, for a given source and lens, is governed by the enclosed mass M. The location of the inner critical line depends on the *gradient* of the gravitational potential ([198]).

The critical lines are important because they represent areas of sky where very high magnifications can be encountered – as high as $\times 30$! For $\simeq 20$ well-studied clusters the location of these lines can be precisely determined for a given source redshift. Accordingly, it is practical to survey just those areas to secure a glimpse of otherwise inaccessibly faint sources boosted into view. The drawback is that, as in an optical lens, the sky area is similarly magnified, so the surface density of faint sources must be very large to yield any results. Regions where the magnification exceed $\times 10$ are typically only 0.1–$0.3\,{\rm arcmin}^2$ per cluster in extent in the image plane and inconveniently shaped for most instruments (Fig. 47). The sampled area in the source plane is then ten times smaller so to see even one magnified source/cluster requires a surface density of *distant* sources of $\sim 50\,{\rm arcmin}^{-2}$.

Two other applications are particularly useful in faint galaxy studies. Firstly, strongly magnified systems at $z \simeq 2$–3 can provide remarkable insight into an already studied population by providing an apparently bright galaxy which is brought within reach of superior instrumentation. cB58, a Lyman break galaxy at $z = 2.72$ boosted by $\times 30$ to $V = 20.6$ ([203, 264]) was the first distant galaxy to be studied with an echellette spectrograph ([186]), yielding chemical abundances and outflow dynamics of unprecedented precision.

More generally, a cluster can magnify a larger area of $\simeq 2$–$4\,{\rm arcmin}^2$ by a modest factor, say $\times 3$–5. This has been effective in probing sub-mm source counts to the faintest possible limits ([215]) and the method shows promise for similar extensions with the IRAC camera onboard Spitzer.

7.2 Creating a Cluster Mass Model

In the applications discussed above, in order to analyze the results, the inferred magnification clearly has to be determined. This will vary as a function of position in the cluster image and the relative distances of source and lens. The magnification follows from the construction of a *mass model* for the cluster.

The precepts for this method are discussed in the detailed analysis of the remarkable image of Abell 2218 taken with the WFPC-2 camera onboard HST in 1995 ([135]). An earlier image of AC 114 showed the important role HST would play in the recognition of multiple images ([214]). Prior to HST, multiple images could only be located by searching for systems with similar

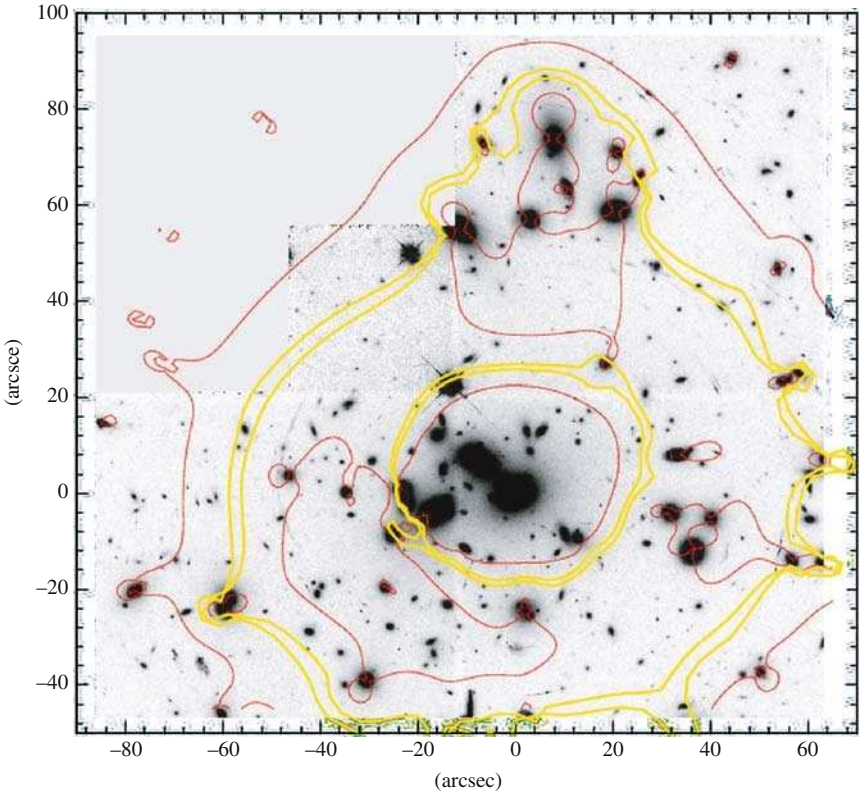

Fig. 47. Hubble Space Telescope image of the rich cluster Abell 1689 with the critical lines for a source at $z \simeq 5$ overlaid in yellow. The narrow regions inbetween the pairs of yellow lines refer to regions where the magnification exceeds $\times 10$

colors, using the fact that lensing is an achromatic phenomenon. HST revealed that morphology is a valuable additional identifier; the improved resolution also reveals the local shear (see Fig. 48).

Today, various approaches are possible for constructing precise mass models for lensing clusters [35, 135, 119]. These are generally based on utilizing the geometrical positions of sets of multiply-imaged systems whose redshift is known or assumed. This then maps the form and diameter of the critical line for a given z. Spectroscopic redshifts are particularly advantageous, as are pairs that straddle the critical line whose location can then be very precisely pinpointed. A particular mass model can be validated by "inverting" the technique and predicting the redshifts of other pairs prior to subsequent spectroscopy [71].

The main debate among cognescenti in this area lies in the extent to which one should adopt a parametric approach to fitting the mass distribution, particularly in relation to the incorporation of mass clumps associated

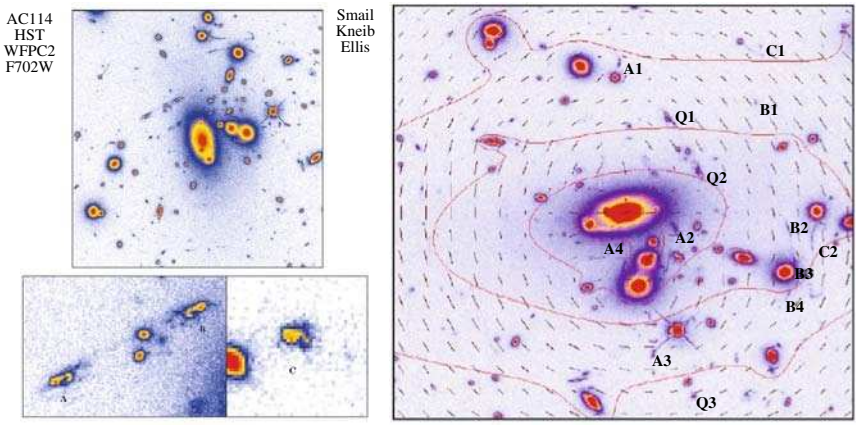

Fig. 48. Hubble Space Telescope study of the rich cluster AC114 (Smail et al. 1995 [214]) (**Top**) Morphological recognition of a triply-imaged source. The lower inset panels zoom in on each of 3 images of the same source. (**Bottom**) Construction of mass contours (red lines) and associated shear (red vectors) from the geometrical arrangement of further multiple images labelled A1–3, B1–4, C1–3, Q1–3

with individual cluster galaxy halos [35]. Reference [231] discuss the likely uncertainties in the mass modeling process arising from the various techniques.

7.3 Lensing in Action: Some High z Examples

Before turning to Lyman α emitters (lensed and unlensed), we will briefly discuss what has been learned from strongly-lensed dropouts.

Figure 49 shows a lensed pair in the cluster Abell 2218 ($z = 0.18$) as detected by NICMOS onboard HST and the two shortest wavelength channels of IRAC [73, 136]. Although no spectroscopic redshift is yet available for this source, three images have been located by HST and their arrangement around the well-constrained $z = 6$ critical line suggests a source beyond $z \simeq 6$ ([136]).

As with the unlensed i-band drop out studies by [79] and [263], the prominent IRAC detections ([73]) permit an improved photometric redshift and important constraints on the stellar mass and age. A redshift of $z = 6.8 \pm 0.1$ is derived, independently of the geometric constraints used by Kneib et al. The stellar mass is $\simeq (5 - 10) \times 10^8\,\mathrm{M}_\odot$ and the current star formation rate is $\simeq 2.6\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$. The luminosity-weighted age corresponds to anything from 40 to 450 Myr for a normal IMF depending on the star formation history. Interestingly, the derived age for such a prominent Balmer break generally exceeds the e-folding timescale of the star formation history (Fig. 49) indicating the source would have been more luminous at redshifts $7 < z < 12$ (unless obscured).

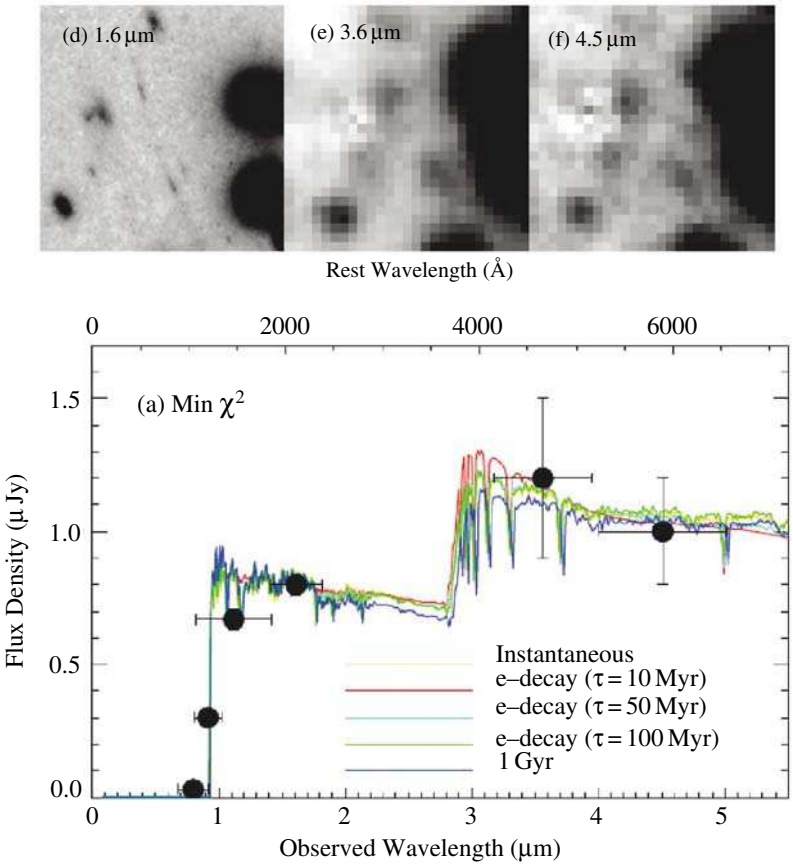

Fig. 49. (**Top**) Lensed pair of a $z = 6.8$ source as seen by NICMOS and IRAC in the rich cluster Abell 2218 Egami et al. (2005) [73]. The pair straddles the critical line at $z \simeq 6$ and a fainter third image at a location predicted by the lensing model has been successfully recovered in the HST data (Kneib et al. 2004 [136]). (**Bottom**) Spectral energy distribution of the source revealing a significant Balmer break and improved estimates of the star formation rate, stellar mass and luminosity weighted age

Given the small search area used to locate this object, such low mass sources may be very common. Accordingly, several groups are now surveying more lensing clusters for further examples of z band dropouts and even J-band dropouts (corresponding to $z \simeq 8$–10). Reference [191] are surveying 6 clusters with NICMOS and IRAC with deep ground-based K band imaging from Subaru and Keck. In these situations, one has to distinguish between magnifications of $\times 5$ or so expected across the 2–3 arcmin fields of NICMOS and IRAC, and the much larger magnifications possible close to the critical lines. Contamination from foreground sources should be similar to what is seen in the GOODS surveys discussed in Lecture 6. The discovery of image

pairs in the highly-magnified regions would be a significant step forward since spectroscopic confirmation of any sources at the limits being probed ($H_{\rm AB} \simeq 26.5 - -27.0$) will be exceedingly difficult.

7.4 Lyman Alpha Surveys

The origin and characteristic properties of the Lyman α emission line has been discussed by my colleagues in their lectures (see also [7, 105, 170, 200]). The $n = 2$ to 1 transition corresponding to an energy difference of 10.2 eV and rest-wavelength of $\lambda 1216$ Å typically arises from ionizing photons absorbed by nearby hydrogen gas. The line has a number of interesting features which make it particularly well-suited for locating early star forming galaxies as well as for characterizing the nature of the IGM.

In searching for distant galaxies, emission lines offer far more contrast against the background sky than the faint stellar continuum of a drop-out. A line gives a convincing spectroscopic redshift (assuming it is correctly identified) and models suggest that as much as 7% of the bolometric output of young star-forming region might emerge in this line. For a normal IMF and no dust, a source with a star formation rate (SFR) of $1\,{\rm M}_\odot\,{\rm yr}^{-1}$ yields an emission line luminosity of $1.5\ 10^{42}\,{\rm ergs\,sec}^{-1}$.

Narrow band imaging techniques (see below) can reach fluxes of $< 10^{-17}$ cgs in comoving survey volumes of $\simeq\ 10^5\,{\rm Mpc}^3$, corresponding to a SFR $\simeq 3\,{\rm M}_\odot\,{\rm yr}^{-1}$ at $z \simeq 6$. Spectroscopic techniques can probe fainter due to the improved contrast. This is particularly so along the critical lines where the additional boost of gravitational lensing enables fluxes as faint as 3×10^{-19} cgs to be reached (corresponding to SFR $\simeq 0.1\,{\rm M}_\odot\,{\rm yr}^{-1}$). However, in this case the survey volumes are much smaller ($\simeq\ 50\,{\rm Mpc}^3$). In this sense, the two techniques (discussed below) are usefully complementary.

Having a large dynamic range in surveys for Lyα emission is important not just to probe the luminosity function of star-forming galaxies but also because it can be used to characterize the IGM. As a resonant transition, foreground hydrogen gas clouds can scatter away Lyα photons in both direction and frequency. In a partially ionized IGM, scattering is maximum at $\lambda 1216$ Å in the rest-frame of the foreground cloud, thus affecting the *blue* wing of the observed line. However, in a fully neutral IGM, scattering far from resonance can occur leading to damping over the entire observed line. Figure 50 illustrates how, in a hypothetical situation where the IGM becomes substantially neutral during $6 < z < 7$, surveys reaching the narrow-band flux limit would still find emitters at $z = 7$. Their intense emission would only be partially damped by even a neutral medium. However, lines with fluxes at the spectroscopic lensing limit would not survive. Accordingly, one possible signature of reionization would be a significant change in the shape of the Lyα luminosity function at the faint end ([95]).

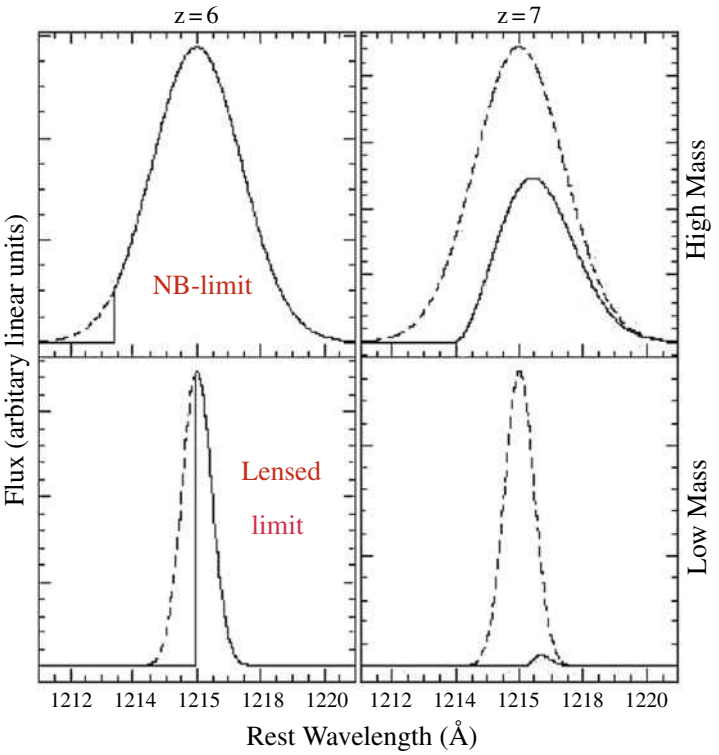

Fig. 50. The Lyman α damping wing is absorbed by neutral hydrogen and thus can act as a valuable tracer of the nature of the IGM. The simulation demonstrates the effect of HI damping on emission lines in high mass and low mass systems (characteristic of sources detected in typical narrow band and lensed spectroscopic surveys respectively) assuming reionization ends inbetween $z = 6$ and 7. The dramatic change in visibility of the weaker systems suggests their study with redshift may offer a sensitive probe of reionization (Courtesy: Mike Santos)

7.5 Results from Narrow Band Lyα Surveys

The most impressive results to date have come from various narrow-band filters placed within the SuPrime camera at the prime focus of the Subaru 8 m telescope [114, 117, 127, 137, 179, 212, 243]. Important conclusions have also been deduced from an independent 4 m campaign ([159]).

Narrow band filters are typically manufactured at wavelengths where the night sky spectrum is quiescent, thus maximizing the contrast. These locations correspond to redshifts of $z = 4.7$, 5.7, 6.6 and 6.9 (Fig. 52). A recent triumph was the successful recovery of two candidates at $z \simeq 6.96$ by [117]. Candidates are selected by comparing their narrow band fluxes with that in a broader band encompassing the narrow band wavelength range. The contrast can then be used as an indicator of line emission (Fig. 53). Spectroscopic follow-up is still

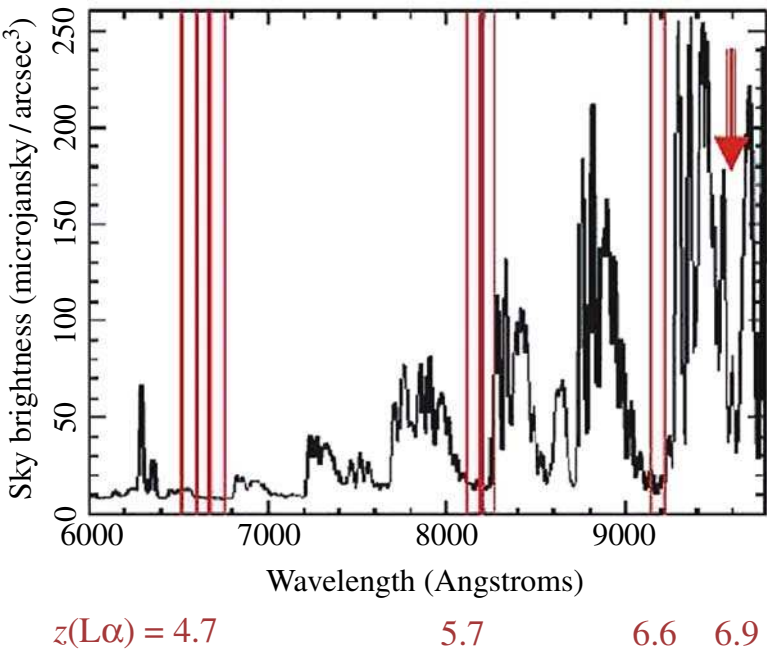

Fig. 51. Night sky spectrum and the deployment of narrow band filters in "quiet" regions corresponding to redshifted Lyman α emission as indicated below. The final optical window, corresponding to $z \simeq 6.9$, was successfully exploited by Iye et al. (2006) [117] to find two sources close to $z \simeq 7$

desirable as the line could arise in a foregound galaxy with [O II] 3727 Å or [O III] 5007 Å emission. The former line is a doublet and the latter is part of a pair with a fixed line ratio, separated in the rest-frame by only 60 Å or so, so these contaminants are readily identified. Furthermore, Lyα is often revealed by its asymmetric profile (c.f. Fig. 51).

Spectroscopic follow-up is obviously time-intensive for a large sample of candidates, hundreds of which can now be found with panoramic imagers such as SuPrimeCam. Therefore it is worth investigating additional ways of eliminating foreground sources. Tanuguchi et al. (2005) combine the narrow band criteria adopted in Fig. 53 with a broad-band $i - z$ drop-out signature. Spectroscopic follow-up of candidates located via this double color cut revealed a 50–70% success rate for locating high z emitters. The drawback is that the sources so found cannot easily be compared in number with other, more traditional, methods. Reference [173] used a narrow – broad band color criterion in the opposite sense, locating sources with a narrow band *depression* (rather than excess). Such rare sources are confirmed to be sources with extremely intense emission elsewhere in the broad-band filter. Such sources, with Lyman α equivalent widths in excess of several hundred Å are interesting because they may challenge what can be produced from normal stellar populations.

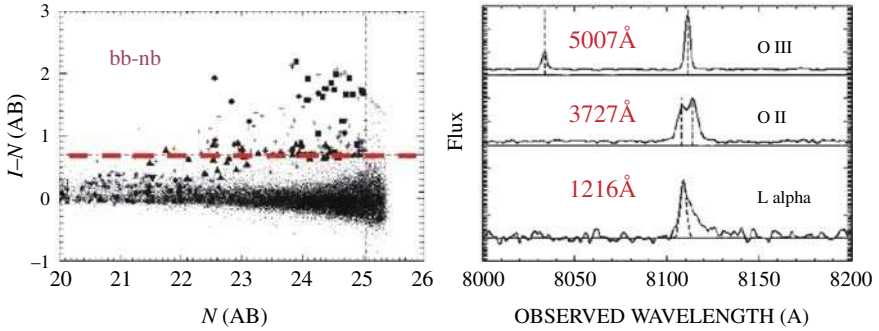

Fig. 52. The two-step process for locating high redshift Lyman α emitters (Hu et al. 2004 [114]). (**Left**) Comparison of broad and narrow band magnitudes; sources with an unusual difference in the sense of being brighter in the narrow band filter represent promising candidates. (**Right**) Spectroscopic follow-up reveals typically three possibilities – [O III] or [O II] at lower redshift, or Lyα often characterized by its asymmetric line profile

Reference [159] were the first to consider the absence of evolution in the Lyα LF as a constraint on the neutral fraction. Although they found no convincing change in the LF between $z = 5.7$ and 6.5, the statistical uncertainties in both LFs were considerable. Specifically, below luminosities of $L_{\mathrm{Ly}\alpha} \simeq 10^{42.5}\,\mathrm{ergs\,sec^{-1}}$ no detections were then available. Reference [114] have also appealed to the absence of any significant change in the mean Lyα profile. Malhotra & Rhoads deduced the neutral fraction must be $x_{\mathrm{HI}} < 0.3$ at $z \simeq 6$ supporting early reionization. However, [95] reanalyzed this constraint and indicated that strong emitters could persist even when $x_{\mathrm{HI}} \simeq 0.5$ (c.f. Fig. 51).

References [127] and [212] have determined statistically greatly improved Lyman α luminosity functions and discuss both spectroscopic confirmed and photometrically-selected emitters (Fig. 53). No decline is apparent in the abundance of low luminosity emitters as expected in an IGM with high x_{HI}; indeed the most significant change is a decline with redshift in the abundance of the most luminous systems. Although the change seems surprisingly rapid given the time interval is only 150 Myr, this is consistent with growth in the halo mass function (Dijsktra et al. 2006).

7.6 Results from Lensed Lyα Surveys

The principal gain of narrow band imaging over other techniques in locating high redshift Lyman α emitters lies in the ability to exploit panoramic cameras with fields of view as large as 30–60 arcmin. Since cosmic lenses only magnify fields of a few arcmin or less, lensing searches are only of practical utility when used in spectroscopic mode. As discussed above, the gain in sensitivity can be factors of $\times 30$ or more, and given the small volumes explored, they

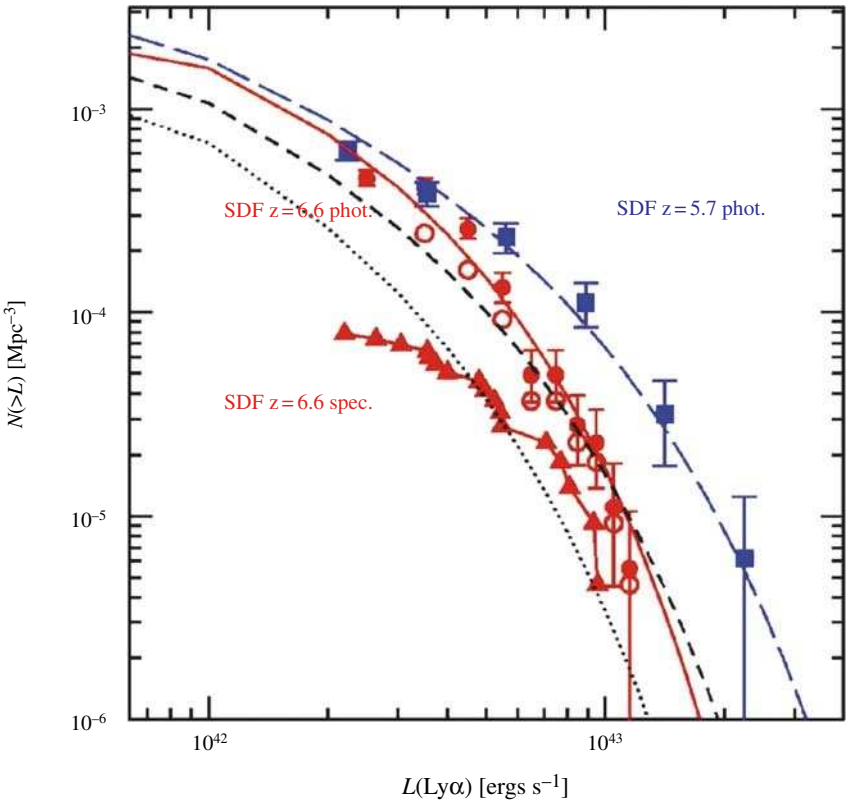

Fig. 53. Comparison of the Lyman α luminosity functions at $z = 5.7$ and 6.5 from the surveys of Kashikawa et al. (2006) [127] and Shimasaku et al. (2005) [212]; both spectroscopically-confirmed and photometric candidates are plotted. The decline in luminous emitters is qualitatively similar to trends seen over $3 < z < 6$ in luminous continuum drop-outs

are primarily useful in testing the faint end of the Lyα luminosity function at various redshifts. A number of workers (e.g. Barkana & Loeb 2004) have emphasized the likelihood that the bulk of the reionizing photons arise from an abundant population of intrinsically-faint sources, and lensed searches provide the only practical route to observationally testing this hypothesis.

A practical demonstration of a blind search for lensed Lyα emitters is summarized in Fig. 54. A long slit is oriented along a straight portion of the critical line (whose location depends on the source redshift). The survey comprises several exposures taken in different positions offset perpendicular to the critical line. Candidate emission lines are astrometrically located on

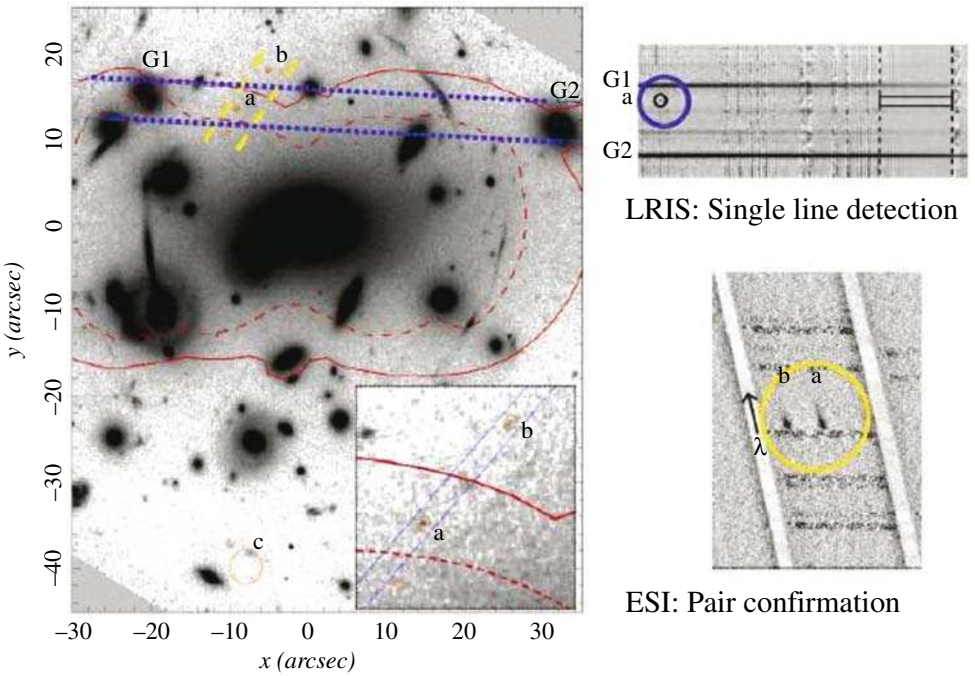

Fig. 54. Critical line mapping in Abell 2218: How it works Ellis et al. (2001) [77]. The red curves show the location of the lines of very high magnification for a source at a redshift $z = 1$ (*dashed*) and $z = 6$ (*solid*). Blue lines show the region scanned at low resolution with a long-slit spectrograph. The upper right panel shows the detection of an isolated line astrometrically associated with (*a*) in the HST image for which a counter image (*b*) is predicted and recovered (*see also inset to main panel*). A higher dispersion spectrum aligned between the pair (*yellow lines*) reveal strong emission with an asymmetric profile in both (*lower right panel*)

a deep HST image and, if a counter-image consistent with the mass model can be located, a separate exposure is undertaken to capture both (as was the case in the source located by [77]). Unfortunately, continuum emission is rarely seen from a faint emitter and the location of a corresponding second image is often too uncertain to warrant a separate search. In this case contamination from foreground sources has to be inferred from the absence of corresponding lines at other wavelengths ([201, 230]).

Using this technique with an optical spectrograph sensitive to Lyα from $2.2 < z < 6.7$, [201] conducted a survey of 9 lensing clusters and found 11 emitters probing luminosities as faint as $L_{Ly\alpha} \simeq 10^{40}$ cgs, significantly fainter than even the more recent Subaru narrow band imaging searches ([212]). The resulting luminosity function is flatter at the faint end than implied for the halo mass function and is consistent with suppression of star formation in the lowest mass halos.

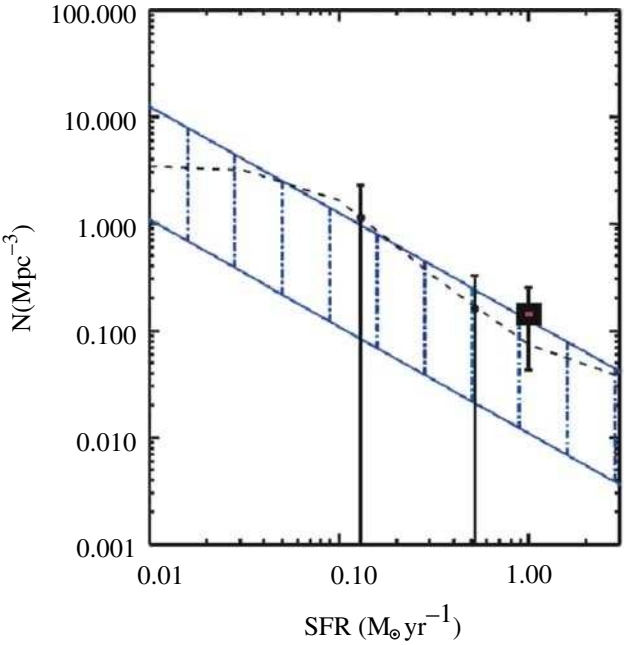

Fig. 55. The volume density of sources of various star formation rates at $z \simeq 8$–10 required for cosmic reionization for a range of assumed parameters (blue hatched region) compared to the inferred density of lensed emitters from the survey of Stark et al. (2006b) [231]. The open red symbol corresponds to the case if all detected emitters are at $z \simeq 10$, the black symbols correspond to the situation if the two most promising candidates are at $z \simeq 10$, and the dashed line corresponds to the 5σ upper limit if none of the candidates is at $z \simeq 10$

Reference [231] have extended this technique to higher redshift using an infrared spectrograph operating in the J band, where lensed Lyα emitters in the range $8.5 < z < 10.2$ would be found (Fig. 55). This is a much more demanding experiment than that conducted in the optical because of the brighter and more variable sky brightness, the smaller slit length necessitating very precise positioning to maximize the magnifications and, obviously, the fainter sources given the increased redshift. Nonetheless, a 5σ sensitivity limit fainter than 10^{-17} cgs corresponding to intrinsic (unlensed) star formation rates of $0.1\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$ is achieved with the 10 m Keck II telescope in exposure times of 1.5 hr per slit position.

After surveying 10 clusters with several slit positions per cluster, 6 candidate emission lines have been found and, via additional spectroscopy, it seems most cannot be explained as foreground sources. Stark et al. estimate the survey volume taking into account both the spatially-dependent magnification (from the cluster mass models) and the redshift-dependent survey sensitivity (governed by the night sky spectrum within the spectral band).

References [157] and [233] have introduced simple prescriptions for estimating the abundance of star forming sources necessary for cosmic reionizations. While these prescriptions are certainly simple-minded, given the coarse datasets at hand, they provide an illustration of the implications.

Generally, the abundance of sources of a given star formation rate SFR necessary for cosmic reionization over a time interval Δt is

$$n \propto \frac{B\, n_H}{f_c\, SFR\, \Delta t} \tag{37}$$

where B is the number of ionizing photons required to keep a single hydrogen atom ionized, n_H is the comoving number density of hydrogen at the redshift of interest and f_c is the escape fraction of ionizing photons. Figure 56 shows the upshot of the [231] survey for various assumptions. The detection of even a few convincing sources with SFR $\simeq$ 0.1–1 $M_\odot\,yr^{-1}$ in such small cosmic volumes would imply a significant contribution from feeble emitters at $z \simeq 10$. Although speculative at this stage given both the uncertain nature of the lensed emitters and the calculation above, it nonetheless provides a strong incentive for continued searches.

7.7 Lecture Summary

In this lecture we have shown how Lyman α emission offers more than simply a way to locate distant galaxies. The distribution of line profiles, equivalent widths and its luminosity function can act as a sensitive gauge of the neutral

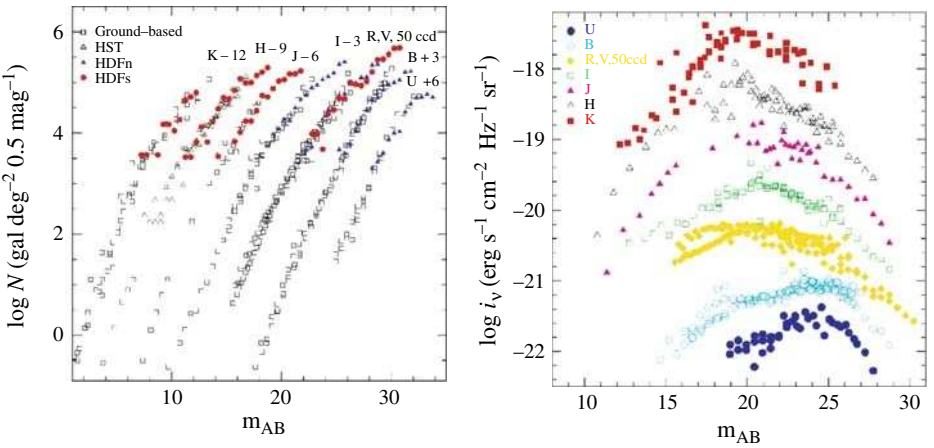

Fig. 56. (**Left**) Differential galaxy counts as a function of wavelength from the compilation of Madau & Pozzetti (2000) [153]. Note the absolute numbers have been arbitrarily scaled for convenience. (**Right**) Magnitude-dependent contribution of the counts to the surface brightness i_ν of the extragalactic night sky. Depending on the waveband, the peak contribution occurs at $m_{AB} \simeq$ 20–25

fraction because of the effects of scattering by hydrogen clouds. Surveys have been undertaken using optical cameras and narrow band filters to redshifts $z \simeq 7$.

However, despite great progress in the narrow-band surveys, as with the earlier i-band drop outs, there is some dispute as to the evolutionary trends being found. Surprisingly strong evolution is seen in luminous emitters over a very short period of cosmic time corresponding to $5.7 < z < 6.5$. And, to date there is no convincing evidence that line profiles are evolving or that the equivalent width distribution of emitters is skewed beyond what can be accounted for by normal young stellar populations. One suspects we will have to push these techniques to higher redshift which will be hard given the Lyα line moves into the infrared where no such panoramic instruments are yet available.

We have also given a brief tutorial on strong gravitational lensing. In about 20 or so clusters, spectroscopic redshifts for sets of multiple images has enabled quite precise mass models to be determined which, in turn, enable accurate magnification maps to be derived. Remarkably faint sources can be found by searching along the so-called *critical lines* where the magnification is high. The techniques has revealed a few intrinsically faint sources and, possibly, the first glimpse of a high abundance of faint star forming sources at $z \simeq 10$ has been secured.

8 Cosmic Infrared Background

8.1 Motivation

In this lecture we examine the role of cosmic backgrounds. First we make an important distinction. We are primarily concerned with extragalactic backgrounds composed of unresolved faint sources rather than the fairly isotropic microwave radiation that comes from the recombination of hydrogen. Such unresolved source background have played a key role in astrophysics.

The pattern of discovery outside the optical and near-infrared spectral windows often goes like the following; the background is first discovered by sensitive detectors which do not have the angular resolution to see if there is any fine structure from faint sources. There is then some puzzlement as to its origin: for example, does it arise from an unforeseen population of sources beyond those already counted? In this manner, the X-ray background was identified as arising from active galactic nuclei and the sub-mm background from dusty star-forming galaxies. In each case we are concerned with separating the counts of resolved sources to some detectability limit with the measured value of the integrated background. Key issues relate to the contribution of resolved sources and the removal of spurious (non-extragalactic) foreground signals. Excellent reviews of the subject have been provided by [107, 146] and [130].

In our case, we are interested in extending source counts of star forming galaxies beyond $z \simeq 6$ and so the question of a near-infrared extragalactic background signal is of greatest relevance. The instruments concerned include cameras on Hubble Space Telescope and Spitzer Space Telescope, but other space missions such as COBE and the Infrared Telescope in Space have provided important results.

8.2 Methodology

To understand the distinction between the *resolved* and *unresolved* components of the background, it is helpful to derive some fundamental relations for galaxy counts.

In the magnitude system, the differential count slope γ with increasing magnitude is

$$\gamma = \frac{\mathrm{d}\ \log\ N(m)}{\mathrm{d}m} \tag{38}$$

where $N(m)$ is the differential number of galaxies per unit sky area (e.g. deg^{-2}) in some counting bin $\mathrm{d}m$.

Now the contribution to the surface brightness of the extragalactic night sky from sources as a given magnitude m is:

$$i_\nu(m) = 10^{-0.4(m+const)}\, N(m) \tag{39}$$

where the 0.4 factor arises from the relationship between flux and magnitude.

And the integrated surface brightness (the extragalactic background light, EBL) is obtained by extending this integral to infinitely faint limits, viz:

$$I_\nu = \int_{-\infty}^{\infty} i_\nu(m)\, \mathrm{d}m \tag{40}$$

The bolometric equivalent of the EBL is then $\int I_\nu\, \mathrm{d}\nu$. The EBL is often alternatively expressed as $\mathrm{EBL} = \int \nu\, I_\nu\, \mathrm{d}\nu$.

Surface brightness is usually expressed in $\mathrm{nW\,m^{-2}\,steradian^{-1}}$ although very early articles refer to some derivative of magnitudes deg^{-2}. As infrared fluxes are often expressed in Janskies ($1\,\mathrm{Jansky} = 10^{-26}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$, a useful conversion is $1\,\mathrm{nw\,m^{-2}\,ster^{-1}} = 3000/\lambda(\mu\,\mathrm{m})\ \mathrm{MJy\ ster^{-1}}$.

Examination of the above relation shows that if $\gamma > 0.4$, the surface brightness contribution from fainter sources outshines that of brighter ones and so the EBL will diverge. By deduction, therefore, the maximum contribution of resolved sources to the background will be where $\gamma \simeq 0.4$. If the slope of the counts turns down below 0.4 at some point, it may seem pointless to speculate that there is much information from fainter unresolved sources. In the case of searching for cosmic reionization however, where the first sources may be a distinct population, continuity in the source counts may not be expected.

For this reason, the interesting quantity is the difference between a measured EBL and the integrated contribution from the faintest resolved sources.

Reference [153] present a careful analysis of the deep optical and near-infrared counts at various wavelengths, mostly from Hubble Space Telescope data prior to the Ultra Deep Field (Fig. 56). Extrapolation of the counts enables the contribution to the integrated light from *known populations* to be evaluated as a function of wavelength. The total EBL from this analysis is $55\,\mathrm{nW\,m^{-2}\,sr^{-1}}$ of which the dominant component lies longward of 1μm. However, it must be remembered that galaxies are extended objects and so it is possible that significant light is lost in the outermost regions of each. As surface brightness is relativistically dimmed by the cosmic expansion, $\propto (1+z)^4$, the contribution from distant sources could be seriously underestimated ([145]).

In a similar fashion to the check we made that the integrated star formation history produces the present stellar mass density (Lecture 4), so it is possible to verify that the bolometric output from stellar evolution should be consistent with the present mass density of stars.

The bolometric radiation density $\rho_{\mathrm{bol}}(t)$ is

$$\rho_{\mathrm{bol}}(t) = \int_0^t L(\tau)\dot{\rho_{\mathrm{s}}}(t-\tau)\mathrm{d}\,\tau \tag{41}$$

where $\dot{\rho_s}$ is the star formation rate per comoving volume.

The integrated EBL is then

$$I_{\mathrm{EBL}} = \frac{c}{4\pi}\int_{t_{\mathrm{F}}}^{t_{\mathrm{H}}} \frac{\rho_{\mathrm{bol}}(t)}{1+z}\,\mathrm{d}t \tag{42}$$

The check is a little bit trickier because a typical star formation history has to be assumed and presumably there is a large variety for different kinds of sources observed at various redshifts. The issue is discussed in detail by [153].

8.3 Recent Background Measurements

The goal for making extragalactic background light (EBL) measures is thus to determine the extent to which the measured value (or limit) exceeds that predicted from extrapolation of the galaxy counts. This might then provide some evidence for a new distant population such as the sources responsible for cosmic reionization.

Various claims of an excess have been made in the 0.3–10 μm wavelength region (Fig. 57). Outside of this window the background appears to be entirely produced by known sources. Such EBL measurements are extremely difficult to make for several reasons. An accurate absolute calibration is essential since the interesting signal is a "DC difference" in surface brightness. The removal

of spurious foregrounds is likewise troublesome: at some wavelengths, the foreground signal greatly exceeds the sought-after effect.

In the optical window, careful experiments have been undertaken by [69, 163] and most recently [22]. The most vexing foreground signals at optical wavelengths arise from airglow (emission in the upper atmosphere which has a time-dependent structure on fine angular scales), zodiacal light (scattering of sunlight by interplanetary dust which varies with the motion of the Earth along the ecliptic plane) and diffuse Galactic light or "cirrus" (gas clouds illuminated by starlight). The wavelength dependence of these foregrounds (including the microwave background itself which is considered a "contaminant" in this respect!) is summarized in Fig. 58. For the critical near-infrared region where redshifted light from early sources might be detected, airglow and zodiacal light are the dominant contaminants.

Experiments have differed in the way these foregrounds might be removed. The use of a space observatory avoids airglow and observations at different Galactic latitudes can be used to monitor or minimize the effect of cirrus. Zodiacal light, the dominant foreground for all mid-infrared studies and all space-based near-infrared studies, is the most troublesome.

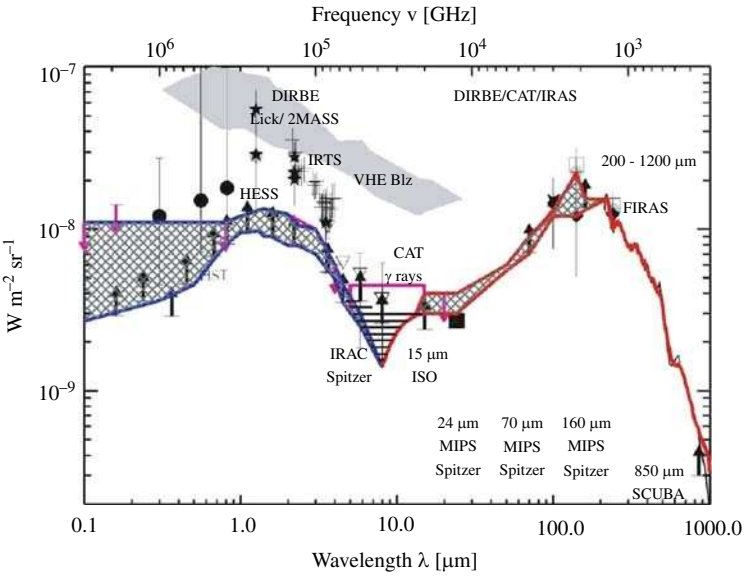

Fig. 57. Summary of recent measures of the extragalactic background light from the compilation of Dole et al. (2006) [63]. The cross-hatched region represents the region of claimed excess between the various background meaasures (labelled points) and the lower limits to the integrated counts (shown as upward arrows in the critical optical-near IR region). Interesting excesses are found in the 0.3–5 μm region

To first order, the zodiacal light can be predicted depending on the geometrical configuration of the ecliptical plane, the earth, sun and target field. Although seasonal variations can be tracked and the spectrum of the zodiacal light is solar, imponderables enter such as the nature and distribution of the interplanetary dust and its scattering properties.

The positive EBL excess measures at optical to near-infrared wavelengths have remained controversial. This is because all such experiments search for a "DC signal" – a small absolute difference between two signals within which foregrounds have to be painstakingly removed. A brief study of the [22] experiment will make the difficulties clear.

Reference [22] claim a significant optical excess at 300, 550 and 800 nm from fields observed with Hubble's WFPC-2 and FOS. Airglow is elminated since HST is above 90 km and sources were removed in each HST image to $V_{\rm AB} = 23.0$. The zodiacal light spectrum was measured simultaneously with a ground-based optical telescope and iteratively subtracted from the HST data.

At 550 nm, the total WFPC-2 background measured after removing all sources was 105.7 ± 0.3 units ($1\,\text{unit} = 10^{-9}\,\text{cgs ster}^{-1}\,\text{Å}^{-1}$). The measured zodi background was 102.2 ± 0.6 units. Galactic cirrus and faint galaxies beyond $V_{\rm AB} = 23.0$ were estimated as 0.8 and 0.5 units respectively. The claimed excess signal at this wavelength is 2.7 ± 1.4 units. Not only is this only a 2σ

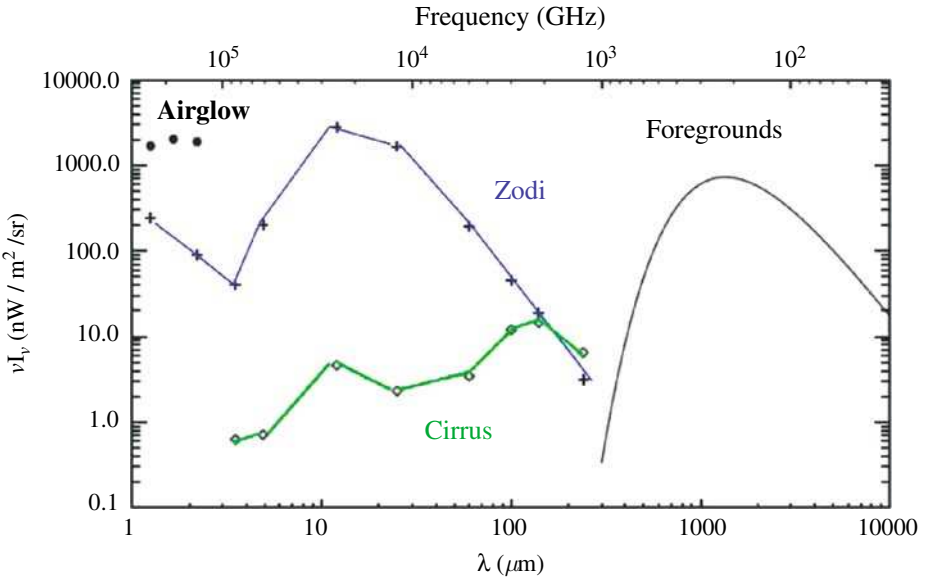

Fig. 58. Wavelength dependence of the dominant foreground signals from the compilation by Kashlinsky (2005) [128]. Airglow and zodiacal light dominate at the wavelengths of interest for stellar radiation from highly redshift sources

detection but it would only require a 2% error in the estimated Zodiacal light signal to be spurious.

Studies of the infrared background advanced significantly with the launch of the COBE satellite which carried DIRBE – a 10 channel photometer operating in the 1–240 μm range with 0.7 degree resolution chopping at 32 Hz onto an internal zero flux surface, and FIRAS – an absolute spectrometer with 7 degree resolution operating in the 100 μm–5 mm range.

DIRBE measured the integrated background at 140 and 240 μm using an elaborate time-dependent model of the Zodiacal light [108, 132]. Reference [206] combined these data with higher resolution IRAS 100 μm maps to improve removal of Galactic cirrus and to measure the temperature of the dust emission. Reference [87] extended the model of Zodiacal light to provide the first detection at 60 and 100 μm and [260] used 2MASS observations to improve Galactic source removal claiming a detection at 2.2 μm.

Reference [63] have undertaken a very elegant analysis of 19,000 Spitzer MIPS images. By centering the images on deep 24 μm sources, they can evaluate the statistical contribution of otherwise inaccessibly faint 70 and 160 μm sources.

Reference [166] has extended these detections to shorter wavelengths using a spectrometer onboard the Infrared Telescope in Space (IRTS). This signal has the tantalising signature of a distant star-forming stellar population – a steeply-rising continuum down to 1 μm and a discontinuity when extended to include optical data (Fig. 57). Reference [165] has argued this signal represents an artefact arising from incorrect foreground removal. He has likewise criticised the [22] optical detections ([164]).

Supposing the DIRBE/2MASS/IRTS detections to be real, what would this imply? Assuming the J-band background ($> 2.5\,\mathrm{nW\,m^{-2}\,ster^{-1}}$) arises from $z \simeq 9$ Lyα emission, [154] calculate the associated stellar mass that is produced. They find an embarrassingly high production rate, corresponding to $\Omega_* = 0.045\,\Omega_\mathrm{B}$; in other words almost all the stars we see today would need to be produced by $z \simeq 9$ to explain the signal. Likewise the ionizing flux produced by such star formation would be in excess of that required to explain the WMAP optical depth.

8.4 Fluctuation Analyses

References [129] and [130] have argued that the difficulties inherent in extracting the EBL signal by the DC method may be alleviated by considering the angular fluctuation spectrum $\delta\ F(\theta)$ and its 2-D Fourier transform $P_2(q)$.

In more detail, the fluctuation in the measured background is

$$\delta\,F(\theta) = F(\theta) - \langle F\rangle = (2\pi)^{-2} \int \delta\,F_\mathrm{q} \exp(=\mathrm{i}\,q\,.\,\theta)\mathrm{d}^2 q \qquad (43)$$

and

$$P_2(q) \equiv \langle |\delta F_q|^2 \rangle \tag{44}$$

The success of the method depends on whether the various foreground contributions to $P_2(q)$ can be readily distinguished from one another. Although advantageous in using independent information from the DC measurements, it is unfortunately not easy to intercompare the experiments or to interpret the fluctuation analyses.

Reference [130] recently applied this method to deep IRAC images. Sources were extracted to reasonably faint limits (0.3 μJy or $m_{\rm AB} \simeq$ 22–25). The pixels associated with each image were masked out in a manner that could be adjusted depending on the significance of the source over the background and its area. The data was split into two equal halves (A, B) and the power spectrum of the signal ($P_{\rm S}(q),\ S = A + B$) was compared to that of the noise ($P_{\rm N}(q),\ N = A - B$). Fig. 59 illustrates the residual signal, $P_{\rm S}(q) - P_{\rm N}(q)$, and the fluctuations $(q^2 P_2(q)/2\pi)^{\frac{1}{2}}$ in 3 IRAC fields as a function of angular scale ($2\pi/q$ arcsec) in the four IRAC channels. Although the positive fluctuations on small scales are consistent with shot noise from the galaxy counts (solid lines in both panels), the excess is particularly prominent on scales of 1-2

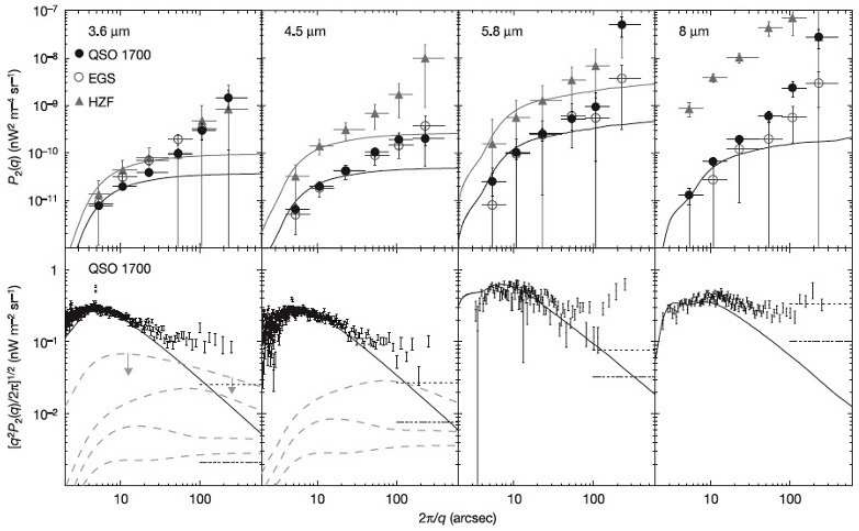

Fig. 59. Detection of fluctuations in the infrared background from the analysis by Kashlinky et al. (2005) [130]. The upper panel shows the angular power spectrum as a function of scale for 3 IRAC fields plotted for all four IRAC channels (3.6 through 8 μm). The solid line shows the effect of shot noise from the galaxy counts. This fits the fluctuations on small scales but there is a significant and consistent excess on large scales. Bottom panels refer to fluctuations for one IRAC field plotted as $(q^2 P_2(q)/2\pi)^{\frac{1}{2}}$ with the predicted contributions from various populations of unresolved distant sources shown as dotted lines

arcmin and consistently in fields of various Galactic latitudes and in all four IRAC bandpasses.

Kashlinksky et al. argue that the excess signal is fairly flat spectrally in νI_ν ruling out any instrumental or Galactic contamination. The constancy of the signal over the wide range in Galactic latitude likewise suggests the signal is extragalactic. However, zodiacal light is difficult to eliminate as a contaminant. Although the signal persists when the Earth is 6 months further around in its orbit, no analysis of the likely zodi fluctuation spectrum is presented.

If the signal persists, what might it mean for sources fainter than 0.3 μJy? Without a detailed redshift distribution, it is hard to be sure. The clustering pattern is remarkably strong given the depth of the imaging data (much deeper fields have subsequently been imaged but no independent analyses have yet been published). In summary, this is an intriguing result, somewhat unexplained and in need of confirmation[8].

8.5 EBL Constraints from TeV Gamma Rays

High energy (TeV) gamma rays are absorbed in the earth's atmosphere and converted into secondary particles forming an "air shower". Cerenkov light is generated – a beam of faint blue light lasting a few $\times 10^{-9}$ sec is produced illuminating an area of 250 m in diameter on the ground.

The significance of searching for TeV gamma rays is that they interact with the sea of 1–10 μm (infrared background) photons via the pair-creation process, viz:

$$\gamma\,\gamma \rightarrow e^+\,e^- \tag{45}$$

producing attenuation in distant sources (blazars) whose spectral energy distributions are assumed to be power laws.

$$\frac{\mathrm{d}\,N}{\mathrm{d}\,E} \propto E^{-\Gamma} \tag{46}$$

The strength of the attention, measured as a change in Γ, for the most distant accessible blazars is thus a measure of the degree of interaction between the gamma rays and the ambient infrared background.

The HESS team [2] have recently analyzed the gamma ray spectrum of a blazar at $z = 0.186$ and fitted its energy spectrum. As a result they can predict an upper limit to the likely infrared background for various assumptions (Fig. 60). The constraint is most useful in the wavelength range 1–10 μm where the associated gamma ray spectral data is of high quality (top axis of Fig. 60). This new method is intriguing. Although in detail the constraint is dependent on an assumed form for the TeV energy spectrum of blazars, for all reasonable assumptions the acceptable background is much lower at 1–4 μm than claimed by the traditional experiments.

[8] Recent developments are discussed by [239] and [131]

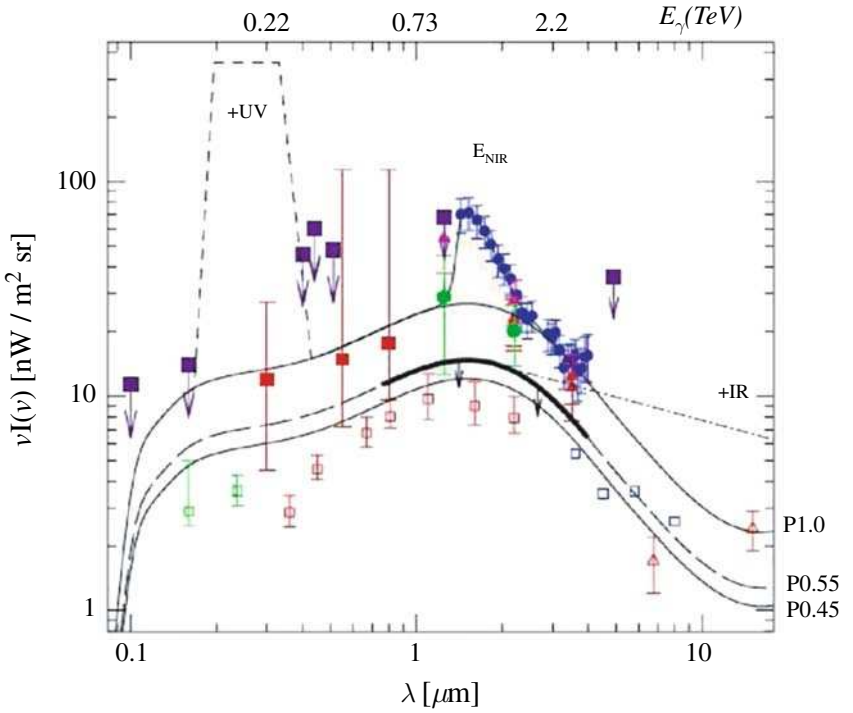

Fig. 60. Constraints on the cosmic infrared background from the degree of attenuation seen in the high energy gamma ray spectrum of a distant blazar Aharonian et al. (2006) [2]. The black curves indicate likely upper limits to the infrared background for various assumptions. This independent method argues the near-IR detections by Matsumoto et al. (2005) [166] (*blue circles*) and others are spurious

8.6 Lecture Summary

In this lecture we have learned that extragalactic background light measurements in principle offer an important and unique constraint on undiscovered distant source populations.

However, the observations remain challenging and controversial because of instrumental effects, related to the precision of absolute calibrations, and dominant foregrounds. In the 1–5 μm spectral regions positive detections in excess of extrapolations of the source counts have been reported. This is interesting because this is the wavelength range where an abundant and early population of faint, unresolved $z > 10$ sources would produce some form of infrared background.

At the moment, the claimed detections at 1–5 μm appear unreasonably strong. Too many stars would be produced by the star formation associated with this background signal and recent ultrahigh energy spectra of distant blazars likewise suggest a much weaker infrared background.

IRAC is a particularly promising instrument for the redshift range $10 < z < 20$ and fluctuation analyses offer a valuable independent probe of the background, although it is hard to interpret the results and compare them with the more
traditional DC level experiments.

Despite the seemlingly chaotic way in which the subject of the cosmic extragalactic background light has unfolded in the past decade, it is worth emphasizing that the motivation remains a strong one. Any information on the surface brightness, angular clustering and redshift range of radiation beyond the detected source counts will be very valuable information ahead of the discoveries made possible by future generations of facilities.

9 Epilogue: Future Prospects

9.1 Introduction

The students attending this lecture course are living at a special time! During your lifetime you can reasonably confidently expect to witness the location of the sources responsible for cosmic reionization, to determine the redshift range over which they were active and perhaps even witness directly the "cosmic dawn" when the first stellar systems shone and terminated the "Dark Ages"!

We have made such remarkable progress in the past decade, reviewed here, that such a bold prediction seems reasonable, even for a cautious individual like myself! We have extended our fundamental knowledge of how various populations of galaxies from $0 < z < 3$ combine to give us a broad picture of galaxy evolution, while extending the frontiers to $z \simeq 7$ and possibly beyond. Certainly many issues remain, including the apparent early assembly of certain classes of quiescent galaxy, the abundance patterns in the intergalactic medium and the apparent "downsizing" signatures seen over a variety of redshifts. However, the progress has been rapid and driven by observations. Accompanying this is a much greater synergy between theoretical predictions and observations than ever before.

We have seen that the question of "First Light" – the subject of this course – is the remaining frontier for observations of galaxy formation. The physical processes involved are poorly understood and thus observations will continue to be key to making progress. In this final lecture, I take out my crystal ball and consider the likely progress we can expect at optical and near-infrared wavelengths with current and near-term facilities ahead of those possible with the more ambitious new observatories such as the James Webb Space Telescope (JWST) and a new generation of extremely large ground-based telescopes.

9.2 The Next Five Years

Panoramic imaging with large optical and near-infrared cameras on telescopes such as Subaru, UKIRT and VISTA will enable continued exploration of the abundance of luminous drop-outs and Lyα emitters over $5 < z < 7$, reducing the currently troublesome issues of cosmic variance. Deeper data will continue to be provided from further fields taken with ACS and NICMOS. With further spectroscopic surveys on large telescopes, we can expect improved stellar mass density estimates at $z \simeq 5$–6 [80]. However, strong gravitational lensing may still be the only route to probing intrinsically fainter sources, particularly beyond $z \simeq 7$.

More detailed characterization of the properties of the most massive galaxies at $z \simeq 5$–6 will also be worthwhile, in addition to continuing to deduce statistical properties such as abundances and luminosity functions.. Laser-guide star adaptive optics is now in widespread use on our 8–10 m class telescopes and, with integral field spectrographs, can be used to probe the resolved dynamics of distant galaxies ([97]). Application of these techniques to the most intense star-forming lensed systems at $z > 4$ ([242], Fig. 61) is already providing unique insight into the physical state of the earliest galaxies. Extending the same techniques with adaptive optics will advance progress and address process on remarkably small scales ($\simeq 200$ pc).

A number of special purpose ground-based instruments are also being developed, both with and without adaptive optics, to probe for $z > 7$ dropouts and Lyman α emitters. These include:

- The Dark Age Z Lyman Explorer (DAZLE, [113][9], a moderate field (7 arcmin) infrared narrow-band imager with airglow discrimination whose goal is to reach a sensitivity of $\simeq 10^{-18}$ cgs in a night of VLT time. This corresponds to a limiting star formation rate of $\simeq 1 M_{\odot}\,\mathrm{yr}^{-1}$ in quiet regions of the airglow spectrum at $z \simeq 7.7$ and 9.9. Given the modest field and the need for dedicated spectroscopic follow-up, a long term campaign is envisaged to conduct a reliable census.
- The Gemini Genesis Survey (GGS [10]) which uses F2T2, a clone of the tunable filter destined for JWST ([208]). This instrument is planned to work behind the Gemini Observatory's multi-conjugate adaptive optics facility and image, in steps of wavelength, regions lensed by foreground clusters. The success of such a strategy will depend on both the angular size and line widths of $z > 7$ Lyman α emitters and the abundance of intrinsically faint examples. Reference [77] found the lensed emitter at $z \simeq 5.7$ was less than 30 milli-arcsec across with a line width of only 100–200 km sec^{-1}. If such tiny emitters are the norm at $z > 7$, the GGS may go much deeper than extant spectroscopic lensed searches ([231]).

[9] http://www.ast.cam.ac.uk/~ optics/dazle/
[10] http://odysseus.astro.utoronto.ca/ggs-blog/?page_id=2

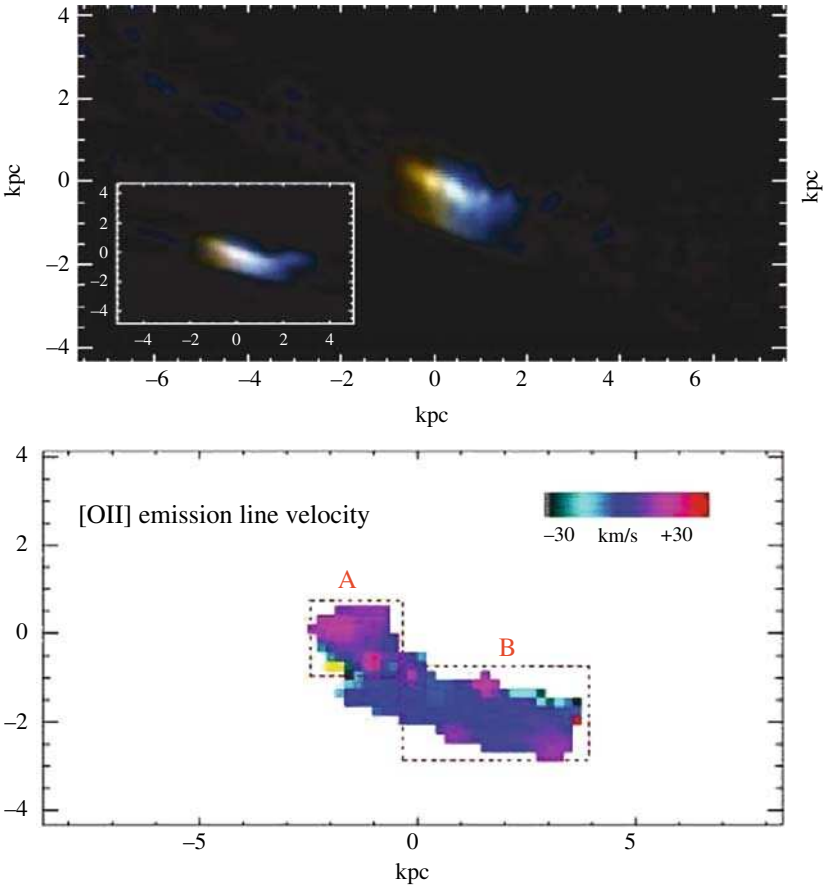

Fig. 61. Detailed studies of a lensed $z \simeq 5$ galaxy (Swinbank et al. 2006 [242]). (*Top*) Reconstructed color HST VI image of the $z = 4.88$ arc in the cluster RCS0224-002. The inset shows the effect of 0.8 arcsec seeing on the reconstruction, thereby demonstrating the advantage of lensing. In the source plane, the galaxy is 2.0×0.8 kpc. (*Bottom*) [O II] velocity field obtained during a 12 hr VLT SINFONI exposure without AO. Spatial comparison with the Lyα field gives clear evidence of significant bi-polar outflows

The exciting redshift range $7 < z < 12$ will also be the province of improved drop-out searches using the instrument WFC3 slated to be installed on Hubble Space Telescope in 2008–9[11]. The infrared channel of this instrument spans 850–1170 nm with a field of $\simeq$ 2 arcmin at an angular resolution of 0.13 arcsec pixel^{-1}. This resolution is coarser than that of adaptive optics-assisted instruments such as F2T2. The principal gain over ground-based instruments will be in deep broad-band imaging free from airglow. The

[11] http://www.stsci.edu/hst/wfc3

survey efficiency is about an order of magnitude better than that of NICMOS. WFC3 also has two infrared grisms which will be helpful in source discrimination.

A major stumbling block at the moment, even at $z \simeq 5$–6, is efficient spectroscopic follow-up of dropout candidates. As we discussed in Sect. 6, photometric redshifts have unfortunately become *de rigeur* in statistical analyses of luminosity densities and luminosity functions [30], yet their precision remains controversial. For z- and J-band dropous beyond $z \simeq 7$, photometric redshifts will be even less reliable. Spitzer detections will be harder and fewer, and the typical source may have only 2–3 detected bands. Candidates may be found in abundance but how will they be confirmed?

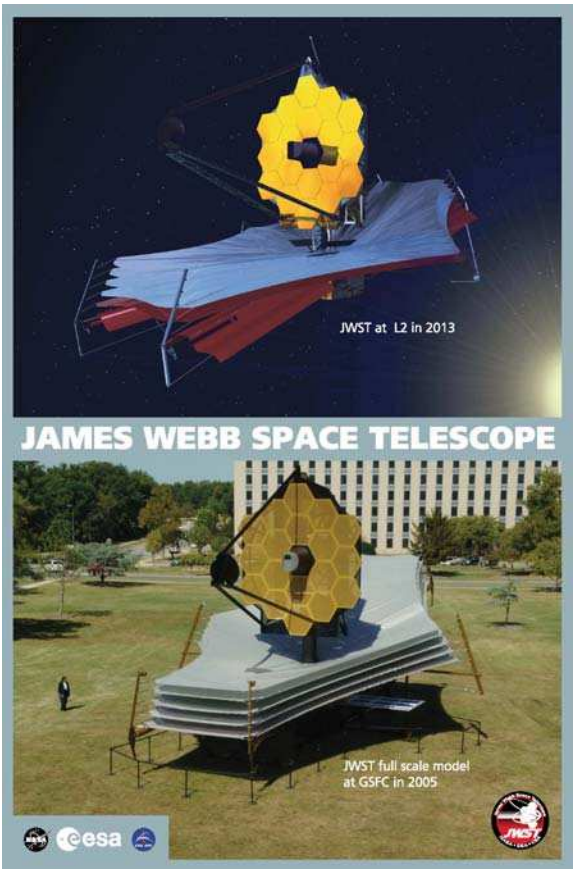

Fig. 62. The 6.5 m James Webb Space Telescope: Then (2013 in orbit) and Now (2005, full scale model)

Fig. 63. The proposed US-Canadian Thirty Meter Telescope (www.tmt.org) now in the detailed design phase

The various 8–10 m telescopes are now building a new generation of cryogenic near-infrared multi-slit spectrographs. It can be hoped that long exposures with these new instruments will be sufficient to break this impasse.

9.3 Beyond Five Years

A number of facilities are being planned, motivated by the progress discussed in these lectures. These include:

- James Webb Space Telescope – a 6.5 m optical-infrared space observatory ([96]) for which a primary mission is the detailed study of the earliest star-forming galaxies (Fig. 62). The facility is currently planned to have a near-infrared imager NIRCam, a spectrograph NIRSpec with both integral field and limited multi-object modes, a tunable filter (as F2T2) and a mid-IR imager, MIRI. Reference [232] has presented a cogent summary of the likely strategies of using this facility, due to be launched in 2013, for studies of the earliest systems.
- Extremely Large Telescopes on the ground including the US-Canadian Thirty Meter Telescope (TMT, see www.tmt.org, Fig. 63) which will have a diffraction-limited near-infrared imager/spectrograph (IRIS) and a adaptive-optics assisted spectrograph (IRMOS) with multiple deployable *integral field units* - area mapping units which can be arranged not only to scan regions surrounding luminous sources but also to undertake multi-object studies in crowded regions (Fig. 64). At the time of writing, TMT is expected to be operational from 2016 onwards.

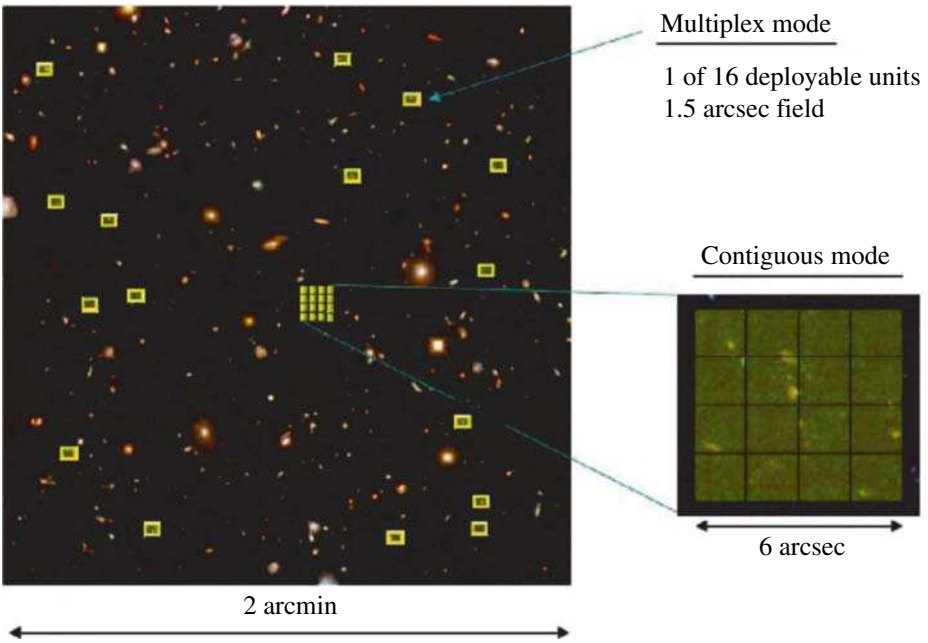

Fig. 64. Flexibility in the deployment strategy for the multiple integral field units (yellow squares) for the proposed TMT infrared multi-object adaptive optics assisted instrument IRMOS. The IFUs can be distributed around the full field of 2 arcmin in classic multi-object mode. Alternatively, the IFUs can be configured together to map a small 6×6 arcsec field at high angular resolution in "blind" mode. Such flexibility will be important in mapping Lyα emission at high redshift in various situations

These will complement redshifted 21 cm line tomography with radio facilities and address the key questions of the escape fraction of photons from star-forming sources and how they create ionized bubbles which merge to cause reionization.

It is quite likely that, by 2013, the redshift range containing the earliest galactic sources, estimated at present to be $10 < z < 20$ perhaps, will have been refined sufficiently by special-purpose instruments on our existing 8–10 m class telescopes. Thus one can surmise that that both JWST and future ELTs will be used for much more challenging work related to the physical process of reionization, as well as the chemical maturity of the most luminous sources found at high redshift.

An obvious partnership between JWST and TMT, for example, which would complement the 21 cm studies, would be to (i) search for the extent and topology of faint Lyα emission in ionized bubbles around JWST-selected luminous star forming galaxies and, (ii) pinpoint early sources for spectroscopic scrutiny so as to identify signatures of Population III stars.

At the present time there are so many imponderables in our knowledge of the earliest sources, that even the design parameters for the ELT instruments is a considerable challenge. How big are the faintest Lyα emitters? What are the typical line widths in km sec^{-1}? How big are the ionized bubbles at a given redshift? And, crucially, what is the surface density of various types of star-forming galaxies. Flexibility in the design of survey strategies will be crucial with instruments such as TMT's IRMOS (Fig. 64).

Any information we can glean on the properties of $z \simeq 10$ sources in the next 5 years will be valuable in optimizing how to move forward when these magnificent new generation telescopes are made available to us. Indeed, it is foolhardy to wait! Time and again, we can retrospectively look back at what we thought we would accomplish with our planned facilities and we always find that we achieved more than we expected!

References

1. Aaronson, M.: Ap. J. **221**, L103 (1978)
2. Aharonian, F., et al.: Nature **440**, 1018 (2006)
3. Adelberger, K., et al.: Ap. J. **505**, 18 (1998)
4. Astier, P., et al.: Astron. Astrophys., **447**, 31 (2006)
5. Bahcall, N., et al.: Science **284**, 1481 (1999)
6. Baldry, I., Glazebrook, K.: Ap. J. **593**, 258 (2003)
7. Barkana, R., Loeb, A.: Phys. Rep. **349**, 125 (2000)
8. Baugh, C., et al.: MNRAS **283**, 1361 (1996)
9. Baugh, C., et al.: Ap. J. **498**, 504 (1998)
10. Baugh, C., et al.: MNRAS **305**, L21 (2005a)
11. Baugh, C., et al.: MNRAS **356**, 1191 (2005b)
12. Becker, G., et al.: ApJ, **662**, 72, (2007)
13. Becker, R.H., et al.: AJ. **122**, 2850 (2001)
14. Beckwith, S., et al.: AJ. **132**, 1179 (2006)
15. Bell, E., et al.: Ap. J. **608**, 752 (2004)
16. Bell, E., et al.: Ap. J. **640**, 241 (2006)
17. Bender, R., et al.: Ap. J. **399**, 462 (1992)
18. Bennett, C., et al.: Ap. J. Suppl. **148**, 97 (2003)
19. Benson, A., et al.: MNRAS **336**, 564 (2002)
20. Benson, A., et al.: Ap. J. **599**, 38 (2003)
21. de Bernadis, P., et al.: Nature **404**, 955 (2000)
22. Bernstein, R., et al.: Ap. J. **571**, 56 (2002)
23. Blain, A., et al.: Phys. Rep. **369**, 111 (2002)
24. Blain, A., et al.: Ap. J. **611**, 725 (2004)
25. Blandford, R., Narayan, R.: ARAA **30**, 311 (1992)
26. Blanton, M.R., et al.: A.J. **121**, 2358 (2001)
27. Blumenthal, G., et al.: Nature **311**, 517 (1984)
28. Bolton, A., et al.: Ap. J. **638**, 703 (2006)
29. Bouwens, R., et al.: Ap. J. **606**, L25 (2004)
30. Bouwens, R., et al.: Ap.J, **653**, 53, (astro-ph/0509641) (2006)
31. Bower, R., et al.: MNRAS **254**, 601 (1992)

32. Bower, R., et al.: MNRAS **370**, 645 (2006)
33. Brinchmann, J., Ellis, R.S.: Ap. J. **536**, L77 (2000)
34. Broadhurst, T., et al.: Nature **355**, 55 (1992)
35. Broadhurst, T., et al.: Ap. J. **621**, 53 (2005)
36. Bruzual, G.: Ph.D. thesis, UC Berkeley (1980)
37. Bundy, K., et al.: Ap. J. **601**, L123 (2004)
38. Bundy, K., et al.: Ap. J. **625**, 621 (2005)
39. Bundy, K., et al.: Ap. J. **651**, 120 (2006)
40. Bunker, A., et al.: MNRAS **355**, 374 (2004)
41. Bunker, A., et al.: New. AR. **50**, 94 (2006)
42. Calzetti, D., et al.: Ap. J. **533**, 682 (2000)
43. Caroll, S., et al.: ARAA **30**, 499 (1992)
44. Chabrier, G.: PASP **115**, 763 (2003)
45. Chapman, S., et al.: Nature **422**, 695 (2003)
46. Chapman, S., et al.: Ap. J. **622**, 772 (2005)
47. Conselice, C., et al.: Ap. J. **628**, 160 (2005)
48. Cowie, L., et al.: AJ. **109**, 1522 (1995)
49. Cowie, L., et al.: AJ. **112**, 839 (1996)
50. Cole, S., et al.: MNRAS **319**, 168 (2000)
51. Cole, S., et al.: MNRAS **326**, 255 (2001)
52. Cole, S., et al.: MNRAS **362**, 505 (2005)
53. Colless, M., et al.: MNRAS **328**, 1039 (2001)
54. Couch, W., et al.: Ap. J. **497**, 188 (1998)
55. Croton, D., et al.: MNRAS **365**, 11 (2006)
56. Daddi, E., et al.: Ap. J. **617**, 746 (2004)
57. Davies, R.L., et al.: Ap. J. **584**, L33 (2001)
58. de Lucia, G., et al.: MNRAS **366**, 499 (2006)
59. Dickinson, M., et al.: Ap. J. **587**, 25 (2003)
60. Dickinson, M., et al.: Ap. J. **600**, L99 (2004)
61. Djorgovski, S., Davis, M.: Ap. J. **313**, 59 (1987)
62. Djorgovski, S., et al.: Ap. J. **560**, L5 (2001)
63. Dole, H., et al.: Astron. Astrophys. **451**, 417 (2006)
64. Dressler, A.: Ap. J. **236**, 351 (1980)
65. Dressler, A., et al.: Ap. J. **313**, 42 (1987)
66. Dressler, A., et al.: Ap. J. **490**, 577 (1997)
67. Driver, S., et al.: Ap. J. **453**, 48 (1995)
68. Drory, N., et al.: Ap. J. **619**, L131 (2005)
69. Dube, R.R., et al.: Ap. J. **232**, 333 (1979)
70. Dunlop, J., et al.: MNRAS, **376**, 1054 (astro-ph/0606192) (2006)
71. Ebbels, T., et al.: MNRAS **295**, 75 (1998)
72. Eddington, A.S., Obs. **42**, 119 (1919)
73. Egami, E., et al.: Ap. J. **618**, L5 (2005)
74. Erb, D., et al.: Ap. J. **644**, 813 (2006)
75. Ellis, R.S.: ARAA **35**, 389 (1997)
76. Ellis, R.S., et al.: MNRAS **280**, 235 (1996)
77. Ellis, R.S., et al.: Ap. J. **560**, L119 (2001)
78. Ellison, S., et al.: AJ. **120**, 1175 (2000)
79. Eyles, L., et al.: MNRAS **364**, 443 (2005)
80. Eyles, L., et al.: MNRAS, **374**, 910 (astro-ph/0607306) (2007)

81. Fall, S., Efstathiou, G.: MNRAS **193**, 189 (1980)
82. Fall, S., et al.: Ap. J. **464**, L43 (1996)
83. Fan, X., et al.: AJ. **125**, 1649 (2003)
84. Fan, X., et al.: ARAA **44**, 415 (2006a)
85. Fan, X., et al.: AJ. **132**, 117 (2006b)
86. Felton, J.E.: AJ. **82**, 861 (1977)
87. Finkbeiner, D., et al.: Ap J **544**, 81 (2000)
88. Fontana, A., et al.: Astron. Astrophys. **424**, 23 (2004)
89. Franx, M., et al.: Ap. J. **587**, L79 (2003)
90. Frayer, D., et al.: AJ. **120**, 1668 (2000)
91. Freedman, W., et al.: Ap. J. **553**, 47 (2001)
92. Fukugita, M., Kawasaki, M.: MNRAS **343**, L25 (2003)
93. Fukugita, M., Peebles, P.J.E.: Ap. J. **616**, 643 (2004)
94. Fukugita, M., et al.: Ap. J. **503**, 518 (1998)
95. Furlanetto, S., et al.: Phys. Reports **433**, 181 (2005)
96. Gardner, J., et al.: Space Sci. Rev., **123**, 485 (astro-ph/0606175) (2006)
97. Genzel, R., et al.: Nature **442**, 786 (2006)
98. Gerhard, O., et al.: AJ. **121**, 1936 (2001)
99. Giavalisco, M., et al.: Ap. J. **600**, 103 (2004)
100. Glazebrook, K., et al.: MNRAS **275**, L19 (1995)
101. Glazebrook, K., et al.: Nature **430**, 181 (2004)
102. Gradshteyn, I.S., Ryzhik, I.M.: Tables of Integrals, Series and Products, Academic Press (2000)
103. Granato, G.L., et al.: Ap. J. **542**, 710 (2000)
104. Gunn, J.E., Peterson, B.A.: Ap. J. **142**, 1633 (1965)
105. Haiman, Z.: Ap. J. **576**, 1 (2002)
106. Hanany, S.: Ap. J. **545**, 5 (2000)
107. Hauser, M., Dwek, E.: ARAA **39**, 249 (2001)
108. Hauser, M., et al.: Ap. J. **508**, 25 (1998)
109. Hoekstra, H., et al.: Ap. J. **647**, 116 (2005)
110. Hopkins, A.M.: Ap. J. **615**, 209 (2004)
111. Hopkins, A.M., Beacom, J.: Ap. J. **651**, 142 (2006)
112. Horgan, J.: The End of Science, Addison-Wesley (1998)
113. Horton, A., et al.: In Ground-based Instrumentation for Astronomy, S.P.I.E. **5492**, 1022 (2004)
114. Hu, E., et al.: AJ. **127**, 563 (2004)
115. Hubble, E. The Realm of the Nebulae, Yale University Press (1936)
116. Hughes, D., et al.: Nature **394**, 241 (1998)
117. Iye, M., et al.: Nature **443**, 186 (2006)
118. Jorgensen, I., et al.: MNRAS **280**, 167 (1996)
119. Jullo, E., et al.: Astron. Astrophys, submitted (2006)
120. Juneau, S., et al.: Ap. J. **619**, L135 (2005)
121. Kaiser, N., et al.: MNRAS **277**, 1 (1987)
122. Kauffmann, G., Charlot, S.: MNRAS **297**, L23 (1998)
123. Kauffmann, G., et al.: MNRAS **264**, 201 (1993)
124. Kauffmann, G., et al.: MNRAS **283**, L117 (1996)
125. Kauffmann, G., et al.: MNRAS **303**, 188 (1999)
126. Kauffmann, G., et al.: MNRAS **341**, 33 (2003)
127. Kashikawa, N., et al.: Ap. J. **648**, 7 (2006)

128. Kashlinsky, A., Phys. Reports **409**, 361 (2005)
129. Kashlinsky, A., et al.: Ap. J. **579**, 53 (2002)
130. Kashlinsky, A., et al.: Nature **438**, 45 (2005)
131. Kashlinsky, A., et al.: ApJL. **654**, 5 (astro-ph/0612445) (2007)
132. Kelsall, T., et al.: Ap. J. **508**, 44 (1998)
133. Kennicutt, R.: Ap. J. **272**, 54 (1983)
134. Kennicutt, R.: ARAA **36**, 189 (1998)
135. Kneib, J.-P., et al.: Ap. J. **471**, 643 (1996)
136. Kneib, J.-P., et al.: Ap. J. **607**, 697 (2004)
137. Kodaira, I., et al.: PASJ **55**, L17 (2003)
138. Kogut, A., et al.: Ap. J. Suppl. **148**, 161 (2003)
139. Kong, X., et al.: Ap. J. **638**, 72 (2006)
140. Koo, D., Kron, R.: ARAA **30**, 613 (1992)
141. Koopmans, L., Treu, T.: Ap. J. **583**, 606 (2003)
142. Kriek, M., et al.: Ap. J. **649**, L71 (2006)
143. Kroupa, P., et al.: MNRAS **262**, 545 (1993)
144. Kroupa, P.: MNRAS **332**, 231 (2001)
145. Lanzetta, K., et al.: Ap. J. **570**, L492 (2002)
146. Leinert, C., et al.: Astron. Astrophys. Suppl. **112**, 99 (1995)
147. Le Fevre, O., et al.: MNRAS **311**, 565 (2000)
148. Lilly, S.J., et al.: Ap. J. **460**, L1 (1996)
149. Lilly, S.J., et al.: Ap. J. **500**, 75 (1998)
150. McCarthy, P.: ARAA **42**, 477 (2004)
151. McCarthy, P., et al.: Ap. J. **614**, L9 (2004)
152. McClure, R., et al.: MNRAS **372**, 357 (2006)
153. Madau, P., Pozzetti, L.: MNRAS **312**, L9 (2000)
154. Madau, P., Silk, J.: MNRAS **359**, L37 (2005)
155. Madau, P., et al.: MNRAS **283**, 1388 (1996)
156. Madau, P., et al.: Ap. J. **498**, 106 (1998)
157. Madau, P., et al.: Ap. J. **514**, 648 (1999)
158. Maddox, S., et al.: MNRAS **242**, 43 (1990)
159. Malhotra, S., Rhoads, J. ApJL. **617**, 5 (2004)
160. Malhotra, S., et al.: Ap. J. **627**, 666 (2005)
161. Malin, D., Carter, D.: Nature **285**, 643 (1980)
162. Mather, J., et al.: Ap. J. **354**, 37 (1990)
163. Mattilia, K., Astron. Astrophys. **47**, 77 (1976)
164. Mattila, K., Ap. J. **591**, 119 (2003)
165. Mattila, K., MNRAS **372**, 1253 (2006)
166. Matsumoto, T., et al.: Ap. J. **626**, 31 (2005)
167. Meier, D.L.: Ap. J. **207**, 343 (1976)
168. Mellier, Y.: ARAA **37**, 127 (2000)
169. Miller, G.E., Scalo, J.: Ap. J. Suppl. **41**, 513 (1979)
170. Miralda-Escude, J., et al.: Ap. J. **501**, 15 (1998)
171. Mobasher, B., et al.: Ap. J. **635**, 832 (2005)
172. Nagamine, K., et al.: Ap. J. **610**, 45 (2004)
173. Nagao, T., et al.: Ap. J. **634**, 142 (2005)
174. Newman J., Davis, M.: Ap. J. **534**, L11 (2000)
175. Norberg, P., et al.: MNRAS **336**, 907 (2002)
176. Oppenheimer, B.D., et al.: In Chemodynamics: First Stars to Local Galaxies, EAS Publication Series, **24**, 157 (astro-ph/0610808) (2007)

177. Ostriker, J., Peebles, P.: Ap. J. **186**, 467 (1973)
178. Ostriker, J., Steinhardt, P.: Nature **377**, 600 (1995)
179. Ouchi, M., et al.: Ap. J. **620**, L1 (2005)
180. Papovich, C., et al.: Ap. J. **640**, 92 (2006)
181. Partridge, B., Peebles, P.J.E. Ap. J. **147**, 868 (1967)
182. Peacock, J.A., et al.: Nature **410**, 169 (2001)
183. Peebles, P.: Large Scale Structure of the Universe, University of Chicago. (1980)
184. Penzias, A.A., Wilson, R.W., Ap. J. **142**, 419 (1965)
185. Perlmutter, S., et al.: Ap. J. **517**, 565 (1999)
186. Pettini, M., et al.: Astrophys. Sp. Sci. **281**, 461 (2002)
187. Postman, M., et al.: Ap. J. **623**, 721 (2005)
188. Pritchet, C.: PASP **106**, 1052 (1994)
189. Reddy, N., et al.: Ap. J. **633**, 748 (2005)
190. Refregier, A.: ARAA **41**, 645 (2003)
191. Richard, J., et al.: Astron. Astrophys. **456**, 861 (2006)
192. Riess, A., et al.: AJ. **116**, 1009 (1998)
193. Rowan-Robinson, M.: The cosmological distance ladder: Distance and time in the universe, W.H. Freeman and Co, (1985)
194. Rubin, V.: PASP **112**, 747 (2000)
195. Rubin, V., et al.:. AJ. **81**, 687 (1976)
196. Rudnick, G., et al.: Ap. J. **599**, 847 (2003)
197. Salpeter, E.: Ap. J. **121**, 161 (1955)
198. Sand, D.J., et al.: Ap. J. **627**, 32 (2005)
199. Sandage, A.: ARAA **43**, 581 (2005)
200. Santos, M.: MNRAS **349**, 1137 (2004)
201. Santos, M., et al.: Ap. J. **606**, 683 (2004)
202. Sargent, M.T., et al.: ApJS. **172**, 434 (astro-ph/0609042) (2007)
203. Seitz, S., et al.: MNRAS **298**, 945 (1998)
204. Scalo, J.: Fund Cosmic Phys **11**, 1 (1986)
205. Schechter, P.: Ap. J. **203**, 297 (1976)
206. Schlegel, D. et al.: Ap. J. **500**, 525 (1998)
207. Schneider, P.: In Gravitational Lensing: Strong, Weak & Micro, Saas-Fee Advanced Course 33 (2006)
208. Scott, A., et al.: In Ground-based and Airborne Instrumentation for Astronomy S.P.I.E. **6269**, 176 (2006)
209. Shapley, A., et al.: Ap. J. **562**, 95 (2001)
210. Shapley, A., et al.: Ap. J. **588**, 65 (2003)
211. Shapley, A., et al.: Ap. J. **626**, 698 (2005)
212. Shimasaku, K., et al.: PASJ **57**, 447 (2005)
213. Shioya Y., et al.: PASJ **57**, 287 (2005)
214. Smail, I., et al.: Ap. J. **440**, 501 (1995)
215. Smail, I., et al.: Ap. J. **490**, L5 (1997)
216. Smith, G., et al.: Ap. J. **620**, 78 (2005)
217. Smoot, G., et al.: Ap. J. **396**, 1 (1992)
218. Somerville, R., Primack, J.: MNRAS **310**, 1087 (1999)
219. Somerville, R., et al.: MNRAS **320**, 504 (2001)
220. Somerville, R., et al.: Ap. J. **600**, L171 (2004)
221. Songaila, A., AJ. **127**, 2598 (2004)
222. Songaila, A., AJ. **130**, 1996 (2005)

223. Songaila, A., AJ. **131**, 24 (2006)
224. Spergel, D., et al.: Ap. J. Suppl **148**, 175 (2003)
225. Spergel, D., et al.: ApJS. **170**, 377 (astro-ph/0603449) (2007)
226. Springel, V., et al.: Nature **435**, 629 (2005)
227. Stanway, E., et al.: Ap. J. **607**, 704 (2004)
228. Stanway, E., et al.: MNRAS **359**, 1184 (2005)
229. Stark, D., Ellis, R.S.: New AR **50**, 46 (2005)
230. Stark, D., et al.: Ap. J. **659**, 84 (astro-ph/0604250) (2007a)
231. Stark, D., et al.: Ap. J. **663**, 10 (2007b)
232. Stiavelli, M.: In Future Research Directions & Visions for Astronomy, S.P.I.E. **4835**, 122 (2002)
233. Stiavelli, M., et al.: Ap. J. **600**, 508 (2004)
234. Steidel, C., et al.: Ap. J. **492**, 428 (1996)
235. Steidel, C., et al.: Ap. J. **519**, 1 (1999a)
236. Steidel, C., et al.: Phil. Trans R. Soc. **357**, 153 (1999b)
237. Steidel, C., et al.: Ap. J. **592**, 728 (2003)
238. Struck-Marcell, C., Tinsley, B.: Ap. J. **221**, 562 (1978)
239. Sullivan, I., et al.: Ap. J. **657**, 37 (astro-ph/0609451) (2007)
240. Sullivan, M., et al.: MNRAS **312**, 442 (2000)
241. Sullivan, M., et al.: Ap. J. **558**, 72 (2001)
242. Swinbank, M., et al.: MNRAS, **376**, 479 (2007)
243. Taniguchi, Y., et al.: PASJ **57**, 165 (2005)
244. Tinsley, B., Fund. Cosmic Phys. **5**, 287 (1980)
245. Toomre, A., Toomre, J.: Ap. J. **178**, 623 (1972)
246. Tran, K.-V., et al.: Ap. J. **627**, L25 (2005)
247. Treu, T., et al.: Ap. J. **591**, 53 (2003)
248. Treu, T., et al.: Ap. J. **622**, L5 (2005)
249. Treu, T., et al.: Ap. J. **640**, 662 (2006)
250. Tully, R.B., Fisher J.R., Astron. Astrophys. **54**, 661 (1977)
251. van der Wel, A., et al.: Ap. J. **631**, 145 (2005)
252. van der Wel, A., et al.: Ap. J. **636**, L21 (2006)
253. van Dokkum, P.: AJ. **130**, 2647 (2005)
254. van Dokkum, P., Ellis, R.S.: Ap. J. **592**, L53 (2003)
255. van Dokkum, P., et al.: Ap. J. **587**, L83 (2003)
256. van Dokkum, P., et al.: Ap. J. **638**, L59 (2006)
257. Vogt, N., et al.: Ap. J. **465**, L15 (1996)
258. Vogt, N., et al.: Ap. J. **479**, L121 (1997)
259. Weiner, B., et al.: Ap. J. **620**, 595 (2005)
260. Wright, E., Reese, E.D.: Ap. J. **545**, 43 (2000)
261. Yan, H., Windhorst, R.: Ap. J. **612**, 93 (2004)
262. Yan, H., et al.: Ap. J. **634**, 109 (2005)
263. Yan, H., et al.: Ap. J. **651**, 24 (2006)
264. Yee, H., et al.: AJ. **111**, 1783 (1996)
265. York, D., et al.: AJ. **120**, 1579 (2001)
266. Zwaan, M.A., et al.: AJ. **125**, 2842 (2003)
267. Zwicky, F., Helv. Physica Acta, **6**, 110 (1933)
268. Zwicky, F.: Phys. Rev. **51**, 290 (1937)

Acknowledgments

First Light by A. Loeb

I thank my young collaborators with whom my own research in this field was accomplished: Dan Babich, Rennan Barkana, Volker Bromm, Benedetta Ciardi, Daniel Eisenstein, Steve Furlanetto, Zoltan Haiman, Rosalba Perna, Stuart Wyithe, and Matias Zaldarriaga. I thank Donna Adams for her highly professional assistance with the latex file, and Dan Babich & Matt McQuinn for their helpful comments on the manuscript.

Cosmological Feedbacks from the First Stars by A. Ferrara

These Lectures are based on the work of many people whose help, inspiration, collaboration, passion and hard work I can hardly properly acknowledge. As a partial recognition let me at least list their names: S. Bianchi, M. Bruscoli, T. Choudhury, B. Ciardi, I. Iliev, S. Gallerani, H. Hirashita, M. Magliocchetti, M. Mapelli, A. Maselli, P. Richter, S. Salvadori, R. Salvaterra, E. Scannapieco, R. Schneider, F. Sigward, M. Valdes. I am indebted to each of them for one reason or another.

Observations of the High Redshift Universe by R. S. Ellis

I thank Daniel Schaerer, Denis Puy and Angela Hempel for inviting me to give these lectures in such a magnificent location with an enthusiastic group of students. I also thank my fellow lecturers, Avi Loeb and Andrea Ferrara and all of the foregoing for their patience in waiting for the completion of

my written lectures. I thank my close colleagues Kevin Bundy, Sean Moran, Mike Santos Dan Stark, and Tommaso Treu for their help and permission to show results in progress as well as Ivan Baldry, Jarle Brinchmann, Andrew Hopkins, John Huchra and Jean-Paul Kneib for valuable input. Finally, I thank Ray Carlberg and his colleagues for their hospitality of the Astronomy Department at the University of Toronto where the bulk of these lectures notes were completed.

Index