

Frontiers of Cosmology

Edited by

Alain Blanchard and
Monique Signore

NATO Science Series

Frontiers of Cosmology

NATO Science Series

A Series presenting the results of scientific meetings supported under the NATO Science Programme.

The Series is published by IOS Press, Amsterdam, and Springer in conjunction with the NATO Scientific Affairs Division

Sub-Series

I. Life and Behavioural Sciences	IOS Press
II. Mathematics, Physics and Chemistry	Springer
III. Computer and Systems Science	IOS Press
IV. Earth and Environmental Sciences	Springer
V. Science and Technology Policy	IOS Press

The NATO Science Series continues the series of books published formerly as the NATO ASI Series.

The NATO Science Programme offers support for collaboration in civil science between scientists of countries of the Euro-Atlantic Partnership Council. The types of scientific meeting generally supported are "Advanced Study Institutes" and "Advanced Research Workshops", although other types of meeting are supported from time to time. The NATO Science Series collects together the results of these meetings. The meetings are co-organized bij scientists from NATO countries and scientists from NATO's Partner countries – countries of the CIS and Central and Eastern Europe.

Advanced Study Institutes are high-level tutorial courses offering in-depth study of latest advances in a field.

Advanced Research Workshops are expert meetings aimed at critical assessment of a field, and identification of directions for future action.

As a consequence of the restructuring of the NATO Science Programme in 1999, the NATO Science Series has been re-organised and there are currently Five Sub-series as noted above. Please consult the following web sites for information on previous volumes published in the Series, as well as details of earlier Sub-series.

<http://www.nato.int/science>
<http://www.springeronline.com>
<http://www.iospress.nl>
<http://www.wtv-books.de/nato-pco.htm>



Series II: Mathematics, Physics and Chemistry – Vol. 187

Frontiers of Cosmology

edited by

Alain Blanchard

Observatoire Midi-Pyrénées,
Toulouse, France

and

Monique Signore

Observatoire de Paris,
Paris, France



Springer

Proceedings of the NATO Advanced Study Institute on
The Frontiers of Cosmology
Cargèse, France
8–20 September 2003

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-3056-8 (PB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-3056-7 (PB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-10 1-4020-3055-X (HB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-10 1-4020-3057-6 (e-book) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-3055-0 (HB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-3057-4 (e-book) Springer Dordrecht, Berlin, Heidelberg, New York

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Printed on acid-free paper

All Rights Reserved
© 2005 Springer
No part of this work may be reproduced, stored in a retrieval system, or transmitted
in any form or by any means, electronic, mechanical, photocopying, microfilming,
recording or otherwise, without written permission from the Publisher, with the exception
of any material supplied specifically for the purpose of being entered
and executed on a computer system, for exclusive use by the purchaser of the work.

Printed in the Netherlands.

Contents

Preface	ix
Acknowledgments	x
1	
Basics of Cosmology	1
<i>Monique Signore and Alain Blanchard</i>	
1. Geometry and Dynamics	2
2. Important quantities needed for observations	5
3. Some solutions of EFL equations:some cosmological models	8
4. The standard Big Bang Nucleosynthesis (SBBN)	12
5. Observations of “primordial abundances”.	16
6. Confrontation of the observed “primordial abundances” to the predictions of the sBBN.	18
7. Conclusions	20
References	21
2	
The X-ray View of Galaxy Clusters	23
<i>William Forman</i>	
1. Observing Clusters in X-rays – the Chandra Observatory	27
2. Regular Clusters XD Cooling Flows	33
3. Physics of Cluster Cores	36
Acknowledgments	41
References	41
3	
Clusters: an optical point of view	43
<i>Christophe Adami</i>	
1. Cluster detections in the optical	44
2. Studies of clusters	47
3. Acknowledgements	54
References	54
4	
Cosmology with Clusters of Galaxies	57
<i>Alain Blanchard</i>	
1. Introduction	57

2.	What is a cluster?	58
3.	The spherical model	59
4.	The mass function	59
5.	Connection to the observations	60
6.	Properties of Clusters and scaling relations	63
7.	Clusters abundance evolution.	65
8.	The baryon fraction	68
9.	Conclusion	72
	References	72
5		
	Astrophysical detection of Dark Matter	75
	<i>S. Colafrancesco</i>	
1.	Signals from the Dark universe	75
2.	Inference probes	76
3.	Physical probes	77
4.	Conclusion	83
	References	83
6		
	Non-thermal and relativistic processes in galaxy clusters	85
	<i>S. Colafrancesco</i>	
1.	Non-thermal and relativistic phenomena in galaxy clusters	86
2.	The origin of cosmic rays in galaxy clusters	90
3.	The astrophysics of cosmic rays in galaxy clusters	94
4.	Conclusions	98
	References	98
7		
	An introductory overview about Cosmological Inflation	101
	<i>Alain Riazuelo</i>	
1.	Introduction	101
2.	The hot Big-Bang scenario and its problems	103
3.	Inflation and inflationary dynamics	109
4.	Basics of cosmological perturbations	117
5.	Inflationary perturbations	128
6.	Basics of quantum field theory	132
7.	Perturbation spectrum	134
8.	Conclusion	137
	References	137
8		
	An introduction to quintessence	139
	<i>Alain Riazuelo</i>	
1.	The two cosmological constant problems	139
2.	A scalar field as dark energy	141
3.	Stability of the $w_Q = \text{Const}$ regime	142
4.	Model building	143
5.	Dark energy and structure formation	145
6.	Observational status	145

<i>Contents</i>	vii
References	147
9	
CMB Observational Techniques and Recent Results	149
<i>Edward L. Wright</i>	
1. Introduction	150
2. Observational Techniques	154
3. Recent Observations	167
4. Summary	170
Acknowledgments	171
References	171
10	
Fluctuations in the CMB	175
<i>Andrew H. Jaffe</i>	
1. Introduction	175
2. Cosmological Preliminaries	176
3. The Last Scattering Surface	178
4. Perturbations on Large and Small Scales	180
5. Oscillations in the Primordial Plasma	183
6. The Power Spectrum of CMB Fluctuations	187
7. The CMB and Cosmological Parameters	188
8. Conclusions	191
Acknowledgments	192
References	193
11	
Supernovae as astrophysical objects	195
<i>Bruno Leibundgut</i>	
1. Some History	195
2. Supernova classification	196
3. Input Energy	199
4. Core-collapse supernovae	200
5. Type Ia supernovae	202
6. Conclusions	202
References	203
12	
Cosmology with Supernovae	207
<i>Bruno Leibundgut</i>	
1. Introduction	207
2. The Hubble constant	208
3. The expansion history of the universe	210
4. Universal acceleration according to Type Ia supernovae	211
5. Characterising dark energy	214
6. Conclusions	215
References	216

Gravitational lensing: from μ -lensing to cosmic shear experiments	219
--	-----

Francis Bernardeau

1. Introduction	219
2. Physical mechanisms	220
3. Gravitational lenses in Cosmology	224
4. Cosmic Shear: weak lensing as a probe of the large-scale structure	230
5. Conclusions and perspectives: cosmic shear in a precision cosmology era	239
References	239

Dark Matter: Early Considerations	241
-----------------------------------	-----

Jaan Einasto

1. Introduction	241
2. Local Dark Matter	242
3. Clusters and Groups of Galaxies	243
4. Masses of Galaxies	245
5. The Nature of Dark Matter	252
6. Summary	256

Acknowledgments	258
-----------------	-----

References	258
------------	-----

Dark Matter and Galaxy Formation	263
----------------------------------	-----

Joseph Silk

1. Challenges of dark matter	263
2. Global baryon inventory	264
3. Confirmation via detailed census of MWG/M31	265
4. Hierarchical galaxy formation	266
5. Unresolved issues in galaxy formation theory	268
6. Resurrecting CDM	269
7. An astrophysical solution: early winds	271
8. Observing CDM via the WIMP LSP	273
9. The future	274
References	277

Non-Baryonic Dark Matter	279
--------------------------	-----

Paolo Gondolo

1. The need for non-baryonic dark matter	279
2. Popular candidates for non-baryonic dark matter	281
3. Neutralino dark matter searches	299
4. Conclusions	326
References	327

Preface

The field of cosmology is currently undergoing a revolution driven by dramatic observational progress and by novel theoretical scenarios imported from particle physics. In particular, two most remarkable results were recently obtained from measurements of the angular spectrum of the fluctuations in the Cosmic Microwave Background (CMB) radiation providing convincing evidence that the Universe is nearly flat and from the Hubble diagram of distant supernovae indicating an accelerating expansion rate, which implies the existence of some dark energy as the dominant component of the Universe. Indeed, the next decade will benefit from high quality data on cosmology from different major experiments and observatories, with a particular important contribution from space missions such as WMAP, Planck Surveyor, XMM and SNAP among others. On one side, cosmologists believe they understand the origin of the main ingredients which allow a coherent description of the Universe from its very early phase, namely inflation, to the actual epoch which accounts for the origin of the primordial fluctuations, allowing predictions of their imprints in the cosmic microwave sky and leading to the large scale structure of the Universe as observed. On the other side, the existence of a non-zero vacuum density is certainly one of the most astonishing results of modern fundamental physics. Understanding its nature and its origin will be one of the major directions of research in the following years. In view of the intensive current activity in the field, a School fully dedicated to these both sides in cosmology was timely. This 11-days NATO Advanced Study Institute took place in the lovely setting of the *Institut d'Études Scientifiques de Cargèse (Corse, France)* and was attended by about 80 participants from several countries. These proceedings contain the papers that were presented during the School and which covered the following fields : quintessence/dark energy; inflation; CMB: anisotropies and polarization; large scale structure; clusters of galaxies; gravitational lensing; galaxy formation; dark matter; supernovae and the accelerating expansion of the Universe.

ALAIN BLANCHARD & MONIQUE SIGNORE

Acknowledgments

The success of the highly productive, friendly and relaxed School **Frontiers of the Universe** is largely due to the local and flawless organization of the *Institut d'Études Scientifiques de Cargèse* and to the support provided by Dr Elisabeth Dubois-Violette, director of the Institut. This school will not have been possible without the financial support from NATO. Finally, we also gratefully acknowledge the financial support of CNRS, INSU, IN2P3, PNC, CNES and Collectivité Territoriale Corse.

Chapter 1

BASICS OF COSMOLOGY

Monique Signore

LERMA, Observatoire de Paris, 61 Av.de l'Observatoire, 75014 Paris, France

Monique.Signore@obspm.fr

Alain Blanchard

LATT, Observatoire Midi-Pyrénées, CNRS 14, Av. Ed. Belin, 31 400 Toulouse, France

Alain.Blanchard@ast.obs-mip.fr

Introduction

We will begin by briefly reviewing the *General Cosmological Framework* in which the following lectures will fit the *Hot Big Bang*!

In practice, on the basis of three observational facts :

- i) the Universe is currently in a state of uniform expansion ,
- ii) the Universe is filled with photons that come from background blackbody radiation at a temperature of about 2.74 K ,
- iii) the Universe is isotropic on large scales i.e. beyond nearly 1000 Mpc, one can construct, from General Relativity, a generic cosmological model known as the *Hot Big Bang* or the *Standard Big Bang Model*.

This model, which ancestor is the Lemaître *Primeval atom*, is but a mere 40 years old and provides a description of some other observations like, in particular, the abundances of light elements.

The paper is organized as follows : in section 2, we briefly review the geometry and dynamics of the Universe and then give the Einstein-Friedman-Lemaître (hereafter EFL) equations; section 3 introduces some important quantities needed for observations; in section 4, we rapidly present some solutions of the EFL equations, i.e. some cosmological models; in section 5, the *Standard Big Bang Nucleosynthesis Model* is described while section 6 shows a statement of observations of primordial abundances; in section 7, we confront the predictions of the Standard Big Bang Nucleosynthesis (hereafter SBBN) model to the observations of the primordial abundances; a brief conclusion is

given in section 8.

Let us also mention the recent review "Precision Cosmology" - and references therein - due to Melchiorri et al. (2003) which develops many topics that are rapidly presented here.

1. Geometry and Dynamics

Geometry of 4-dimensional space-time

The fundamental idea of geometrical theory of gravity starts from the fact that we can assign four coordinates to any event observed in our vicinity, for instance in Cartesian coordinates (x, y, z, t) . Locally, space appears flat. However this does not prejudge of the global shape of space : local observations put us in the same situation that lead people to think the earth was flat. Let us take the line element of a homogeneous 3D space which can be shown to be :

$$dl^2 = r^2(d\theta^2 + \sin^2 \theta d\phi^2) + \frac{dr^2}{1 - k \left(\frac{r}{R}\right)^2}$$

where k is $-1, 0, 1$ accordingly to whether space is hyperbolic, flat or spherical. R is a characteristic size (in the spherical case, that is the radius of the 3D-sphere embedded in a 4D space).

We can add the time as the fourth coordinate, to build the equivalent of the Minkowski space-time element. We then get the Robertson-Walker line element after the change of variables $\frac{r}{R} \rightarrow r$:

$$ds^2 = -c^2 dt^2 + R(t)^2 [r^2(d\theta^2 + \sin^2 \theta d\phi^2) + \frac{dr^2}{1 - kr^2}] \quad (1.1)$$

Topology

The above line element depends on the local shape of space: the curvature (i.e. the value of k) is only a local property of space, its geometry, but does not tell us the *global* shape of space. For instance, the Euclidean plane is an infinite flat surface while the surface of a cylinder is a 2D-space which is flat everywhere but is finite in one direction. Identically, we may in principle derived the local geometry of space through General Relativity. It does not prejudge of the global topology of space. Only direct observations would allow to test what the topology actually is. Of course this will not be possible on scales larger than what can be observed (the horizon). We can therefore hope to prove that the Universe is finite, if it is small enough, but we could never know whether we are in a finite Universe of which the scale is larger than the horizon, or whether we are in an infinite Universe.

Dynamics

The function $R(t)$ which appears in the RW line element, is totally independent of any geometrical consideration. It can be specified only within a theory of gravity. Although General Relativity (GR hereafter) is at the starting point of modern cosmology, it is often of little use in practice as in most cases we are in the weak field regime, for which Newton theory is sufficient. Therefore, this lecture will say almost nothing about GR. The basic equation of GR relates the geometrical tensor G_{ij} to the energy-momentum tensor T_{ij}

$$G_{ij} = R_{ij} - \frac{1}{2}g_{ij}R = 8\pi GT_{ij} \quad (1.2)$$

There exists a coordinates system, called the comoving coordinates, in which the matter is at rest, and the tensor T_{ij} is diagonal with $T_{00} = \rho$ and $T_{11} = T_{22} = T_{33} = p$, ρ being the density and p the pressure. A fundamental aspect of GR is that the source of gravity includes explicitly a term coming from the pressure : $\rho + 3p/c^2$. Finally, there is an analog of the Gauss theorem, that is the Birkhoff's theorem: if the matter distribution is spherical then the evolution of the radius of a given shell of matter does depend only on its internal content.

From the above rules, we can easily derive the equation for $R(t)$. Let us consider a spherical region of radius a in a homogeneous distribution of matter. The equivalent Newtonian acceleration is:

$$\frac{d^2a}{dt^2} = g$$

with the acceleration being generated by the “mass” of the above spherical region $M(a)$:

$$g = -\frac{GM(a)}{a^2} = -\frac{4}{3}\pi G(\rho + 3p/c^2)a$$

The density term includes the effect of kinetic energy ($E = mc^2!$), so that energy conservation can be written inside the volume of the sphere, and elementary thermodynamics gives:

$$d(E_t) = d(\rho V c^2) = -pdV$$

leading to :

$$\dot{\rho} = -3\left(\frac{p}{c^2} + \rho\right)\frac{\dot{a}}{a}$$

From these two equations, the pressure can be eliminated, and, after having multiply both term by \dot{a} , the differential equation can be easily integrated. This leads to the following equation:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{Kc^2}{a^2}$$

The last term corresponds to the constant of integration associated to the total energy of the sphere (and varies as a^2). Its value depends on the initial conditions. Furthermore, it expresses a link between the geometry and the material content of the Universe, which cannot be specified in the Newtonian approach we had and can be justified only within the framework of GR. The form of the above equation is independent of the radius a of the sphere and we shall therefore admit that the equation still holds for the quantity $R(t)$, the constant K being then the constant k which is involved in the Robertson-Walker metric element :

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G\rho}{3} - \frac{kc^2}{R^2(t)}$$

In order to specify completely the function $R(t)$, one needs an equation of state for the content of the Universe. The two cases often seen in cosmology are the dust case ($p = 0$) and the radiation dominated regime ($p = \frac{1}{3}\rho c^2$).

Vacuum and the Cosmological Constant

The vacuum is a particular medium, and one should wonder what is the state equation of this medium. Naively, one would think that the equation will be $p = 0$ and $\rho = 0$. However, let's try to derive the vacuum equation of state from first principles. As in classical thermodynamics let's assume that we have a piston with vacuum in it. We also assume that simple vacuum ($p = 0, \rho = 0$) is present outside). The energy inside the piston is $E = \rho_v c^2 V$. If the volume changes by a small amount the net energy change is:

$$dE = d(\rho_v V c^2) = \rho_v c^2 dV$$

this change is equal to the work of the pressure:

$$dE = -p_v dV$$

so the equation of state is:

$$p_v = -\rho_v c^2$$

As one can see, the conditions $p \geq 0$ and $\rho \geq 0$ guaranty that the simple solution is the only one. However, there is nothing which imposes these conditions for the vacuum, and we can therefore decide to keep such a possible term. This can be directly introduced in the equation governing $R(t)$ or after having introduced the following constant:

$$\Lambda = 8\pi G\rho_v$$

Such a term is called the cosmological constant, and has been historically introduced by Einstein, as a modification of his original theory. We have then

recovered the Einstein-Friedmann-Lemaître (EFL) equations:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G\rho}{3} - \frac{kc^2}{R^2} + \frac{\Lambda}{3} \quad (1.3)$$

and

$$\dot{\rho} = -3\left(\frac{p}{c^2} + \rho\right)\frac{\dot{R}}{R} \quad (1.4)$$

Remarks :

1) later, we will see that observations favor an accelerating Universe indicating the existence of "dark energy" which can be identified with the vacuum energy or with something called the quintessence (Wetterich, 1988; Ratra & Peebles, 1988; Caldwell et al., 1997; Peebles & Ratra, 2002).

2) very often, some "reduced" quantities can be used in the literature :

$H = \frac{\dot{R}}{R}$, the Hubble parameter,

$\Omega_M = \Omega = \frac{8\pi G\rho}{3H^2}$ the density parameter,

$q = -\frac{\ddot{R}R}{\dot{R}^2}$, the deceleration parameter,

$\Omega_\lambda = \lambda = \frac{\Lambda}{3H^2}$, the (reduced) cosmological constant,

$\alpha = -\Omega_k = \frac{kc^2}{H^2R^2}$, the curvature parameter. Quantities are labeled by 0 when they are referred to their present value. For instance the present day value of the density parameter Ω is Ω_0 . The EFL equations then reads:

$$\Omega_M + \Omega_k + \Omega_\lambda = 1$$

or:

$$\alpha = \Omega + \lambda - 1$$

so that the "radius of the Universe" can be written :

$$R = \frac{c}{H} \frac{1}{\sqrt{|\alpha|}}$$

2. Important quantities needed for observations

In this section we only need to work in the framework of a geometrical theory of space-time, in which the trajectories of light rays are assumed to be the null geodesics. Let's have a comoving spherical coordinates system (r, θ, ϕ, t) with the observer at the origin of the spatial coordinates $(r = 0, \theta = 0, \phi = 0, t_0)$. Let's assume the observed source is emitting light at the coordinates $(r_S = 0, \theta = 0, \phi = 0, t_S)$, and let's $r(t)$ be the trajectory of the photons emitted. As this trajectory is a null geodesic, we have:

$$c^2 dt^2 - R^2(t) \frac{dr^2}{1 - kr^2} = 0$$

so the variables can be separated and the integration over r is analytical:

$$\int_{t_S}^{t_0} \frac{cdt}{R(t)} = \int_0^{r_S} \frac{dr}{(1 - kr^2)^{1/2}} = S_k^{-1}(r_S)$$

with:

$$S_k(r_S) = \begin{cases} \sin(r_S) & \text{if } k = +1 \\ r_S & \text{if } k = 0 \\ \sinh(r_S) & \text{if } k = -1 \end{cases}$$

When the distance is small in front of R_0 we just have $S_k^{-1}(r) = r$.

The Redshift

In order to derive the observed frequency ν_0 of the light coming from a source emitting at the frequency ν , we consider the trajectory of the light ray emitted at the time $t_S + \frac{1}{\nu}$. As the source is comoving, we have:

$$S_k^{-1}(r_S) = \int_{t_S}^{t_0} \frac{cdt}{R(t)} = \int_{t_S+1/\nu}^{t_0+1/\nu_0} \frac{cdt}{R(t)}$$

$$\frac{\nu_0}{\nu} = \frac{\lambda_S}{\lambda_0} = \frac{R_S}{R_0} = \frac{1}{1+z}$$

where z is the redshift. This is the standard formula for the cosmological shift of the frequencies. This result shows that the redshift z is a natural consequence of the expansion.

The proper distance

In GR, space changes with time, so that the “intuitive” distance between two points changes with time. Therefore the various methods to measure the distance from the source to a point give different answers. See, for instance, the “Table 1: distances in cosmology” (Melchiorri et al., 2003). The proper distance – between the source and the observer – can be seen as a distance measured by a set of rulers. The distance element is given by :

$$dl^2 = ds^2 = R(t)^2 \frac{dr^2}{1 - kr^2}$$

so that the proper distance is :

$$D = R(t)S_k^{-1}(r_S)$$

The fact that this distance changes with time is the observational consequence of expansion. We can now examine how this length changes with time :

$$\dot{D} = \dot{R}S_k^{-1}(r)dt$$

so that the source is *actually recessing* from the observer with a speed:

$$V = \frac{\dot{R}}{R}D = HD$$

the fact that this speed could be larger than the speed of light should not be considered as a problem: this speed can be measured but cannot help to transport information faster than light. When the speed is small, the Doppler frequency shift is :

$$\frac{\delta\nu}{\nu} = \frac{\dot{R}}{R}\delta t = H\frac{D}{c} = \frac{V}{c}$$

so that the shift is exactly the one corresponding to the Doppler shift associated with the above velocity. The redshift is therefore purely a kinematical effect.

The angular distance

Let's suppose we observe a ruler orthogonal to the line of sight. The extremities of the rulers have the coordinates $(r, 0, 0, t_S)$ and $(r, \theta, 0, t_S)$. The proper length l between the extremities is:

$$l^2 = -ds^2 = R(t_S)^2 r^2 \theta^2$$

We then obtain the relation between the angle θ and the length l . The angular distance is defined by:

$$D_{\text{ang}} = R(t_S)r$$

The luminosity distance

Let's assume that we observe a source with an absolute luminosity L through a telescope with a diameter d . Now we can chose a coordinates system which is centered on the source. Let θ be the angle between rays reaching the opposite side of the telescope. We have $d = R(t_0)r\theta$. The energy emitted by the source that reaches the telescope is :

$$s = \frac{L}{4\pi} \times \frac{\pi\theta^2}{4}$$

When observed, the energy of photons has been shifted by $1/(1+z)$ but also the frequency at which they arrive is reduced by the same factor. Therefore the flux (energy per unit time and unit surface) one gets is:

$$f = \frac{s}{\pi l^2/4} \frac{1}{(1+z)^2} = \frac{L}{4\pi R(t_0)^2 r^2 (1+z)^2} = \frac{L}{4\pi D_{\text{lum}}^2}$$

This relation provides the luminosity distance:

$$D_{\text{lum}} = R(t_0)r(1+z) = R(t_S)r(1+z)^2 = D_{\text{ang}}(1+z)^2$$

Distance along the line of sight

We consider here the length along the path of a photon. The length element is

$$dl = cdt = c \frac{dR}{\dot{R}} = -\frac{c}{H(z)} \frac{dz}{1+z}$$

This relation is useful to write the volume element.

The age of the universe

The general expression of time interval is :

$$dt = \frac{dR}{\dot{R}} = -\frac{1}{H(z)} \frac{dz}{1+z}$$

For instance, in the pressure-less case, the look back time, i.e. the time since the epoch corresponding to the redshift z is:

$$\tau(z) = \frac{1}{H_0} \int_1^{1+z} \frac{d\xi}{\xi \sqrt{\Omega_0 \xi^3 - \alpha_0 \xi^2 + \lambda_0}}$$

where $\xi = 1 + z$. When $\lambda \leq 0$ it can be shown from the expression of \ddot{R} that

$$t_0 = \tau(\infty) \leq \frac{1}{H_0}$$

In the case $\Omega = 1$, the age of the universe is just $\frac{2}{3} H_0^{-1}$.

Let us only remark that excellent progress has been recently made toward the measurement of H_0 through the Hubble Space Telescope. Consistently with current estimates, the result favors the following value: $72 \pm 2 \pm 7$ km/s/Mpc (Freedman & Turner, 2003).

In general, there is no direct information on t_0 . But there are several astrophysical objects for which we believe an age can be derived. The most common is the age of globular clusters. Since Hipparcos results, the age of globular clusters is now generally believed to be in the range 10 – 13 Gyr, with an uncertainty of 1.5 Gyr (Chaboyer, 1998). Let us also mention that the star CS 31082-001 shows an age of about 12.5 ± 3 Gyr, based on the decay of U-238 (Cayrel et al., 2001). These values are consistent with age constraints obtained from WMAP (Spergel et al, 2003).

3. Some solutions of EFL equations:some cosmological models

The various possible theories provide different functions $R(t)$ and through the above tests may in principle be distinguished by observations. However,

it is easy to check that the difference occurs only at high redshift. In practice these tests are not discriminant because the observations are difficult and because high redshift objects are younger.

One must also underline that in order to solve the EFL equations, one has to specify the relation between p and ρ i.e. an equation of state.

There are two important regimes : the matter dominated one and the radiation dominated one.

In the matter dominated one, the pressure is negligible. Energy conservation then leads to :

$$\dot{\rho} = -3\rho \frac{\dot{a}}{a} \text{ giving } \rho a^3 = \text{cste}$$

While in the pressure dominated case

$$\dot{\rho} = -4\rho \frac{\dot{a}}{a} \text{ giving } \rho a^4 = \text{cste}$$

This constant can be set into the EFL equation. The solutions are not analytical in the general case, but when $\lambda_0 = 0$, or when $\Omega_0 = 0$, or when $\alpha_0 = 0$.

Case ($\lambda_0 = 0, p = 0$)

When the cosmological constant is zero, there are three types of solutions:

a) when the density is above the critical density:

$$\rho > \rho_c = \frac{3H_0^2}{8\pi G} = 2.h^2 10^{-29} \text{ g/cm}^3$$

the spatial solution is the spherical space. The function $R(t)$ grows from zero to a maximum value then a collapse phase follows to zero.

b) when the density is equal to the critical density, $R(t)$ is simple:

$$R(t) = R_0 \left(\frac{3}{2} H_0 t \right)^{2/3} = R_0 \left(\frac{t}{t_0} \right)^{2/3}$$

with $t_0 = \frac{2}{3} H_0^{-1} = 1/(6\pi G \rho_c)^{1/2}$

c) when the density is below the critical density, the function $R(t)$ grows from zero to infinity. It is easy to check from EFL equations that the function $R(t)$ behaves like t when R is large.

Identically, the behavior of $R(t)$ can be found when $t \rightarrow 0$ independently of the model: $R(t) \propto t^{2/3}$.

Finally, the relation between the comoving coordinates r and the redshift can be expressed:

$$R_0 r = \frac{c}{H_0} \frac{2}{\Omega_0^2} \frac{\Omega_0(1+z) + 2 - 2\Omega_0 - (2-\Omega_0)\sqrt{1+\Omega_0 z}}{1+z}$$

This is known as the Mattig relation. Others useful quantities can be found in Weinberg (1972). See also Melchiorri et al. (2003) and references therein.

Case ($\lambda_0 > 0, p = 0$)

There are much more possibilities when a cosmological constant is allowed. To specify a cosmological model, it is customary to specify two "observables": Ω_0 and q_0 . For instance, the Modern Cosmological view of the Friedman models is summarized in figure 1. One must note that a strong motivation for the existence of a non-zero cosmological constant comes from observations :

- i) Recent results from the cosmic microwave background anisotropies (CMBA) observations (Maxima, Boomerang, DASI, Archeops, WMAP) : the Universe is nearly flat! One combined with HST Hubble constant measurements request a non-zero cosmological constant (Page et al.; 2003). ii) in 1998, two independent groups - the Supernova Cosmology Project (SCP) and the High z Supernova Search (HzS) - announced a spectacular result based on observations of distant type-Ia supernovae : the Universe is accelerating ! For a critical summary, see Signore & Puy 2001 and for a more recent study, see the lecture given by B.Leibundgut.
- i) Recent results from the cosmic microwave background anisotropies (CMBA) observations (Maxima, Boomerang, DASI, Archeops, WMAP) : the Universe is flat! For a very recent study, see the lecture given by E.Wright. See the figure 1 where the point: "We are here" is put by taking into account SNIa and CMBA results.

Radiation dominated case

As we have seen the density associated with radiation evolves accordingly to $\rho_\gamma a^4 = \text{cste}$. It is clear that this radiation term will be dominant over the matter term for small a , that is at the "beginning". In this regime, if we neglect other terms in the EFL equation, we have:

$$\dot{R} = \left(\frac{8\pi G \rho_1}{3} \right)^{1/2} \frac{R_1^2}{R} \quad (1.5)$$

so that $R \propto t^{1/2}$

An important application of cosmic background radiation is to provide a test of the reality of the expansion: if we are able to measure the temperature of the background at higher redshift it should scale accordingly to:

$$T(z) = T_0(1 + z)$$

It is actually possible to measure the temperature of the background radiation through the observations of the ratio of molecular lines: the ratio of population on two levels for which the difference of energy is only a few Kelvin and provides a sensitive way to actually measure the temperature of the background. Such lines can be detected in the optical domain. Actually the first detection

The Pressure-less Friedmann Models of the Universe
with a cosmological constant

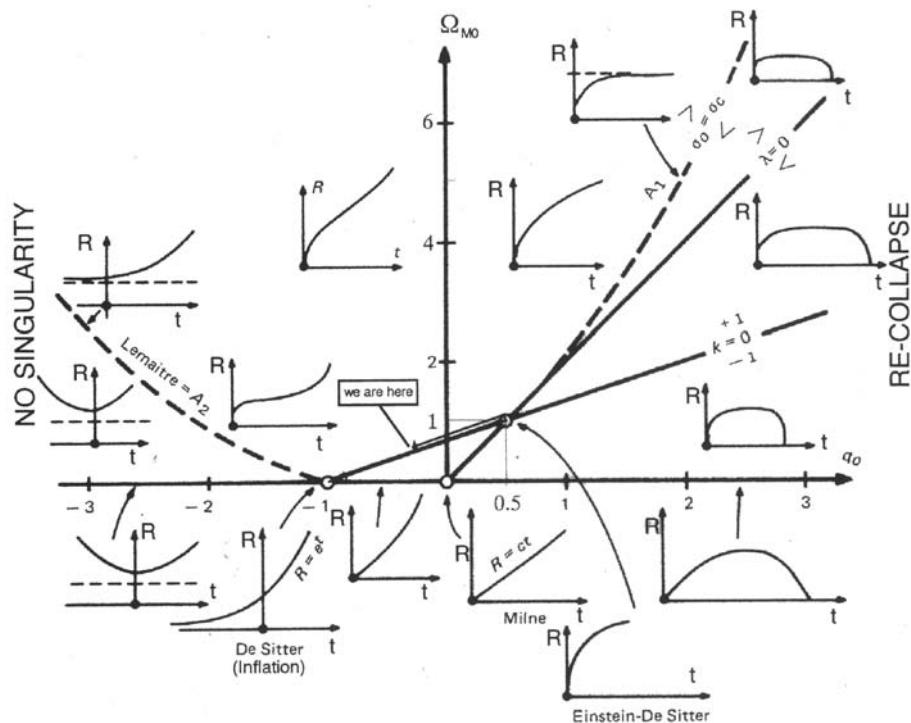


Figure 1.1. Evolution of the scale-factor R with the time, as predicted by EFL equations, as a function of the today values of the parameters Ω (or Ω_M) and q . One notes that at the left of the dashed line (Lemaître limit) there is no initial singularity; on the right side of the plot there are recollapsing models. See Melchiorri et al. 2003.

of the CBR was obtained by this method. It has also been successfully applied to distant QSO's and the result are consistent with the expanding picture (Songaila et al, 1994. See also Lu et al.,1996; Roth & Bauer, 1999; Srianand et al., 2000; Levshakov et al., 2001).

A different approach has been recently attempted by Battistelli (2002), based on an idea proposed by Fabbri & Melchiorri (1978): the signal produced by CMB radiation when interacting with the hot gas in clusters of galaxies is extremely sensitive to the CMB temperature.

Therefore, the most precise verification of the basic cosmological framework now comes from the observations of the spectrum and of the anisotropies of the CMB.

In the following we will consider other observational relics of the early Universe : the primordial abundances of light nuclei.

4. The standard Big Bang Nucleosynthesis (SBBN)

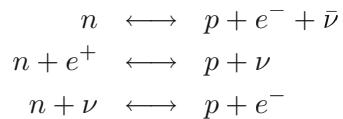
The agreement between the predictions of the SBBN model for the abundances of deuterium, helium-4 , helium-3 and lithium-7 and the observations of the “primordial abundances” of these light elements is one of the successes and therefore one of the cornerstones of the Big Bang Cosmology.

The model has only one parameter, the baryon-to-photon ratio, η . These abundances, which depend only on η , span some nine orders of magnitude. The synthesis of these light elements strongly depends on the physical conditions of the early Universe, at T about 1 MeV, when t is around 1 second.

Here, let us only remark that for a precise description of the state of the thermal equilibrium and of the conditions of decoupling of the various particles of the early Universe see Signore & Puy (1999) and more recently Melchiorri et al. (2003) .

The neutron-to-proton ratio.

When $T > 1$ MeV, neutrinos, electrons, positrons were still in equilibrium through the ($n-p$) reactions :



The ratio of the neutron and proton densities is given approximately by the Saha relation :

$$\frac{n}{p} = \exp\left(-\frac{Q}{T}\right) \quad (1.6)$$

where Q is the proton-neutron mass difference :

$$Q = m_n - m_p = 1.293 \text{ MeV}$$

At very high T ($\gg 1 \text{ MeV}$), the ratio $\frac{n}{p}$ is around 1. As the temperature falls, the ratio $\frac{n}{p}$ follows the equilibrium curve until the reaction rate $\Gamma(T)$ falls below the expansion rate $H(T)$; at the temperature T^* such that :

$$\Gamma(T^*) = H(T^*) \quad (1.7)$$

the ratio $\frac{n}{p}$ will get “frozen” at the value :

$$\left(\frac{n}{p}\right)^* = \exp\left(-\frac{Q}{T^*}\right)$$

Calculations of $\Gamma(T)$, $H(T)$ lead to the value :

$$T^* = 1.8 \text{ MeV} \quad (1.8)$$

at t around 1 second, with :

$$\left(\frac{n}{p}\right)^* = \exp\left(-\frac{Q}{T^*}\right) \sim \frac{1}{6} \quad (1.9)$$

Due to occasional weak interactions, eventually dominated by free neutron-decays, the ratio $\frac{n}{p}$ decreases until : $t = t_{nuc}$, for which the nucleosynthesis occurs and then :

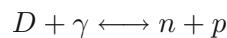
$$\left(\frac{n}{p}\right)_{nuc} \sim \frac{1}{7} \quad (1.10)$$

Helium synthesis.

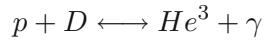
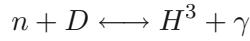
As soon as n/p leaves its equilibrium curve i.e. for : $\frac{n}{p} = \left(\frac{n}{p}\right)^* = \frac{1}{6}$, neutrons and protons collide to make deuterium :



But as long as $T > 0.1 \text{ MeV}$, (or $t < 100$ seconds), the photons have enough energy to photodissociate deuterium through the following reaction :

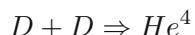


a reaction which is faster than the other possible reactions :



Therefore not much helium is produced as long as $T > 0.1$ MeV.

Below this temperature, photodesintegration becomes inefficient, then, more complex nuclei than deuterium are built through the following reactions :



A rough estimate of the He^4 production can be done:

$$Y(He^4) = (4nHe)/(n_{tot}) = 4(n_n/2)/(n_n + n_p)$$

$$= 2(n_n/n_p)/(1 + n_n/n_p) = (2n/p)/(1 + n/p)$$

with the value of (n/p) of $1/7$ at $t = t_{nuc}$, therefore :

$$Y(He^4) \sim 0.25$$

To build heavier elements than He^4 , collisions of rare deuterium, tritium, He^3 with He^4 are required. But, most of deuterium, tritium and He^3 are burned to He^4 . In fact, very little synthesis of heavier elements will be done. More precisely, we will see that the calculated abundances of He^3 and deuterium are of the order of 10^{-4} to 10^{-5} , while those of Li^7 are of the order of 10^{-9} to 10^{-10} .

Heavier elements are produced in truly negligible quantities since there are no stable elements with $A = 5$ and $A = 8$ that could serve as intermediate steps. In fact, there are only some traces of beryllium-9 and boron-11.

As the Universe continues to expand and cool, the density and the temperature of the Universe continue to decrease. Very soon, there is no more possible nuclear reactions. At $t \simeq 5$ minutes, the epoch of BBN is definitively over! Let us now briefly present the predictions of the sBBN model.

Calculated abundances of the light elements.

As said above, the sBBN model depends on only one parameter η , the ratio of the number of baryons n_b to the number of photons n_γ :

$$\eta = (n_b/n_\gamma) \quad \text{and} \quad \eta_{10} = 10^{10} \eta \quad (1.11)$$

The predicted abundances will depend only on η or η_{10} .

Moreover, since the number of photons of the Universe is known now-a-days from the Cosmic Microwave Background :

$$n_\gamma = 411 \pm 2 \text{ photons cm}^{-3} \quad (1.12)$$

η specifies the dimensionless present-day baryon density Ω_{b0} through:

$$\Omega_{b0}h^2 = 3.65 \times 10^{-3}\eta_{10} \quad (1.13)$$

Where the subscript "0" on any quantity denotes its present value and h is defined through: $h = H_0 / (100 \text{ km/sec/Mpc})$.

The initial computer code due to Wagoner (1973) has been improved by Kawano and his collaborators who have incorporated new measurements, and revised estimates of nuclear cross sections. This code has been made available by Kawano (1992) and has become for a long time the standard code for Big Bang Nucleosynthesis studies. Now, many groups use various "up-to-date" codes and publish their predictions under a graphic form i.e. a traditional plot of primordial abundances as functions of the nucleon-to-photon ratio, η or Ω_b .

On the other hand, Sarkar (1996) and Hogan (1997) give the following fitting formulae :

- the predicted fraction of total baryon mass in helium-4 :

$$Y_p = 0.235 + 0.012 \ln\left(\frac{\eta_{10}}{2}\right) \left(\frac{\eta_{10}}{2}\right)^{-0.2} + 0.011[1 - \left(\frac{\eta_{10}}{2}\right)^{-0.2}] \pm 0.0006 \quad (1.14)$$

- the abundance by number of deuterium :

$$(D/H)_p = 15.6 \times 10^{-5 \pm 0.03} \times \left(\frac{\eta_{10}}{2}\right)^{-1.6} \quad (1.15)$$

- the abundance of lithium-7 :

$$1.06 \times 10^{-10 \pm 0.1} \left[\left(\frac{\eta_{10}}{2}\right)^{-2.38} + 0.28 \left(\frac{\eta_{10}}{2}\right)^{2.38} \right] \quad (1.16)$$

Let us only notice that the theoretical errors are dominated by uncertainties in nuclear rates.

Let us only note that after 5 minutes - most of neutrons are in helium-4 nuclei, most protons are free; we see also that much smaller amounts of deuterium, helium-3 and lithium-7 are synthetized. Moreover, a lower density, growing Coulomb barriers, a stability gap at masses 5 and 8 will work against the formation of heavier elements. This elemental composition will be unchanged until the formation of the first stars. See the three vertical panels of

figure 2. The four curves show the abundance ratios predicted for sBBN, from the calculations due to Burles, Nollett & Turner (2001).

5. Observations of “primordial abundances”.

The primordial abundances of the light elements are not measured easily and simultaneously. The main difficulties come from systematic uncertainties in inferring abundances from observations and in modeling their chemical evolution since the Big Bang.

Primordial abundance of helium-4.

i) *Observations*

Helium-4 can be observed in galactic and extragalactic HII regions - regions of hot and ionized gas - using either optical or radio recombination lines. However the best determinations come from observations of $\text{HeII} \rightarrow \text{HeI}$ recombination lines in extragalactic HII regions.

ii) *Conversion of observed lines into abundances.*

Deriving an abundance from the observed lines should be straightforward. However, corrections must be applied to compensate for excitation effects.

iii) *Correction for chemical evolution since the Big Bang.*

In stars, most of the hydrogen is converted into helium-4 which is converted into heavier elements. Then, an excess of helium-4 can return into the interstellar medium. Therefore, one must account for a possible helium enrichment.

iv) *The “primordial abundance” of helium-4.*

There is a lot of recent and independent observations and analyses leading to some values of Y_p . In particular, the value inferred by Izotov and Thuan (1998) is :

$$Y_p = 0.244 \pm 0.002 \pm 0.005 \quad (1.17)$$

Note that the first uncertainty is statistical while the second is systematic.

Let us also give the lower value found by Olive, Steigman & Skillman (1997) :

$$Y_p = 0.238 \pm 0.002 \pm 0.005 \quad (1.18)$$

Primordial abundance of lithium-7

i) *Observations*

Lithium absorption lines can be relatively easily observed and measured in stars. There are more than 100 hot, population-II halo stars which have been observed since the first measurements of Spite and Spite (1982)

ii) *Conversion of observed lines in abundances.*

The observed lines gives the stellar photospheric abundance via the standard stellar atmospheric technique. For stars with a surface temperature $T > 5500$ K and a metallicity less than about 1/20th the solar metallicity, the abundances practically show no dispersion (the famous “lithium plateau” for such stars.)

iii) *Correction for chemical evolution since the Big Bang.*

The problem arises in relating the observed abundance to the cosmic abundance, i.e. the abundance of the star when it formed. There is a possibility that lithium has been depleted in these stars though the lack of dispersion in the lithium data - at least in the hotter stars - limits the amount of depletion. On the other hand, since the Big Bang, lithium-7 has been also produced through spallation of nuclei by cosmic rays and also synthesized in novae and type-II supernovae - via the neutrino process. Ryan et al. (2000) have undertaken some observations - in particular studies of possible correlations between Li and Fe - to constrain the evolution of lithium-7.

iv) *The “primordial abundance” of lithium-7.*

Finally, the lithium data (Ryan et al, 2000) lead to the following “primordial abundance” :

$$(Li/H)_P = (1.23 \pm 0.06^{+0.68}_{-0.32}) \times (10^{-10}) \quad (1.19)$$

Primordial abundance of deuterium.

Observations for deuterium present certain difficulties : all deuterium is primordial, but some of this deuterium has been destroyed. During a long time, astronomers considered the pre-solar system abundances of :

$$\frac{He^3}{H} \quad \text{and} \quad \frac{He^3 + D}{H}$$

from measurements on meteoritic near-surfaces samples, the solar wind and lunar soil samples to obtain, by difference, the pre-solar D/H abundance. They also observed deuterated molecules, such as HD, in the atmosphere of planets; for example, there is a recent measurement of HD in the atmosphere of Jupiter which is perfectly consistent with the above pre-solar measurements of D/H. There are also many measurements of the D/H ratio in the interstellar medium done first with the *Copernicus satellite*, then with the *Hubble Space Telescope* and more recently with FUSE (the *Far Ultraviolet Spectroscopic Explorer* launched in June 1999). But we do not know the history of the Galaxy well enough to reconstruct the primordial D from Galactic observations, while it is possible to detect D in high-redshift, low-metallicity quasar absorption systems.

In fact, the D/H has been recently measured in several high- z hydrogen clouds. These hydrogen clouds are “seen” by their distinctive Lyman-alpha features in the spectra of several QSOs. Here, we consider the new results reported by Kirkman et al. (2003) : they announce the detection of D in an absorption system at $z=2.526$ in the spectrum of Q1243+3047 ; they show how they have improved the exploration of some factors which determine the measurement errors . We adopt their best estimate of the primordial D/H values towards five QSOs :

$$(D/H)_p = (2.40 - 3.22) \times 10^{-5} \quad (1.20)$$

6. Confrontation of the observed “primordial abundances” to the predictions of the sBBN.

Since the beginning of the 1980’s cosmologists introduce the notion of : “concordance interval” for $\Omega_b h^2$, ρ_b or η , for which predicted and measured abundances of helium-4, lithium-7 and deuterium are consistent within their uncertainties.

Then, in the 1990’ s, the accurate determination of the primeval deuterium abundance changes the strategy. Schramm and his collaborators introduced the “deuteronomy” (see Schramm & Turner, 1997). Also, because the Big Bang deuterium production decreases rapidly with baryon density and its post Big Bang history (or chemical evolution) is simple – stars only destroy D – primordial deuterium became the best baryometer. It pegged the baryon density and led to accurate predictions for the other light elements.

On this figure 2, we see the four plots of the primordial abundances of ^4He , D, ^3He and ^7Li as predicted by the standard model of Big Bang nucleosynthesis (Burles, Nollett & Turner, 2001). We see also the boxes which indicate the observed light elements :

- for ^4He , the larger is from Olive et al. (1997), the smaller from Izotov &

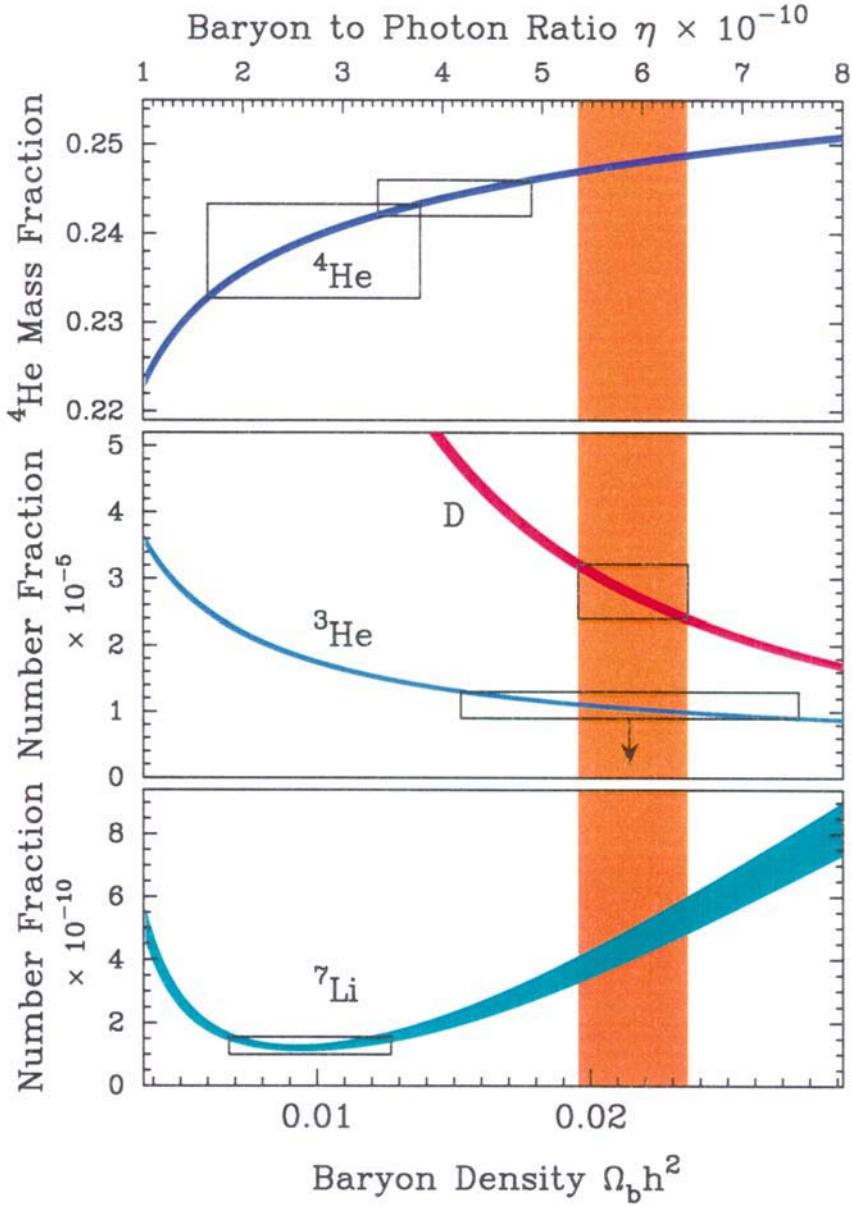


Figure 1.2. A/H for deuterium, helium-3, lithium-7 and Y_p versus Ω_b : - the four curves show the predicted abundances by the sBBN model, - the horizontal boxes show the various measurements, - the vertical band covers the D/H data. From Kirkman et al. 2003

Thuan (2000),

- for D ,the box is the mean from the five QSOs of Kirkman et al. (2003),
- for ^3He , the box is from Bania et al. (2002),
- for ^7Li , the box is from Ryan et al. (2000),

Again, we see the vertical band, inferred by the deuterium measurements, which can pin down the baryon density (Kirkman et al. 2003):

$$\Omega_b h^2 = 0.0194 - 0.0234 \quad (1.21)$$

We expected that all the boxes should overlap the vertical band that covers the D/H data. Finally, we see that they do not : measurements of primordial ^3He are consistent but all ^7Li and most ^4He measurements prefer lower $\Omega_b h^2$. Probably, because of systematic errors, Kirkman et al. dixit ! Much more observational work and analysis of data are needed .

Nevertheless, for the moment, we prefer the $\Omega_b h^2$ value from D/H measurements due to Kirkman et al.2003, than the lower values from ^4He and ^7Li . Indeed Deuterium is the most sensitive "baryometer" among the light nuclei. Finally, the baryon density can be constrained efficiently from CMB measurements (Le Dour et al., 2000), leading to values in agreement with Deuterium measurements See in particular recent constraints obtained from BOOMERANG, DASI, MAXIMA, VSA, CBI, ARCHEOPS and WMAP data. : de Bernardis (2002), Netterfield et al. (2002), Prike et al. (2002), Rubiño-Martin et al.(2003), Sievers et al. (2003), Wang et al. (2002), Benoît et al. (2002), Bennett et al. (2003), Spergel et al. (2003) and the lecture given by E.Wright in this volume.

7. Conclusions

In summary, the thermal history of the early Universe is very simple. It just assumes a global isotropic and uniform Universe. In its simplest version - no structure of any kind on scales larger than individual particles - the contents of the Universe are determined by "standard elementary physics" : i) global expansion governed by GR, ii) particles interactions governed by the "Standard Model " of Particle Physics , iii) distributions of particles governed by the laws of Statistical Physics.

Now, there is a raft of questions to be addressed (Peebles, 2003) like :

- why is the Universe expanding ?
- why is there today a net asymmetry between matter and anti-matter ?
- why do the various forms of energy and matter (baryonic matter,non-baryonic matter , dark energy) have almost the same density today ?
- when did the first objects formed ?
- what is the physics relevant to the early universe ?

And many other questions which should be considered as challenges for future researches.

References

- Bania, T. M., Rood, R. T. & Balser, D. S., 2002, Nature, 415, 549.
- Battistelli, E. S., et al. 2002, ApJL, 580, L101
- Battistelli E.S. et al., astro-ph/0208027
- Bennett, C. L., et al. 2003, ApJS, 148, 97
- Benoît, A., et al. 2003, A&A, 399, L19
- Burles, S., Nollett, K. M. & Turner M. S., 2001, ApJ, 552, L1.
- Caldwell, R. R., Dave, R., & Steinhardt, P. J. 1998, Phys. Rev. Lett., 80, 1582
- Cayrel, R., 2001, Nature, 409, 691.
- Chaboyer, B., 1995, ApJ 444 L9.
- de Bernardis, P., et al. 2002, ApJ, 564, 559
- Fabbri, R. & Melchiorri, F., 1979, ApJ 78, 376
- Freedman, W. L. & Turner, M. S. 2003, Rev. Mod. Phys., 75, 1433
- Harrison E.R., 1981, "COSMOLOGY", Cambridge University Press, Cambridge.
- Hogan, C. J., 1998, in "18th Texas Symposium", Olinto et al. eds., World Scientific Singapore, p.15.
- Izotov, Y. I. & Thuan, T. X., 1998, ApJ 500, 188.
- Kavano, L., 1992, FERMILAB PUB 92/04A
- Kirkman, D., Tytler, D., Suzuki, N., O'Meara, J. M., & Lubin, D. 2003, ApJS, 149, 1
- Le Dour, M., Douspis, M., Bartlett, J. G., & Blanchard, A. 2000, A&A, 364, 369
- Levshakov, S. A., Dessauges-Zavadsky, M., D'Odorico, S., & Molaro, P. 2002, ApJ, 565, 696
- Lu, L. et al., 1996, ApJS, 27, 475.
- Melchiorri, F., Olivo-Melchiorri, B. & Signore, M., 2003, Riv. Nuovo Cim. , in press.
- Netterfield, C. B. et al., 2002, ApJ, 571, 604, astro-ph/0104460
- Olive, K. A., Steigman, G. & Skillman, E. D., 1997, ApJ., 483, 788.
- Page, L., et al. 2003, , 148, 233
- Peebles, P.J.E. 2003, astro-ph/0311435
- Peebles, P. J. & Ratra, B. 2003, Rev. Mod. Phys., 75, 559
- Pryke, C. et al., 2002, ApJ. 568, 46, astro-ph/0104490
- Ratra, B. & Peebles, P. J. E., 1988, Phys. Rev. D37 3406.
- Roth, K. C. & Bauer, J. M. 1999, ApJL, 515, L57
- Rubiño-Martin, J. A., et al. 2003, MNRAS, 341, 1084
- Ryan, S. G. et al., 2000, ApJ, 530, L57
- Sarkar, S., 1996, Rep.Prog.Phys., 59, 1493
- Schramm, D. N. & Turner, M. S. 1996, Nature, 381, 193
- Sievers, J. L., et al. 2003, ApJ, 591, 599
- Signore, M. & Puy, D., 1999, New Astr.Rev., 43, 185

- Signore, M. & Puy, D., 2001, New Astr.Rev., 45, 409
Songaila, A. et al., 1994, Nature, 371,43.
Spergel, D. N., Verde, L., Peiris, H. V. et al. 2003, ApJS 148, 175
Spite, F. & Spite, M., 1982, A&A 357, 115.
Srianand, R., Petitjean, P., & Ledoux, C. 2000, Nature, 408, 931
Wagoner R. V., 1973, ApJ 179, 343.
Wang, X. et al., 2002, Phys. Rev. D65, 123001.
Wetterich, C., 1988, Nucl. Phys. B302, 668.
Weinberg, S., 1972, *Gravitation and Cosmology*, New york: Wiley.

Chapter 2

THE X-RAY VIEW OF GALAXY CLUSTERS

William Forman

*Smithsonian Astrophysical Observatory
Harvard-Smithsonian Center for Astrophysics
Cambridge, MA, USA
wrf@cfa.harvard.edu*

Introduction

Clusters of galaxies are the most massive collapsed systems in the Universe. A typical luminous cluster (e.g., Coma cluster) is filled with a hot, 100 million degree, low density (10^{-3} cm^{-3}) gas. In addition to the optically luminous galaxies and diffuse X-ray gas, clusters are dominated by dark matter. The X-ray gas, relaxing on the relatively short sound crossing time, $\tau_{Mpc} = 6.6 \times 10^8 (T/10^8) \text{ yrs}$ (where $\tau = D/c_s$ and $c_s^2 = \gamma P/\rho$) is an effective tracer of this unseen dark matter.

Cluster studies provide a unique window on the Universe. Although clusters are the most massive collapsed systems in the Universe, with dynamical timescales of order 10^9 yrs in their cores, they are relatively young and “remember” the conditions from which they formed. Also, clusters form from rare overdensities. Therefore, cluster properties and numbers are sensitive to cosmological parameters.

Clusters form from large volumes. For a Universe with a mean mass density of $3 \times 10^{-30} \text{ gm cm}^{-3}$ (30% of closure density; $\rho_c = 3H_0^2/8\pi G$, with $H_0=70 \text{ km s}^{-1}$), a rich cluster with a mass of $10^{15} M_\odot$ forms from a sphere with a radius of $\sim 20 \text{ Mpc}$. Since the dominant process in the formation of the cluster whose mass consists of cold dark matter and baryons is gravitation and the formation of collapsed objects by gravity alone should not affect the ratio of the mass components, it is believed that the mass components of today’s clusters are representative of the Universe (e.g., White et al. 1993).

The mass fluctuation spectrum gives rise to collapsed objects from galaxies to groups to rich clusters. The systems dominated by old stars – early type galaxies, elliptical-dominated groups, and rich clusters – all have gaseous halos

(Forman et al. 1985, Helsdon et al. 2001, Jones & Forman 1999). Table 2.1 summarizes the properties of these systems.

Table 2.1. Typical Properties of Gas Rich Systems

	Galaxies	Groups	Clusters
L_x (ergs s ⁻¹)	10^{40-42}	10^{42-43}	10^{43-46}
kT (keV)	0.5-1.0	1-3	2 -15
M_{total} (M_\odot)	$10^{11} - 10^{12}$	$10^{12} - 10^{13}$	$10^{13} - 10^{15}$

Unique X-ray Cluster Properties

Why use X-rays for cluster studies since they are relatively difficult to observe and require space-based observatories? As clusters provide a unique window on the Universe, X-ray observations provide a unique window on clusters.

X-ray observations detect the dominant baryonic component in clusters. While clusters were first detected optically, the dominant baryonic component in clusters is the X-ray emitting gas. Despite its low mean density, (10^{-2} to 10^{-3} cm⁻³), the gas mass in a rich cluster exceeds the optically luminous matter in the galaxies by a factor of 3-5. Thus, to “see” the bulk of the baryons that have temperatures of 10 – 100 million degrees, X-ray observations provide the only direct approach.

X-ray observations detect deep gravitational potentials. The X-ray emission from galaxy clusters is optically thin thermal bremsstrahlung. The luminosity of a cluster is given as:

$$L_x = \int n_e n_p \Lambda(T, Z) dV \quad (2.1)$$

where n_e and n_p are the electron and proton densities, $\Lambda(T, Z)$ depends both on gas temperature and elemental abundance. The integral is taken over the cluster volume. Since the cluster luminosity depends on the square of the density, X-ray observations are less affected by superpositions of smaller objects along the line of sight. Because the Universe is highly structured – clusters form at the intersections of filaments and sheets and groups and smaller clusters lie along filaments and sheets – the probability of superposition of substructures is much higher than in a uniform Universe. However, the X-ray luminosity of clusters has a steep dependence on gas temperature (e.g., $L_{bol} \propto T_{gas}^{2.5}$) and for a gravitationally collapsed cluster the gas temperature has a strong dependence on galaxy velocity dispersion ($T_{gas} \propto \sigma^2$), then smaller groups have much lower X-ray luminosities compared to rich clusters and hence, projection effects are not as severe in X-rays as in the optical.

While the steep dependence of X-ray luminosity on the mass of the system (equivalently σ^2) facilitates the study of luminous, rich clusters, poor groups with low masses (and low velocity dispersions) are inherently X-ray faint.

Precision of X-ray observations. X-rays do not penetrate the Earth's atmosphere and so all X-ray observations are performed from satellites (or balloons for hard X-rays). However, compared to optical observations, X-ray studies are not intrinsically limited by the number of cluster member galaxies. As an X-ray observation is made longer and longer, more and more X-ray photons are detected and measurements made from the observation can be made more and more precise (up to the systematic calibration of the detectors).

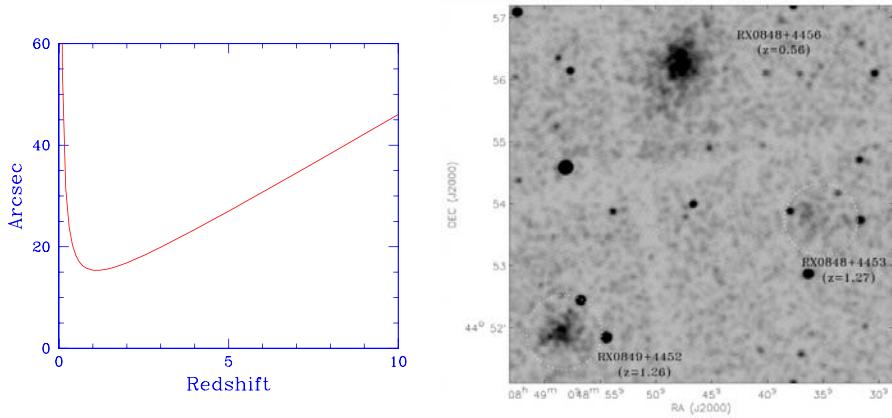


Figure 2.1. (left: a) The angle corresponding to a 100 kpc cluster core diameter as a function of redshift z (for a cosmology with $\Omega_m = 0.3$, $\Lambda = 0.7$, and $H_0 = 70 \text{ km s}^{-1}$). For moderate angular resolution X-ray telescopes, cluster cores are resolved at *all* redshifts. (right: b) X-ray image of a deep, 180 ksec Chandra observation. In the image, three clusters are detected with redshifts $z = 1.26$, 1.27 , 0.58 (from Stanford et al. 2001). The figure emphasizes the ease of selecting extended cluster emission in X-ray images.

X-ray Clusters Resolved at All Redshifts Another key feature of X-ray cluster studies is that rich clusters are resolved by moderate resolution telescopes at all redshifts. For example, the core of a rich cluster of diameter 100 kpc at $z = 1$ would subtend an angle of $15''$ (for a cosmology with $\Omega_m = 0.3$, $\Lambda = 0.7$, and $H_0 = 70 \text{ km s}^{-1}$). The variation of angular diameter of a cluster core as a function of redshift is shown in Fig. 2.1a and shows a minimum that can be resolved by present day telescopes.

The efficacy of X-ray detection of distant clusters is shown in Fig. 2.1. The figure shows three clusters with redshifts ranging from $z = 0.58$ to $z = 1.27$. The X-ray sources are easily seen and readily identified as extended compared

to the point-like sources spread over the field. X-ray surveys are one of the best ways to detect moderately distant clusters.

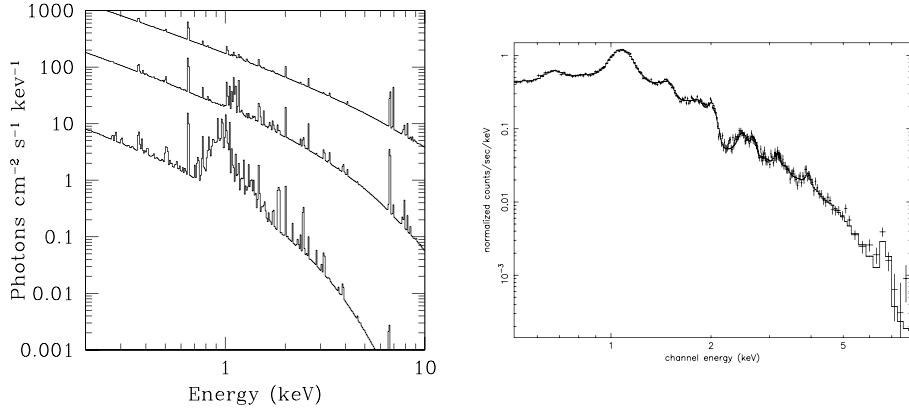


Figure 2.2. (a) Model spectra from optically thin gas of solar abundance with temperatures of 1 (lower graph), 3 (middle graph), and 10 keV (top graph). The characteristic exponential dependence on gas temperature is seen in each model as is the decline in importance of line emission with increasing gas temperature. (b) The observed Chandra spectrum of a region of 2 keV gas surrounding the dominant galaxy M87 in the Virgo cluster. Prominent line contributions include: O (shallow peak from 0.6–0.7 keV), Ne and Fe (large peak at 1.1 keV), Mg and Fe (at 1.5 keV), Si (2 keV), S (2.5 keV), Ar (just above 3 keV), Ca (3.9 keV), Fe (6.7 keV). The sharp decrease in the spectrum at 2.2 keV is produced by a decrease in the mirror effective area.

Optically thin thermal emission is “simple” and well-understood. The X-ray emission process is well-understood physically. The emission process is thermal bremsstrahlung with contributions from emission lines (that become less important as the temperature increases):

$$dN(E)/dE \propto n_e n_p \exp(-E/kT) + \text{line emission} \quad (2.2)$$

where E is the photon energy and n_e and n_p are the electron and proton densities. The exponential dependence on temperature allows accurate determination of the gas temperature and the line emission contributions allow measurement of elemental abundances. Typical spectra are shown in Fig. 2.2 and show 1) the exponential dependence on gas temperature and 2) the decreasing contribution from emission lines as the gas temperature increases and atoms become more fully ionized. Integrated over energy, the emissivity of hot gas ($t > 3 \times 10^7$ K) is approximated by $\epsilon \approx 3 \times 10^{-27} T_{\text{gas}}^{1/2} n_p^2 \text{ erg cm}^{-3} \text{ s}^{-1}$.

A variety of computer codes are used to describe X-ray spectra. Combined with fitting routines that account for the telescope and detector responses, the

codes, coupled to fitting tools, allow determination of the gas temperature and abundances of individual elements, when the statistical precision of the observation is high. Popular codes include those by Raymond & Smith (1976, 1977) with numerous updates, Mewe, Gronenschild, van den Oord (1985; MEKAL; Mewe-Kaastra-Liedahl) with revisions to low temperature emission lines and including a total of 2131 spectral lines and APEC, with additional emission lines and improvements in wavelengths. Other packages are Chianti, Cloudy, SPEX, and XSTAR. Fig. 2.2b shows the results from the software package XSPEC using the MEKAL code with a fit to the X-ray spectrum of the central cluster galaxy M87. Surrounded by hot gas with a temperature of about 2 keV, the gas shows the exponential continuum with a set of emission lines.

1. Observing Clusters in X-rays – the Chandra Observatory

The key features of the Chandra Observatory are shown in Fig. 2.3 and include the solar arrays for power, aspect cameras for reconstructing the pointing direction, the mirror assembly for collecting photons and creating the high angular resolution image, the instrument module that houses the two Chandra detectors, and the optical bench that maintains the alignment of the mirrors and detectors.

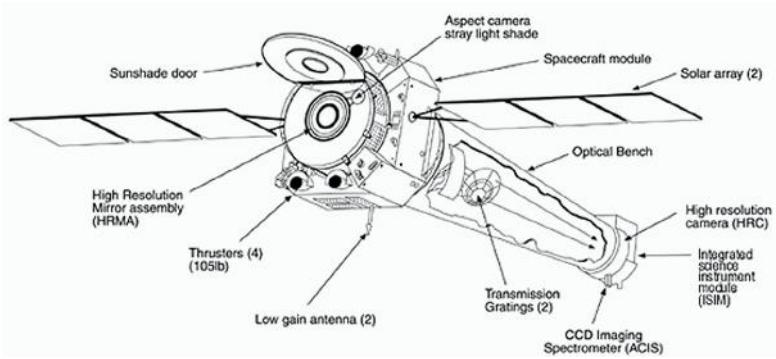


Figure 2.3. A schematic of the Chandra Observatory showing the typical components including the solar arrays, aspect cameras, mirror assembly, instrument module, and optical bench. The Chandra Observatory weighs 4800 kg and orbits the Earth every 64 hours in a 10,000 by 140,000 km orbit that extends 1/3 of the distance the moon. Including the solar panels, the observatory's dimensions are 14m by 20m.

The polishing, construction and alignment of the Chandra mirrors and optical bench were critical for achieving high angular resolution. First, because of the need for grazing incidence reflection, the effective telescope area is small

compared to the polished area. For Chandra, the polished area is 200,000 cm² while the effective area is only 1000 cm². The polishing and figuring of the telescope are such that if the mirror were enlarged to the size of Earth, the largest imperfection would be only 2 m high. At their final cleaning the mirrors had 1 dust particle per 1000 cm² of surface and were aligned to better than 0.1" over their three meter length (1/50 the width of a human hair). The mirrors are coated with iridium to enhance their reflectivity.

As with most high angular resolution telescopes, the Chandra mirror module consists of four nested parabola-hyperbola mirrors pairs. X-ray photons are reflected at small grazing incidence angles ($\sim 1^\circ$). While the telescope maintains its pointing position reasonably accurately, the high angular resolution images are produced by time tagging the arrival of each photon (characterized by its position on the detector and a measure of its energy). The image is then constructed by using the aspect solution, derived from the motions of stars in the aspect camera, to locate the detected photons at their proper position in sky coordinates.

Detectors

The two detectors on Chandra complement each other. The High Resolution Camera (HRC) provides a large field of view, high time resolution, but little energy resolution. The Advanced CCD Imaging Spectrometer (ACIS) provides very modest time resolution, good spectral resolution, and modest field of view.

The primary components of the HRC are two Micro-Channel Plates (MCP). They each consist of a 10 cm square cluster of 69 million lead-oxide glass tubes that are about 10 micrometers in diameter (1/8 the thickness of a human hair) and 1.2 millimeters long. The tubes have a special coating that causes electrons to be released when the tubes are struck by X-rays. These electrons are accelerated down the tube by a high voltage, releasing more electrons as they bounce off the sides of the tube. By the time they leave the end of the tube, they have created a cloud of thirty million electrons. A crossed grid of wires detects this electronic signal and allows the position of the original X-ray to be determined with high precision.

The Chandra Advanced CCD Imaging Spectrometer (ACIS) is an array of charged coupled devices (CCD's). This instrument is especially useful because it can simultaneously make X-ray images, and at the same time, accurately measure the energy of each incoming X-ray. It is the instrument of choice for studying temperature and abundance variations across extended X-ray sources.

Cluster Morphologies

X-ray clusters can be classified into several broad groups. Forman & Jones (1992) used Einstein X-ray observations of over 200 clusters (see Fig. 2.4)

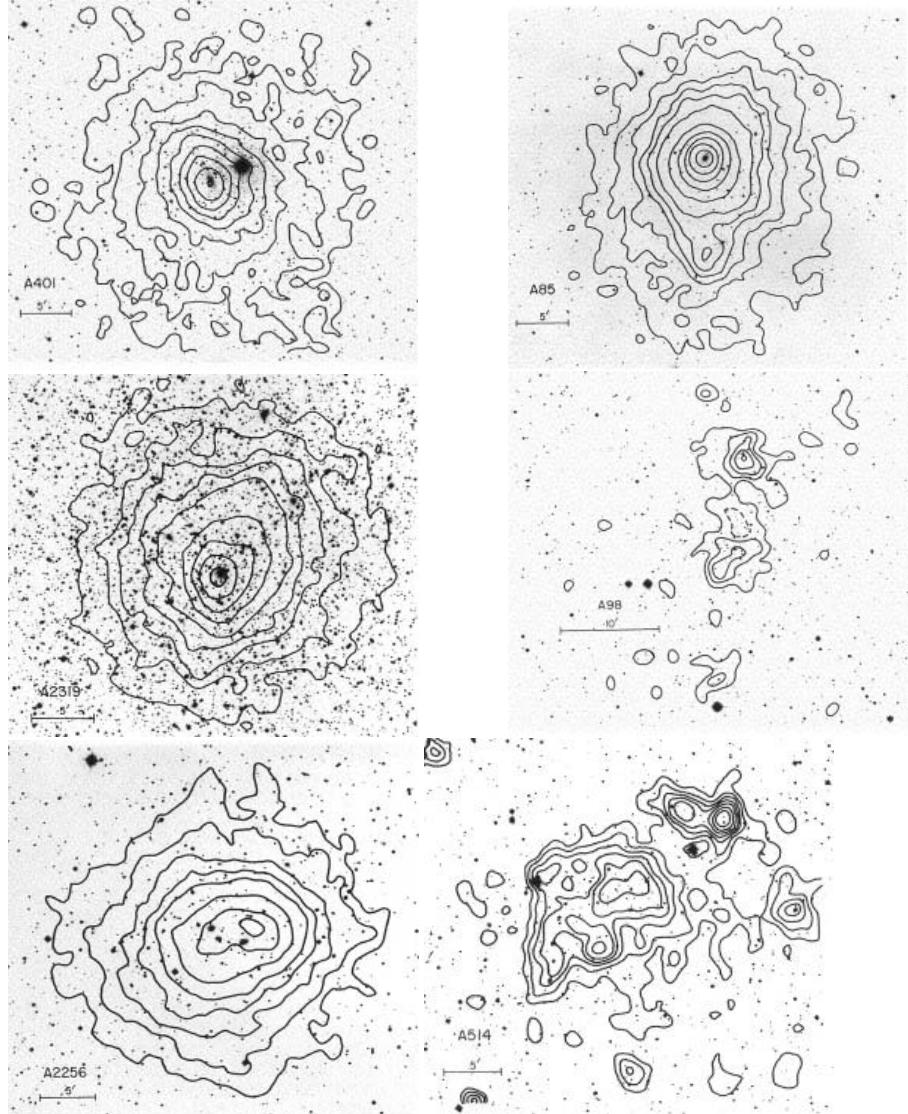


Figure 2.4. Cluster contour maps superposed on optical images (see Jones & Forman 1999 for details). The clusters illustrate the morphological classes described in the text and listed in Table 2.2. The last class, “G”, where the cluster emission is dominated by that from individual galaxies, is not shown.

to define seven morphological classes. These classes are “single” clusters in which no substructure or departures from symmetry are found, “double” (clusters with two subclusters of comparable size and luminosity), “primary with small secondary” (clusters in which one of the two subclusters is at least two

times brighter), “complex” (clusters with more than two subclusters), “elliptical” (clusters whose X-ray contours are elliptical, rather than circular), “offset center” (clusters whose peak emission does not lie at the cluster center as determined from lower surface brightness emission and “galaxy” (in which the emission is dominated by a single galaxy). Table 2.2 lists the cluster type, the number and percentage of clusters in each morphological class, and their mean X-ray luminosity.

The morphological classification of clusters is closely related to the study of substructure. All except the “single” and “galaxy” classes contain clusters with significant structure. While the categories of “double”, “complex”, and “primary with small secondary” show obvious structure, the “elliptical” and “offset center” clusters generally also are structured systems as demonstrated by detailed X-ray and optical studies of clusters in these categories including Coma, A754, and A2256 (e.g., Mohr et al. 1995).

As Table 2.2 shows, substructure in clusters is indeed common. Those classes clearly exhibiting substructure (double “d”, primary with small secondary “p”, and complex “c”), comprise 22% of the classifiable sample. Including elliptical (“e”) clusters (e.g. A2256-like) brings the fraction of clusters with substructure to 36%. Finally, including clusters with offset centers (“o”, e.g. A2319-like) gives a fraction of 41%. The limited resolution of the Einstein IPC and the bias in classifying faint or poorly observed clusters as centrally peaked and symmetric suggest 41% is a lower limit to the number of clusters with substructure.

Cluster Formation

Clusters form and grow from small density fluctuations by hierarchical merging where small groups merge to form larger clusters. Before the advent of X-ray imaging observatories, clusters were generally believed to be old, relaxed systems. X-ray observations, followed by extensive optical studies (e.g., Geller and Beers 1982), showed that while the cores of rich clusters are dynamically old, mergers continue to the present day. A particularly clear description of the growth of clusters is shown in Fig. 2.5 which shows the merging history of a rich cluster derived by Cavaliere, Menci & Tozzi (1999). The figure emphasizes that clusters undergo many small mergers, rare major mergers, and that in between these merger events, cluster reside in approximate equilibrium. Cavaliere et al. referred to this type of evolution as “punctuated equilibrium”. It is worth noting that major mergers between two large clusters involve up to 10^{63} ergs and are the most energetic events in the Universe since the Big Bang.

A Classic Supersonic Merger - 1E0657 1E0657 ($z = 0.296$) was discovered by Tucker et al. (1995) as part of a search for X-ray bright, but optically poor, clusters. The Chandra image of 1E0657 shows the classic properties of

Table 2.2. Frequencies of Cluster Morphological Classes

X-ray Morphological Class	Example	Number	Percent	Mean L_x [†]
S – single symmetric peak	A401	120	56	29.3
O – offset center	A2319	10	5	44.5
E – elliptical	A2256	31	14	30.8
C – complex, multiple structures	A514	27	13	9.6
D – double (roughly equal components)	A98	13	6	22.1
P – primary with a small secondary	A85	7	3	19.9
G – primarily galaxy emission	A2666	7	3	0.4

[†] $\times 10^{43}$ ergs s⁻¹

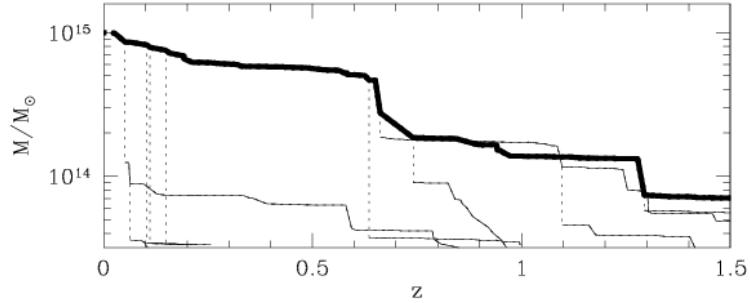


Figure 2.5. A Monte Carlo calculation showing the merging history of a DM halo with final mass comparable to the Coma cluster ($10^{15} M_\odot$) from Cavaliere, Menci & Tozzi (1999). The solid heavy line shows the mass as a function of redshift of the primary cluster. The lighter solid lines show the growth of subclusters that eventually merge into the main cluster. The merger epochs are indicated by the vertical dotted lines. The figure shows that there are episodes of near equilibrium punctuated by major merging events.

a supersonic merger (see Markevitch et al. 2002b and Markevitch et al. 2003a for a detailed discussion). The image (see Fig. 2.6) shows a dense (cold) core moving to the west after having traversed the core of the main cluster. Leading the cold, dense core is a density discontinuity that appears as a shock front. The shock is confirmed by the spectral data since the gas to the east (trailing the shock) is hotter than that in front of the discontinuity (see Fig. 2.6. The detailed gas density parameters confirm that the “bullet” is moving to the west with a velocity of 4500 ± 1100 km sec⁻¹, approximately 3.2 times the sound speed of the ambient gas.

Markevitch et al. (2003a), see also Clowe et al. (2003), combined the dark matter distribution from weak lensing observations with the X-ray observations to derive upper limits on the dark matter cross section. The strongest limit was derived from the constancy of the mass-to-light ratio of the merg-

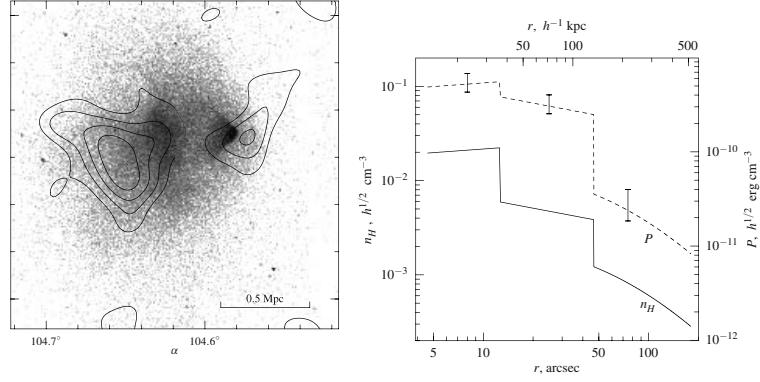


Figure 2.6. (left) Overlay of the mass (derived from weak lensing) contours on the X-ray image (see Markevitch et al. 2003a). The gas bullet lags behind the dark matter subcluster. Combining lensing observations, which give the dark matter distribution combined with X-ray and optical properties, provide limits on the cross section of dark matter. (right) The X-ray surface brightness and gas temperature provides measurements of the gas density and pressure across the shock and across the surface of the gas “bullet”. The pressure jump across the shock and the dense core of the bullet are clearly seen.

ing subclump compared to that of the cluster (and other clusters). If the dark matter had a sufficiently large cross section, the dark matter in the merging subcluster would be lost as it traversed the main cluster while the non-interacting galaxies, that provide the light, would remain part of the subcluster. From the measured values of the mass-to-light ratios, Markevitch et al. (2003a) showed that $\sigma/m < 1 \text{ cm}^2 \text{ gm}^{-1}$. This is to be compared, for instance, to the estimates of $\sigma/m \sim 1 - 100 \text{ cm}^2 \text{ gm}^{-1}$ for self-interacting dark matter (Spergel & Steinhart 2000).

Cold Fronts – A3667, the prototype The study of “cold fronts”, contact discontinuities between cooler and hotter gas, began with the launch of Chandra (Markevitch et al. 2000; Vikhlinin et al. 2001a, b). Cold fronts provide a unique opportunity to explore cluster physics. In a study of A3667, Vikhlinin et al. (2001a, b) derived the ram pressure of the ICM on the moving cold front from the gas density and gas temperature. In turn, the ram pressure yielded a measurement of the cold front velocity. The factor of two difference in pressures between the free streaming region and the region immediately inside the cold front implied a cloud velocity of $1430 \pm 290 \text{ km s}^{-1}$ (Mach 1 ± 0.2). In addition, Vikhlinin et al. (2001b) showed that the “edge” of the cold front in A3667 is very sharp – the width of the front was less than $3.5''$ (5 kpc). This sharp edge requires that energy transport across the edge be suppressed, presumably by magnetic fields. Without such suppression, the density discontinuity at the “edge” would be broader since the relevant Coulomb mean free

path for electrons is several times the width of the cold front. Furthermore, Vikhlinin et al. observed that the cold front appears sharp only over a sector of about $\pm 30^\circ$ centered on the direction of motion, while at larger angles, the sharp boundary disappears. The disappearance can be explained by the onset of Kelvin-Helmholtz instabilities, as the ambient ICM gas flows past the moving cold front. To suppress the instability over the inner $\pm 30^\circ$ requires a magnetic field parallel to the boundary with a strength of $7 - 16 \mu\text{G}$ (Vikhlinin et al. 2001b), assuming that the width of the front is very small.

2. Regular Clusters XD Cooling Flows

In the sections above, we have seen that present epoch clusters are currently undergoing small and large mergers. We have focussed on the most spectacular examples of mergers. However, many clusters are regular. As Table 2.2 shows 56% of clusters are “single”, 5% are round, but with offset centers, 14% are elliptical, and 3% have a main peak with a small secondary. When observed with the high angular resolution and large effective area of Chandra or XMM-Newton, clusters show many irregularities, but for many systems the basic regularity of the cluster is its dominant feature. Regular clusters are a profoundly important class and allow all the promise of X-ray observations to be applied to deriving fundamental cluster properties.

Gas and Galaxies in Equilibrium in a Cluster

We begin with an introduction to a simple model first proposed by Cavaliere and Fusco-Femiano (1976). This model assumes that the gas and galaxies are in equilibrium within the same gravitational potential. Through a measure of the X-ray surface brightness, the model relates the gas temperature to the cluster velocity dispersion, measured from the galaxies.

For gas and galaxies in the same gravitational potential defined by $M_{grav}(r)$:

$$\frac{dP_{gal}}{dr} = \frac{-\rho_{gal}GM_{grav}(r)}{r^2} \quad (2.3)$$

and

$$\frac{dP_{gas}}{dr} = \frac{-\rho_{gas}GM_{grav}(r)}{r^2} \quad (2.4)$$

Combining these two equations, we eliminate the mass:

$$\frac{1}{\rho_{gal}} \frac{dP_{gal}}{dr} = \frac{1}{\rho_{gas}} \frac{dP_{gas}}{dr} \quad (2.5)$$

For isothermal gas and galaxies, $P_{gal} = nkT_{gal} = \frac{1}{3}\rho_{gal}v^2$ and $P_{gas} = nkT_{gas} = \rho kT_{gas}/\mu m_p$:

$$\frac{v^2}{3} \frac{1}{\rho_{gal}} \frac{d\rho_{gal}}{dr} = \frac{kT_{gal}}{\mu m_p} \frac{1}{\rho_{gas}} \frac{d\rho_{gas}}{dr} \quad (2.6)$$

$$\beta \frac{d \ln \rho_{gal}}{dr} = \frac{d \ln \rho_{gas}}{dr} \quad (2.7)$$

where $\beta = \mu m_p v^2 / 3kT_{gas}$.

$$\ln \left(\frac{\rho_{gas}}{\rho_{gas}(0)} \right) = \beta \ln \left(\frac{\rho_{gal}}{\rho_{gal}(0)} \right) \quad (2.8)$$

or

$$\frac{\rho_{gas}}{\rho_{gas}(0)} = \left(\frac{\rho_{gal}}{\rho_{gal}(0)} \right)^\beta \quad (2.9)$$

If we next approximate the galaxy distribution as

$$\rho_{gal} = \rho_{gal}(0)(1 + r/a)^{-3/2} \quad (2.10)$$

then the gas distribution is given by

$$\rho_{gas} = \rho_{gas}(0) (1 + (r/a)^2)^{-3\beta/2} \quad (2.11)$$

The X-ray surface brightness is found by integrating the square of the gas density along the line of sight:

$$S(r) = S(0) (1 + (r/a)^2)^{-3\beta+1/2} \quad (2.12)$$

Thus, by fitting the gas surface brightness distribution, one can measure $\beta = \mu m_p v^2 / 3kT_{gas}$, the ratio of the energy per unit mass in galaxies to that in the gas. With the gas temperature, one can then estimate the galaxy velocity dispersion or with both the gas temperature and galaxy velocity dispersion, one can test the simple model for the gas. With a measurement of the X-ray luminosity to determine the normalization, one can derive the central gas density and, with the surface brightness profile, determine the gas density distribution.

Mass distribution

Independent of any model, one can use the X-ray observables, gas density and temperature, to derive the *total gravitating mass* of a cluster. Thus, the elusive dark matter that dominates the cluster gravitational potential can be mapped from X-ray observations.

We start with the equation of hydrostatic equilibrium

$$\frac{dP_{gas}}{dr} = \frac{-\rho_{gas}GM_{grav}(r)}{r^2} \quad (2.13)$$

and the ideal gas law $P_{gas} = nkT = \rho kT / \mu m_p$. $M_{grav}(r)$ can be determined from gas properties alone:

$$\frac{k}{\mu m_p} \left(T \frac{d\rho}{dr} + \rho \frac{dT}{dr} \right) = \frac{-\rho GM_{grav}(r)}{r^2} \quad (2.14)$$

Solving for $M_{grav}(r)$

$$M_{grav}(r) = \frac{-kT}{G\mu m_p} \left(\frac{r}{\rho} \frac{d\rho}{dr} + \frac{r}{T} \frac{dT}{dr} \right) r \quad (2.15)$$

$$M_{grav}(r) = \frac{-kT}{G\mu m_p} \left(\frac{d \ln \rho}{d \ln r} + \frac{d \ln T}{d \ln r} \right) r \quad (2.16)$$

Thus, by measuring the gas temperature T , and the slope of the gas density distribution (or that of the surface brightness distribution), the radial distribution of the total gravitating mass can be found.

Applications to galaxy clusters

Fig.2.7 shows the basic observables as derived from a Chandra observation of the regular cluster A478 (Sun et al. 2003). From the basic observations, the three dimensional gas temperature, gas density, and gas pressure distributions can be derived. In addition, as already mentioned, with sufficiently long observations, the heavy element abundance distributions can be measured by studies of the energy spectrum (e.g., see Fig.2.2).

A478. The simple observations of gas and total mass can have profound implications. For example, since the mass components of clusters are believed to be representative of those in the Universe, the observed baryon fraction can be used to perform cosmological tests. A study of the luminous distant ($z = 0.451$) cluster RXJ1347.5-1145 is illustrated in Fig. 2.8 which shows the derivation of the gas and total mass and from these the “universal” baryon fraction (Allen et al. 2002). Since the baryon fraction depends on cosmological parameters, by measuring cluster parameters at a variety of redshifts, very strong constraints can be derived. Similarly, by studying a well defined sample of clusters, the evolution of the baryon mass function, derived through the X-ray observations, can constrain cosmological parameters (e.g., Vikhlinin et al. 2003).

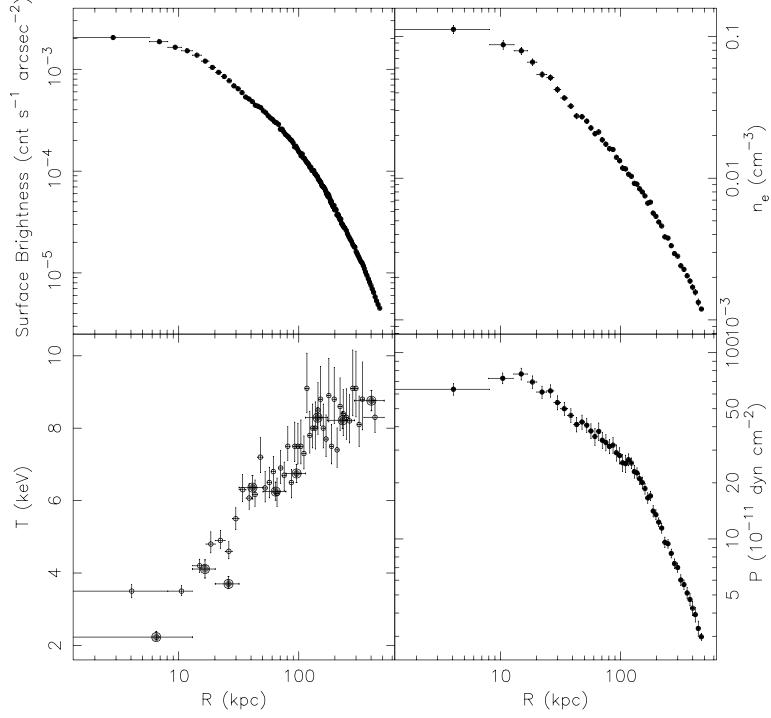


Figure 2.7. The X-ray observables surface brightness profile and projected gas temperature and their simple derivatives, deprojected temperature, gas density, and gas pressure for the regular galaxy A478 (Sun et al. 2003). (top left) Surface brightness (0.5 - 5 keV) distribution, (upper right) electron density, (lower left) gas temperature, (lower right) gas pressure. Only statistical uncertainties are shown (1σ random errors). In the temperature profile, the data points with small circles are projected temperatures, while the data points with large filled circles are deprojected (three dimensional) values.

3. Physics of Cluster Cores

As mentioned above, many clusters have surface brightness distributions that are regular with strong peaks on a bright, central, often cD, galaxy. In the Einstein X-ray survey of 215 clusters, 64% of clusters contained a bright X-ray peak, centered on an optically bright galaxy (Jones & Forman 1999). These systems, described as **X-ray Dominant** (XD) clusters, are those that have high central gas densities and hence short cooling times. In the absence of energy input, the most remarkable XD systems are calculated to be depositing mass at rates as high as $1000 M_\odot \text{ yr}^{-1}$ (e.g., Edge et al. 1994). For many years, these centrally, peaked, XD clusters have been described with a “cooling flow” model (Fabian & Nulsen 1977, Cowie & Binney 1977). However, observations with XMM-Newton have shown that mass deposition rates in cooling

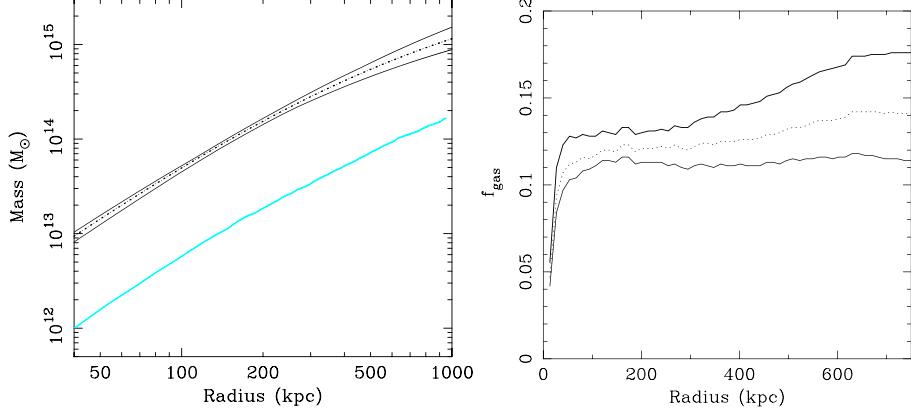


Figure 2.8. (left) The gas (lower curve) and total mass (upper curve) profiles (with 1σ range) for RXJ1347.5 (see Allen et al. 2002). gas mass profile. (right) The gas mass fraction as a function of radius derived as the ratio of the X-ray measured gas and total masses.

flow clusters are at least five times smaller than expected in the standard model (Peterson et al. 2003 and references therein). This requires considerable energy input to compensate for radiative losses.

The phenomenon of “cooling flows”, described for luminous clusters, also applies to other systems with gaseous atmospheres. In particular, galaxy groups, the poorer cousins of rich clusters, (e.g., Helsdon et al. 2001) and individual early type galaxies with their luminous, X-ray coronae (Forman, Jones & Tucker 1985) also frequently have short gas cooling times at their centers (Thomas et al. 1986).

To compensate for the radiative losses, a variety of mechanisms have been proposed. We discuss cluster mergers, thermal conduction (Tucker & Rosner 1983, Bertschinger & Meiksin 1986, Gaetz 1989, David, Hughes & Tucker 1992, Zakamska & Narayan 2003), and energy input from AGN in the central galaxy (Rosner & Tucker 1989, Böhringer & Morfill 1989, Binney & Tabor 1995, Tucker & David 1997, Churazov et al. 2001, Churazov et al. 2002).

Cluster mergers are energetic and can provide large amounts of energy. A major merger can yield up to 10^{63} ergs, thus making cluster mergers, the most energetic events in the Universe since the Big Bang. Unfortunately, cluster mergers are sporadic and are not a steady source of energy (see Fig. 2.5).

Thermal conduction was previously discounted as an energy source since many believed it was easy to suppress. However, new work by Narayan & Medvedev (2001) showed that for a sufficiently chaotic magnetic field, the conductivity can be as high as 20% of the Spitzer value. Applied to a sample of ten clusters, Zakamska & Narayan (2003) found that such thermal conduction could explain the observed gas density and temperature profiles for only

five of a sample of ten clusters. For the remaining five clusters, thermal conduction, even at the full Spitzer value, was insufficient to replace the energy lost from radiation. Thus, while conduction may be important in some clusters or in parts of clusters, it cannot be the entire solution to the replenishment of radiated energy in “cooling flow” clusters.

Energy Input from Active Nuclei

One of the most promising energy sources for replenishing the energy radiated by hot gas in clusters, groups, and individual early type galaxies is energy input from supermassive black holes that lie in galaxy nuclei. One early clue that AGN were an important source of energy was the high frequency of radio activity in the central cluster galaxy of XD clusters (70%) compared to that in galaxies with non-peaked X-ray emission (20%) (Burns 1990).

Using radio studies, Owen, Eilek & Kassim (2000; see also Binney 1999) pioneered the view that the mechanical power produced by the supermassive black hole at the center of M87 was more than sufficient to compensate for the energy radiated in X-rays. Tabor & Binney (1993) and Binney & Tabor (1995) developed models without mass deposition and included energy injection from the central AGN. Heinz, Reynolds & Begelman (1998; see also Reynolds, Heinz & Begelman 2001) modelled shock heating of the IGM by an expanding radio source. Churazov et al. (2001; see also Kaiser & Binney 2003, Bruggen 2003, De Young 2003, Kaiser 2003) argued that the morphology of the X-ray and radio observations could be explained by radio emitting plasma bubbles buoyantly rising through the hot X-ray emitting gas. These buoyant bubbles could uplift the coolest gas and provide energy input as bubble enthalpy is converted to kinetic energy, then thermalized into the gas in the bubble wake.

With Chandra, more direct evidence for the importance of AGN energy input to hot gas in early type galaxies, groups, and clusters was manifest in the frequent detection of plasma cavities in the hot gas. For NGC1275, the active galaxy at the center of the Perseus cluster, the cavities evacuated of hot gas by the radio emitting relativistic plasma had been detected with ROSAT (Böhringer et al. 1993). With Chandra, similar cavities were detected around galaxies (e.g., M84, Finoguenov & Jones 2001), in groups (e.g., HCG62, Vrtilek et al. 2002), and in other rich clusters (e.g., McNamara et al. 2000, Fabian et al. 2002, Blanton et al. 2003).

Since the cavities were surrounded by bright rims of cool gas, alternatives to shock heating models were proposed. Churazov et al. (2001; see also Kaiser & Binney 2003, Bruggen 2003, De Young 2003, Kaiser 2003) used M87 to study the effects of buoyant bubbles on the surrounding gas. They argued that buoyant bubbles could uplift the coolest gas and provide energy input as bubble enthalpy is converted to kinetic energy, then thermalized into the gas in the bubble wake.

Churazov et al. (2002) developed a model to explain the self-regulation of the accretion process around supermassive black holes. In this simple model, the supermassive black hole lies at the bottom of the gravitational potential surrounded by the lowest entropy gas. When cooling dominates, the gas entropy decreases and accretion increases. The higher accretion rate increases the energy output of the central black hole and the energy input to the gas increases, its entropy decreases, and the accretion declines.

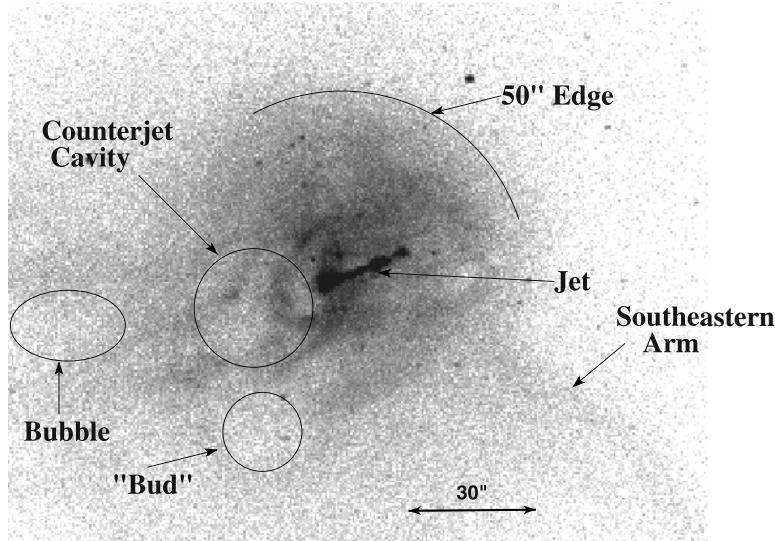


Figure 2.9. The central region of M87 as seen by the Chandra ACIS-S detector in the energy band 0.5 to 2.5 keV. The core region shows the jet, the counterjet cavity, a “bud” emanating from the southeastern edge of the counterjet cavity, and the beginning of the southwestern arm.

M87, the dominant central galaxy in the nearby Virgo cluster is a shows the detailed interaction between the X-ray emitting gas and the central supermassive black hole. The Chandra images also show cavities and filaments in the eastern and southwestern X-ray arms (see Young et al. 2002 and Forman et al. 2004).

The M87 core shows remarkable structures (Fig. 2.9) including the counterjet cavity and a bubble (labeled “bud” in Fig. 2.9) erupting from the counterjet cavity, corresponding to the southeastern extension of the bright core in the 6 cm radio observations of Hines et al. (1989).

On larger scales, Fig. 2.10 shows a remarkable variety of structure including:

- a nearly azimuthally symmetric ring of emission with a leading edge at a radius of 14 kpc ($3'$) most prominent to the north and northwest

- a second partial ring of enhanced emission, just beyond the 14 kpc ring, at a radius of 17 kpc
- the prominent eastern and southwestern arms. These arms are prominently seen as cooler gas in the XMM-Newton temperature map
- at radii beyond the 14 kpc ring, the division of each arm into two filaments. For the southwestern arm, the filaments (labeled **S1** and **S2** in Fig. 2.10b) turn east while for the eastern arm, the two filaments (labeled **E1** and **E2** in Fig. 2.10b) turn north.
- a southern arc, at a radius of approximately 37 kpc ($\sim 8'$) (also seen in the ROSAT HRI and XMM-Newton images in Forman et al. 2003)

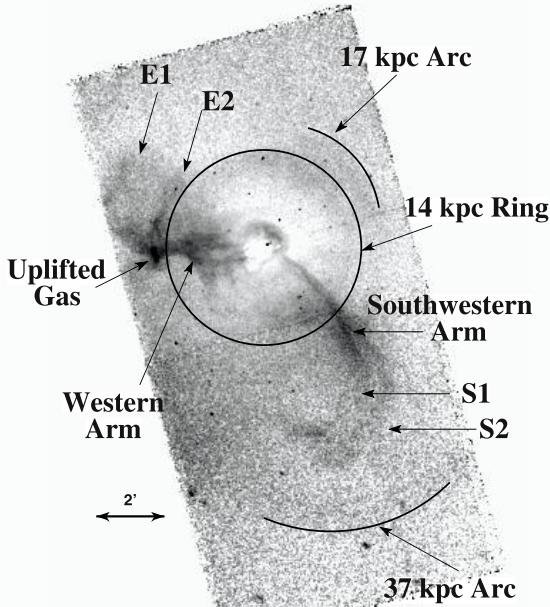


Figure 2.10 The flat-fielded Chandra image in the energy band 0.5-2.5 keV with a King model subtracted to remove the large scale radial surface brightness gradient. Many faint features are seen including 1) the prominent eastern and southwestern arms and their bifurcation (**E1**, **E2** and **S1**, **S2** identify the extensions of the eastern and southwestern arms), 2) the 14 kpc ($3'$) ring, 3) the 17 kpc ($3.75'$) arc, and 4) the faint southern 37 kpc ($8'$) arc.

The 14 kpc ring is the clearest example of a shock-driven feature in M87 (Fig. 2.10). A simple model was used to calculate the energy in an initial outburst that gave rise to the 14 kpc ring. Assuming a single outburst into a hydrostatic, isothermal atmosphere described by a power-law gas density profile which matches that observed, we found that an energy deposit of 8×10^{57} ergs about 10^7 yrs ago can produce both the observed gas density and gas temperature profiles.

A second partial ring is seen to the west at a radius of 17 kpc ($3.75'$) extending over $\sim 60^\circ$ in azimuth (see Fig. 2.10b). To form this surface brightness enhancement, a disturbance traveling at the sound speed would have originated

approximately 4×10^6 years before the event that created the 14 kpc ring. The amplitude of the 17 kpc ring is comparable to that of the 14 kpc ring and therefore would require a similar amount of injected energy.

With the detailed X-ray studies we are beginning to see the solution to the “cooling flow” puzzle and how energy from the central black hole interacts with the surrounding atmosphere of hot gas.

Acknowledgments

This work was supported by the Smithsonian Astrophysical Observatory.

References

- Allen, S., Schmidt, R., and Fabian, A. 2002, MNRAS, 335, L7
- Bertschinger, E. & Meiksin, A. 1986, ApJL, 306, L1
- Binney, J. & Tabor, G. 1995, MNRAS, 276, 663
- Blanton, E., Sarazin, C., McNamara, B. 2003, ApJ, 585, 227
- Böhringer, H. & Morfill, G. 1988, ApJ, 330, 609
- Böhringer, H., Voges, W., Fabian, A., Edge, A. & Neumann, D. 1993, MNRAS, 264, L25
- Bruggen, M. 2003, ApJ, 592, 839
- Burns, J. 1990, AJ, 99, 14
- Cavaliere, A. & Fusco-Femiano, R. 1976, A&A, 49, 137
- Cavaliere, A., Menci, N., Tozzi, P. 1999, MNRAS, 308, 599
- Churazov, E., Bruggen, M., Kaiser, C., Böhringer, H. & Forman, W. 2001, ApJ, 554, 261
- Churazov, E., Sunyaev, R., Forman & Böhringer, H. 2002, MNRAS, 332, 729
- Clowe, D., Gonzalez, A., and Markevitch, M. 2003, ApJ, astro-ph/0312273
- Cowie, L. & Binney, J. 1977, ApJ, 215, 723
- David, L., Hughes, J. & Tucker, W. 1992, ApJ, 394, 452
- De Young, D. 2003, MNRAS, 343, 719
- Edge, A. et al. 1994, MNRAS, 270, L1
- Fabian, A. & Nulsen, P. 1977, MNRAS, 180, 479
- Finoguenov, A. & Jones, C. 2001, ApJL, 547, L107
- Forman, W., Jones, C. and Tucker, W. 1985, ApJ, 293, 102
- Forman, W. et al. 2004, submitted to ApJ, astro-ph/0312576
- Gaetz, T. 1989, ApJ, 345, 666
- Geller, M. and Beers, T. 1983, PASP, 94, 421
- Helsdon, S., Ponman, T., O’Sullivan, E. & Forbes, D. 2001, MNRAS, 325, 693
- Hines, D., Owen, F. & Eilek, J. 1989, ApJ, 347, 713
- Jones, C. and Forman, W. 1999, ApJ, 511, 65
- Kaiser, C. 2003 MNRAS 343 1319
- Kaiser, C. & Binney, J. 2003, MNRAS, 338, 839

- Kraft, R. et al. 2003, ApJ, 592, 129
 Markevitch, M. et al. 2000, ApJ, 541, 542
 Markevitch, M. et al. 2002, ApJ, 567, L27
 Markevitch et al. 2003a, submitted to ApJ, astro-ph/0309303
 Markevitch et al. 2003b, ApJL, 586, L13
 McNamara, B. et al. 2000, ApJL, 534, L135
 Mewe, R., Gronenschild, E., and van den Oord, G. 1985, A&AS, 62, 197
 Mohr, J., Evrard, A., Fabricant, D. and Geller, M. 1995, ApJ, 447, 8
 Owen, F., Eilek, J. & Kassim, N. 2000, ApJ, 543, 611
 Peterson, J. et al. 2003, Proceedings of “The Riddle of Cooling Flows in Clusters of Galaxies” eds. Reiprich, Kempner, Soker, astro-ph/0310008
 Raymond, J. and Smith, B. 1976, ApJ, 204, 290
 Raymond, J. and Smith, B. 1977, ApJS, 35, 419
 Rosner, R. & Tucker, W. 1989, ApJ, 338, 761
 Spergel, D. and Steinhart, P. 2000, Phys. Rev. Lett., 84, 3760
 Stanford, S. et al. 2001, ApJ, 552, 504
 Sun, M. Jones, C., Murray, S., Allen, S., Fabian, A., Edge, A. 2003, ApJ, 587, 619
 Thomas, P., Fabian, A., Arnaud, K., Forman, W. & Jones, C. 1986, MNRAS, 222, 655
 Tucker, W., Tananbaum, H., and Remillard, R. 1995, ApJ, 444, 532
 Tucker, W. & Rosner, R. 1983, ApJ, 267, 547
 Tucker, W. & David, L. 1997, ApJ, 484, 602
 Vikhlinin, A., Markevitch, M., and Murray, S. 2001a, ApJ, 551, 160
 Vikhlinin, A., Markevitch, M., and Murray, S. 2001b, ApJL, 549, L47
 Vikhlinin, A. et al. 2003, ApJ, 590, 15
 Vrtilek, J., Grego, L., David, L., Ponman, T., Forman, W., Jones, C. & Harris, D. 2002, APRB17107
 White, S., Navarro, F., Frenk, C. and Evrard, A. 1993, Nature, 366, 429
 Young, A., Wilson, A., Mundell, C. 2002, ApJ, 579, 560
 Zakamska, N. & Narayan, R. 2003, ApJ, 582, 162

Chapter 3

CLUSTERS: AN OPTICAL POINT OF VIEW

Christophe Adami

LAM, Traverse du Siphon, 13012 Marseille, France

Abstract

Clusters are useful cosmological tools which allow to understand how large scale structures and galaxies evolve in the Universe. I will review some of the methods based on optical observations of clusters of galaxies to derive cosmological properties, describing methods and limitations to find clusters and to study these structures depending on cluster characteristics.

Introduction

Clusters of galaxies are very useful cosmological tools. Detection of complete samples at redshifts greater than 1 allow for example combined cosmological constraints on both Ω_m and Ω_Λ . These constraints are independent from usual constraints using Supernovae or the microwave background. Clusters are also ideal cosmological laboratories to study structure evolution from galaxies to large scale structures. The key-question is therefore to detect cluster samples as complete as possible and to obtain the required data to be able to study the clusters themselves.

At optical wavelengths, detection of clusters is efficient at low and moderate redshifts (roughly speaking, up to $z=0.5$). At higher redshifts, however, background and foreground galaxies begin to erase virialized overdensities on the sky and such detections at optical wavelengths require complex treatments (e.g. Postman et al. 1996).

In X-rays, detections of such distant structures are easier because they appear as extended sources and because the field is not providing such extended sources at these wavelengths in general. The contamination by field X-rays point sources is, however, sometimes very strong. Another problem is to measure the precise redshift of these extended sources at the same time. In any case, optical observations are needed to complete X-ray observations in order to confirm the nature of the source and to estimate its redshift.

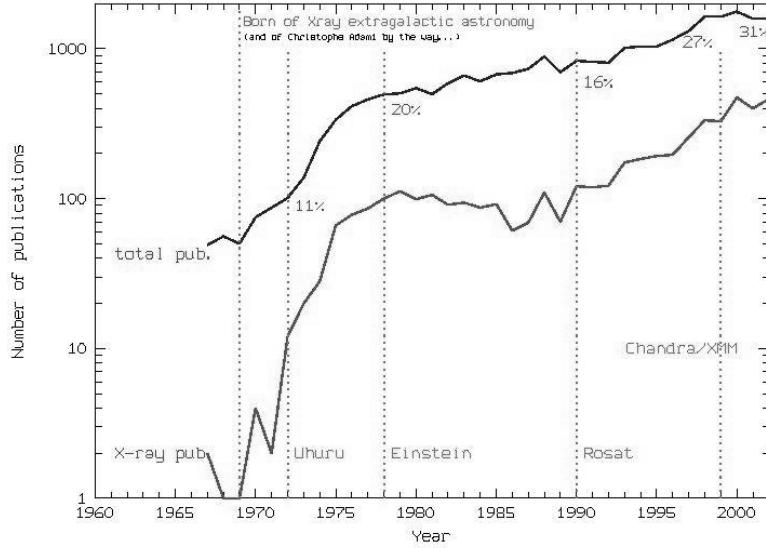


Figure 3.1. Variation of the number of all and X-ray cluster-related publications as a function of year. Percentages are the proportion of X-ray cluster related publications compared to the total number.

The ideal solution is therefore to combine X-ray and optical surveys as they provide complementary approaches (see for example Adami et al. 2000a or Donahue et al. 2001). Let us note that future SZ cluster surveys will similarly have to be completed with optical data. The need for such a combined approach has been understood since the 80s. This is illustrated by Fig. 1 that shows the parallel evolution of cluster publications using X-rays or optical data since 1980 (with the arrival of modern X-ray satellites as Einstein, Rosat, Chandra and XMM-Newton).

I will now present a (incomplete) summary of what can be done with clusters from an optical point of view.

1. Cluster detections in the optical Adaptive filters

Besides the nearby cluster detections ($z \sim 0.1$) that are using simple galaxy density criteria, the combination of such galaxy densities over the sky and of the shape of the galaxy luminosity function allows cluster detections up to $z \sim 1$

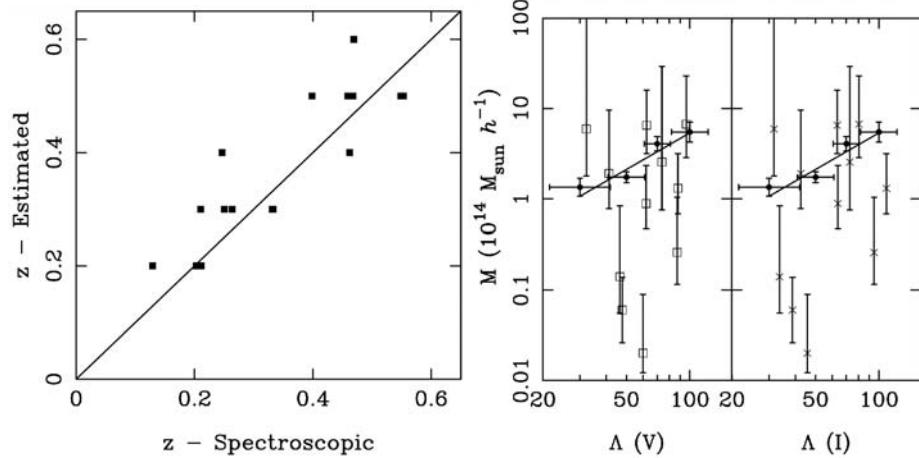


Figure 3.2. Left: estimated versus measured redshifts of the COP subsample (Holden et al. 2001) of the PDCS clusters (Postman et al. 1996). Right: measured mass versus estimated richness from Girardi et al. (1998).

(e.g. PDCS, Postman et al. 1996). The same methods also allow an estimate of the redshift and of the richness of these structures. However, while the redshift estimate is most of the time robust, the richness is sometimes overestimated (e.g. Holden et al. 2001 or Girardi et al. 1998 and Fig. 2). This is a problem since the mass of the clusters is a direct ingredient to constrain Ω_m and/or Ω_Λ using cluster counts with redshift (e.g. Oukbir & Blanchard 1992).

Red sequence

Clusters are (at least partially) virialized structures. These quite old structures are therefore partially populated with old galaxy populations (elliptical and lenticular galaxies). These galaxies trace a well defined sequence in a color/magnitude diagram of cluster galaxies. This is called the “Red sequence” (e.g. Fig. 3 for the Coma cluster: Savine 2002 and for the CIG J0848+4453 $z=1.27$ cluster: Van Dokkum et al. 2001). Systematic searches for such sequences in blind galaxy surveys allows to detect old large scale structures (assimilated most of the time with what we call a “cluster of galaxies”) (e.g. Andreon et al. 2003).

However, this method is mainly efficient at low and moderate redshifts (z lower than 1). At high redshifts, foreground contaminations hide possible red sequences by adding field galaxies homogeneously populating any color/magnitude diagram. Another problem of this method is that at any redshift you will lose any young structure or group (as younger structures will

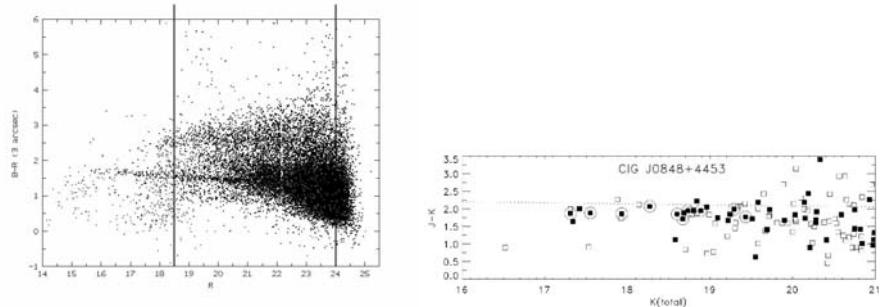


Figure 3.3. Left: color magnitude diagram for the Coma cluster (Savine 2002). The red sequence is well visible with its horizontal pattern. Right: color magnitude diagram for the CIG J0848+4453 $z=1.27$ cluster (Van Dokkum et al. 2001).

have a younger galaxy population and therefore a less well defined red sequence).

Blind spectroscopic surveys

The best way to detect compact structures along the line of sight would be to measure redshifts for all galaxies in the region of interest. This method is, in principle, independent of the distance and characteristics of the structures. For example, the new generation of Multi Object Spectrographs (MOS) mounted on 8m telescopes (e.g. VIMOS, GMOS, DEIMOS) allows to measure simultaneously several hundreds of spectra in a single exposure down to $I \sim 24$ (e.g. Le Fèvre et al. 2003). These measures analyzed for example with friend-of-friend algorithms (e.g. Rizzo et al. 2004) allow cluster detections up to $z \sim 1.2$. The redshift limit depends only on the magnitude limit and on the success rate when building the redshift catalog.

However, in practice, even with these very efficient MOS, it is still impossible to cover very large areas on the sky. We can assume a limit of a few 10deg^2 (e.g. Le Fèvre et al. 2003). The number of clusters is then limited to a few tens up to $z \sim 1$. Moreover, lowering the magnitude limit or sampling more sparsely the galaxies will decrease very rapidly the detection efficiency (when detecting only the brighter galaxies of the clusters)(see Fig.4).

Detections using photometric redshifts

The main limitation of the previous method is the time needed to detect (and to treat) optical spectroscopic data. A possible solution to this problem is to use photometric redshifts. Using five bands observed in photometry only (for example from the B to the J band) allows robust galaxy redshift determinations

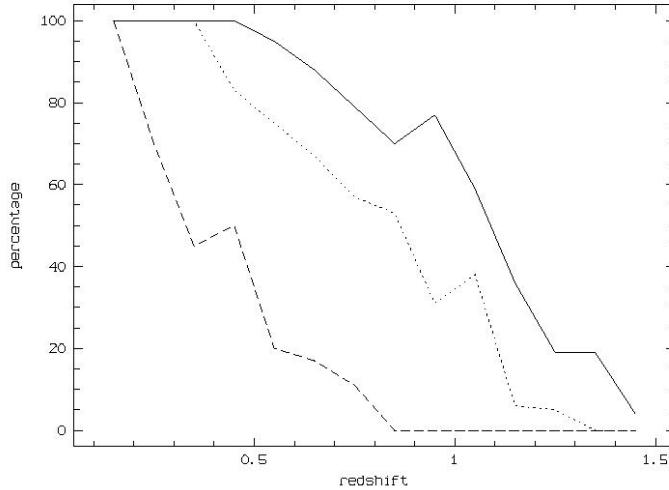


Figure 3.4. Cluster detection efficiency for blind redshift surveys down to $I=24$. Solid line is for a 100% complete survey, dotted line is for a 50% complete survey. Dashed line is for a 33% complete survey down to $I=22.5$.

from $z \sim 0.2$ to very high z (the upper limit comes from the depth of the photometric catalog and not from the photometric redshift method as long as we have near infrared photometric data)(e.g. Bolzonella et al. 2002 and Fig. 5).

The uncertainty on the redshift estimate using this method is, however, typically 0.1. With such uncertainties, friend-of-friend algorithms become inefficient. The best method to detect clusters is to use photometric redshifts to cut the galaxy samples in several redshift slices and, then, to use classical galaxy density maps in order to detect structures. This method simply increases the contrast between the structure and the field galaxies.

This method requires, however, very precise photometric measurements as photometric redshift estimates are quite versatile. Moreover, the redshift estimates depend on the galaxy types, ages, on the environmental effects acting in clusters and on the available photometric bands. For example, near infrared data are necessary to compute efficiently redshifts at z greater than 1 and wide field near infrared imagers (e.g. the future Wircam at CFHT) are not yet available.

2. Studies of clusters

Dynamical aspects

Cluster evolution. Most of the time, we need to know what is the virialization state of the clusters we observe. This is useful to compute the mass

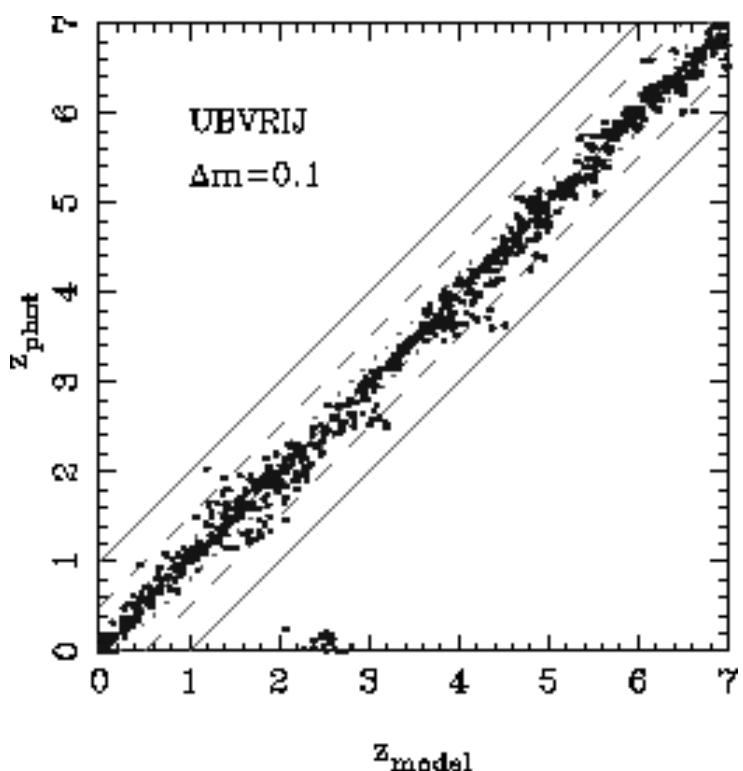


Figure 3.5. Spectroscopic redshift versus photometric redshift for UBVRJ data.

of the cluster and also to understand how this cluster is actually forming and evolving. Measuring spectroscopic redshifts allows, beside the knowledge of the cluster redshift, to have access to such informations.

We can draw a very simplified picture of a $z \sim 0$ cluster in the optical (e.g. Sarazin 1986). If a cluster is virialized, on the one hand bright and/or red cluster galaxies (mainly early types) will have small velocity dispersions and will be central objects with circular orbits. On the other hand, faint and/or blue cluster galaxies (mainly late types) will be external objects with radial orbits in the cluster. This is verified for nearby clusters (e.g. ENACS survey: Adami et al. 1998a)(Fig. 6).

Using now more distant structures (e.g. Adami et al. 2000a), we show that such structures are generally dynamically younger at z greater than 0.5 (Fig. 7). Behaviour of cluster galaxies is typically similar from $z=0$ to $z=0.4$. For higher redshifts, the simplified scenario described earlier does not apply completely: all galaxy types have higher velocity dispersions, and only very bright galaxies exhibit low velocity dispersions (as expected assuming that the brighter a galaxy is, the faster it virializes inside a cluster: e.g. Sarazin 1986).

M/L ratio. Using redshifts to compute a velocity dispersion, we can access to the mass of a cluster (assuming enough redshifts are available). Combining this estimate with photometry of the cluster, we can compute the Mass to Light ratio (M/L hereafter). The M/L ratio is very important as it is a way to measure the dark matter abundance in clusters. It is linked to Ω_m with:

$$\Omega_m = (M/L)_{cluster} / (\rho_c / j)$$

with ρ_c the critical density of the Universe and j the luminosity density of the Universe.

The M/L ratio also puts in evidence unexpected trends. For example, using ENACS data (Adami et al. 1998b) we can show a clear dependence of M/L regarding cluster velocity dispersion (Fig. 8). It means basically that the most massive the cluster, the higher the M/L, or in other words, the most massive the cluster, the higher the value of Ω_m .

Photometric aspects

From a cosmological point of view, a key ingredient of current models is the shape of the power spectrum. It relates directly structures to the mass function of the Universe and also to the luminosity function of these structures (e.g. Press & Schechter 1974). Measuring galaxy luminosity functions is therefore a way to estimate the shape of the power spectrum. Such luminosity functions could be computed using field galaxies but they are generally easier to draw in a cluster because nearly all galaxies are expected to be at the same redshift (in a cluster by definition).

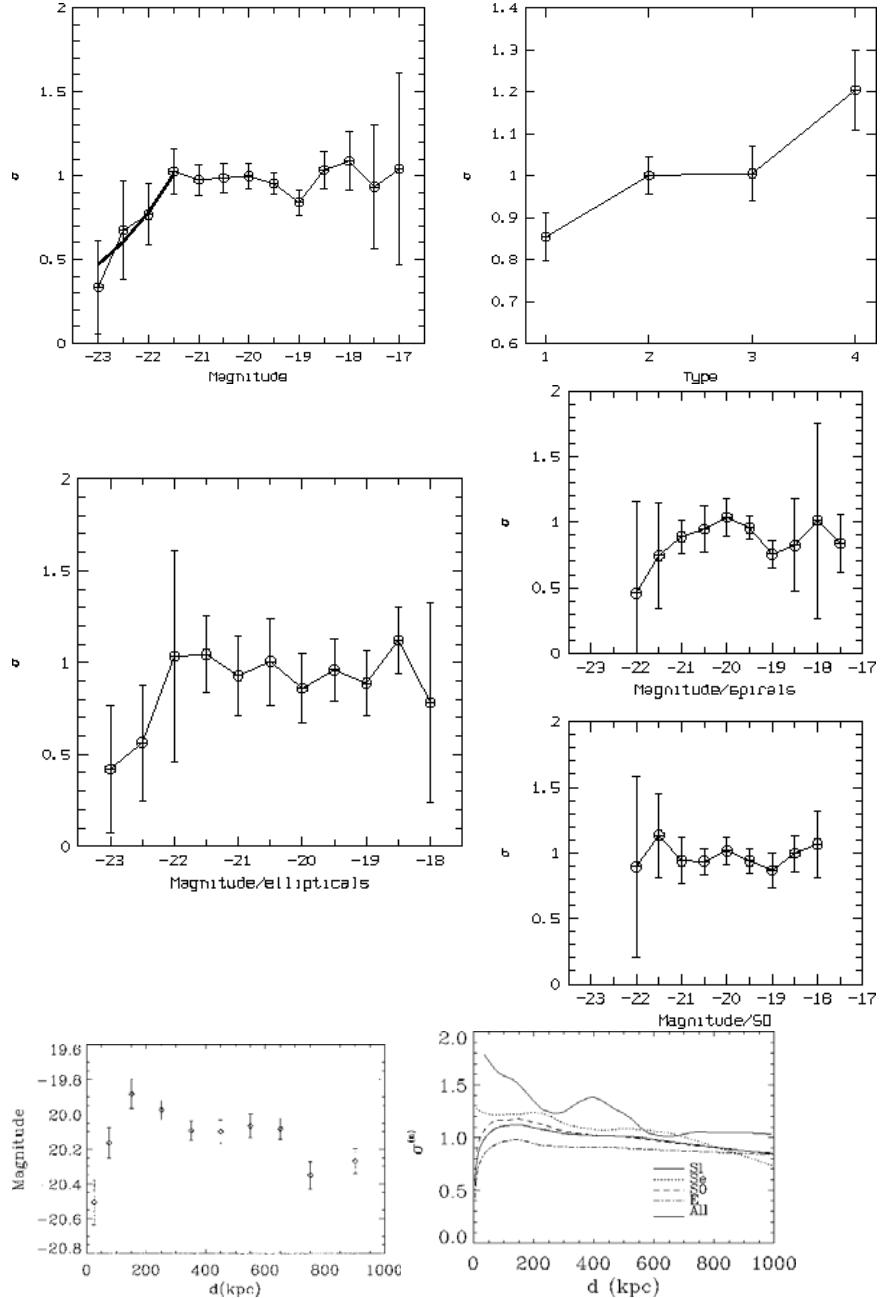


Figure 3.6. From top to bottom: (1) left: normalized velocity dispersion versus magnitude for galaxies in the nearby ENACS clusters; right: normalized velocity dispersion versus morphological type for galaxies in the nearby ENACS clusters (1: ellipticals, 2: S0, 3: early spirals, 4: late spirals). (2) normalized velocity dispersion versus magnitude for galaxies in the nearby ENACS clusters and as a function of morphological type. (3) left: mean magnitude of galaxies in the nearby ENACS clusters as a function of the cluster centric distance; right: normalized velocity dispersion versus clustercentric distance.

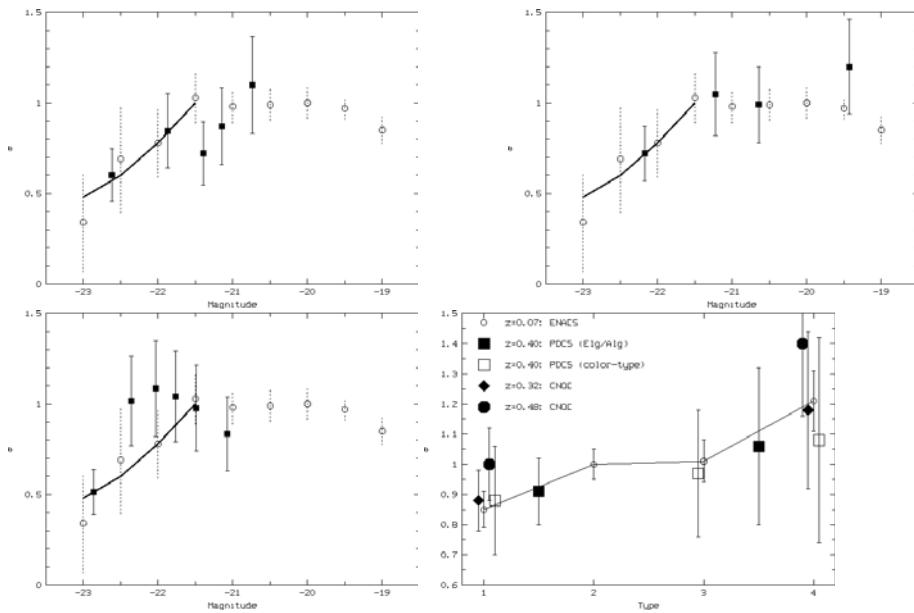


Figure 3.7. From top to bottom: (1) left: normalized velocity dispersion versus magnitude for galaxies in the nearby ENACS clusters (circles) and for $z=0.32$ cluster galaxies (1) right: normalized velocity dispersion versus magnitude for galaxies in the nearby ENACS clusters (circles) and for $z=0.4$ cluster galaxies (2) left: normalized velocity dispersion versus magnitude for galaxies in the nearby ENACS clusters (circles) and for $z=0.48$ cluster galaxies (2) right: normalized velocity dispersion versus morphological type (1: ellipticals, 2: S0, 3: early spirals, 4: late spirals) and as a function of redshift. We clearly see the departure from the mean behaviour for z greater than 0.48.

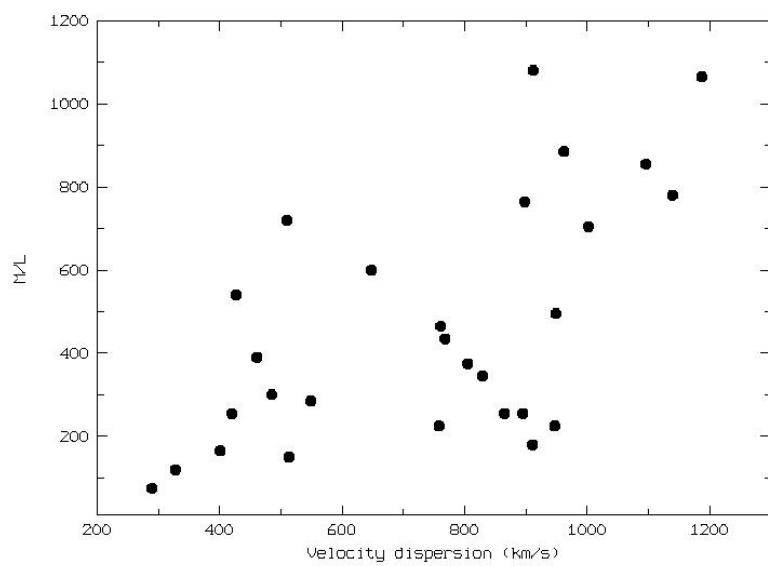


Figure 3.8. M/L ratio for a subsample of ENACS clusters versus velocity dispersion.

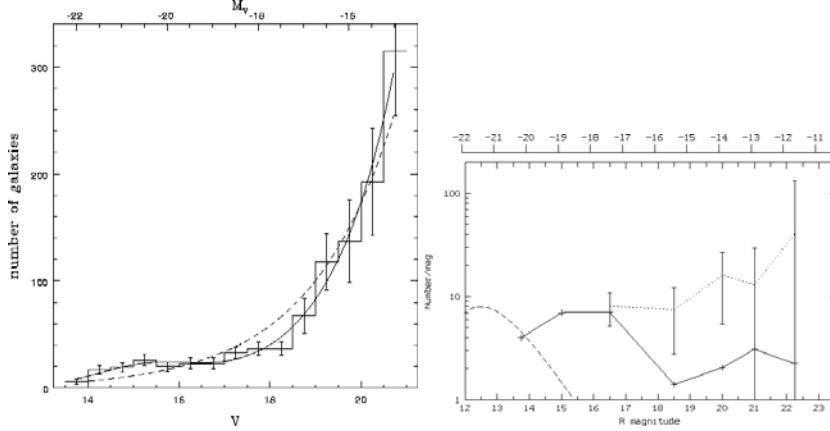


Figure 3.9. Left: (1) Bright end of the Coma GLF: we clearly see the bump at bright magnitudes and the whole just after. This is interpreted as the accretion or the merging of moderately faint galaxies in order to form brighter galaxies (Lobo et al. 1997). (2) Right: solid line: Coma inner core GLF, short dashed line: Global Coma GLF: we clearly see the cut-off with magnitude in the Coma cluster inner core.

However, a cluster is a complex place where environmental effects as galaxy merging, harassment, disruptions play a key role. It is therefore important to understand the link between cluster physic and the galaxy luminosity function (GLF hereafter). Beside the “pure” cosmological interest, it is also a very good way to understand how a structure has evolved. We can for example put in evidence the feeding of bright galaxies by fainter galaxies (e.g. Lobo et al. 1997, Fig 10), the excess of faint galaxies in clusters compared to the field (Durret et al. 2002) or the destruction of faint galaxies in the cluster center (e.g. Adami et al. 2000b and Fig. 9).

A simple way to compute a cluster GLF is first to count galaxies per unit of magnitude. After that, we have to remove foreground and background galaxies. For nearby clusters and for the bright part of the GLF, the situation is quite favourable. Only shallow galaxy surveys are needed, providing a small number of background galaxies. We can remove statistically these galaxies by considering field counts in a nearby area, or using photometric or spectroscopic redshifts. The case is, however, much more difficult for distant clusters or when computing the faint end of the GLF at any redshift. The background is very high and the objects are faint. The number of background/foreground objects is therefore higher and it is more difficult to measure spectroscopic redshifts for all these galaxies. We cannot use easily field counts to remove statistically the background/foreground galaxies because the cluster counts are completely dominated by the field counts and even a small uncertainty (cosmic variance for example) on field counts could change completely the results. The last pos-

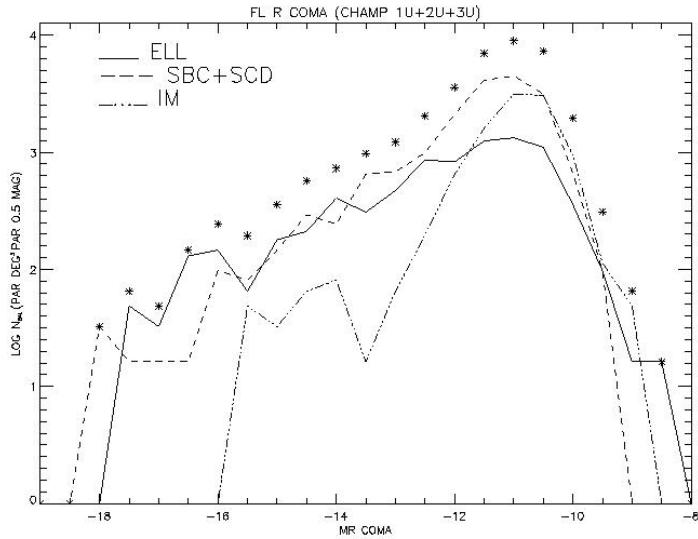


Figure 3.10. crosses: global Coma GLF from Savine 2003, solid, short dashed and long dashed lines: Coma GLF for various spectro-morphological types.

sibility is to use photometric redshifts. These are more difficult to compute because galaxies are fainter, but this method is probably the most efficient, as long as enough photometric bands are available. Moreover, this method provides, besides the redshift, an estimate of the galaxy spectro-morphological type. This is not exactly the morphological type, but it could be used as a quite reliable estimate. We show for example preliminary results for the Coma cluster (Savine 2003 and Fig. 10) showing that elliptical galaxies contribute mainly at bright magnitudes while late type galaxies contribute mainly at faint magnitudes.

3. Acknowledgements

C. Adami thanks F. Durret for useful comments and the organizers for this nice school in a nice place at a nice period and with nice food!

References

- Adami C., Holden B., Castander F., et al., 2000a, AJ 120, 1
 Adami C., Ulmer M., Durret F., et al., 2000b, A&A 353, 930

- Adami C., Biviano A., Mazure A., 1998a, A&A 331, 439
Adami C., Mazure A., Katgert P., Biviano A., 1998b, A&A 336, 63
Andreon S., 2003, A&A 409, 37
Bolzonella M., Pello R., Maccagni D., 2002, A&A 395, 443
van Dokkum P., Stanford S., Holden B., et al., 2001, ApJ 552, L101
Donahue M., Mack J., Scharf c., et al., 2001, ApJ 552, L93
Durret F., Adami C., Lobo C., 2002, A&A 393, 439
Le Fèvre, O., Vettolani G., Maccagni D., et al., 2003, Msngr 111, 18
Girardi M., Giuricin G., Mardirossian F., et al., 1998, ApJ 505, 74
Holden B., Adami C., Nichol R., et al., 2001, AJ 120, 23
Lobo C., Biviano A., Durret F., et al., 1997, A&A 317, 385
Oukbir D., Blanchard A., 1992, A&A 262, L21
Postman M., Lubin L., Gunn J., et al., 1996, AJ 11, 615
Press W., Schechter P., 1974, ApJ 187, 425
Rizzo D., Adami C., Bardelli S., et al., 2004, A&A 413, 453
Sarazin C., 1986, Rev. Mod. Phys. 58, 1

Chapter 4

COSMOLOGY WITH CLUSTERS OF GALAXIES

Alain Blanchard

LATT, OMP, CNRS, UPS, 14, Av Ed. Belin, 31 400 Toulouse

alain.blanchard@ast.obs-mip.fr

Abstract

Clusters constitute one major source of informations for cosmology. In the present paper, I concentrate on the vision on X-ray clusters from a theorist point of view and more specifically on the way they can be used for the determination of the matter density parameter of the universe Ω_m . I describe the usual modeling of X-ray clusters for cosmological applications and the scaling relations that are used in this modeling. I then discuss several cosmological tests based on clusters properties. The first method uses the evolution of the abundance of clusters with redshift, the second one make use of the baryon fraction in clusters. The former method needs reliable estimates of the local temperature distribution function as well as at high redshift. The latter can provide a direct estimate of Ω_0 provided that we have a reliable estimate of the baryon content of clusters and a good estimate of the primordial baryonic abundance Ω_b . Finally, I will illustrate an iconoclast point of view accordingly to which no definitive conclusions have been achieved by these methods. Although low density universes are favored by most authors, I argue that present day observations can also be interpreted as providing evidence for a high density universe.

1. Introduction

The first evidence for the existence of dark matter has been provided by dynamical measures performed on the Coma cluster, in ~ 1930 by F. Zwicky. Since that time, our understanding of clusters has greatly increased. There is nearly 100 times more mass in clusters than in the stars that can be seen within them. However, there is much more baryons seen in X-ray clusters in form of hot gas than in stars. The discovery of this hot gas through its X-ray emission has revolutionized the study of clusters. Indeed X-ray observations allows to measure gas density and gas temperatures with a high accuracy and they are likely to provide the most accurate mass measurements. Clusters therefore provide a fascinating laboratory for cosmological studies: their stellar, bary-

onic, metals and dark matter contents can be accurately estimated, their spatial distribution can be measured on very large scale, and their evolution can be followed up to redshift $z \sim 1$.

2. What is a cluster?

From an observational point of view clusters were first identified as galaxy concentrations containing several hundred of galaxies. These concentrations were identified on Schmidt plates. With the advent of optical spectroscopic studies it has become clear that those concentrations were actually corresponding to concentrations in space, although from time to time they were rather spurious projections. X-ray telescopes have given a new vision on these objects. The progress on the observational properties of clusters was certainly driven by the X-ray observatories (Einstein, ROSAT, ASCA, Chandra, XMM) but also by the possibility to observe clusters at various wavelengths, bringing different type of informations, like velocity dispersion, weak lensing mass estimates, Sunyaev-Zeldovich signal and, potentially, their distribution functions. The actual correspondence between optical and X-ray clusters is generally good, although some debate remains on which of the two visions is the more relevant. Therefore there is some problem in defining a cluster and this problem is becoming more and more serious for lower mass systems, conventionally called groups. This problem is leading to some practical difficulties. For instance, some objects archived in BAX, the X-ray clusters database (<http://bax.ast.obs-mip.fr>; Sadat et al., 2004), are not identified as clusters in NED or Simbad. This difficulty reappears in numerical simulations.

During the last twenty years numerical simulations have brought considerable progresses on the question of clusters, and these progresses are certainly an important factor in their growing role in cosmology. As a result from numerical studies, it has been shown that analytical modelings were able to catch their main average properties, mainly their mass estimation and their mass function. Therefore, despite the fact that clusters are non linear objects they can be used directly to compare the expected properties in a given scenario to the growing set of observational data.

However, the same difficulty that observers meet in defining a cluster exist for theorists to define clusters in a numerical simulation: typical numerical simulations handled several millions dark matter particle and a similar number of gas particle when hydro-dynamical processes are taken into account; the actual distribution of dark matter, at least on non linear scales is very much like a fractal, for which the definition of an object is somewhat conventional! Different algorithms are commonly used to define clusters. Friend of friend is commonly used because of its simplicity, however its relevance to observations is very questionable, especially for low mass systems. On the analytical side

the description of a non-linear structure becomes very rapidly complex. For this reason the spherical model is the most commonly used.¹.

3. The spherical model

A spherical overdensity let say at time of recombination will grow under its own gravity and will eventually reach a maximum radius and then collapses back. During the actual collapse, non radial trajectories will develop, forbidding to end up to a singularity. The virial argument allows to derive that the final radius is half of the radius at maximum. This process is supposed to provide a stationary solution, a “virialized” state, when the spherical model would have collapse back to a singularity. The nonlinear contrast density Δ_v at this epoch is defined as the contrast at virialization. In an Einstein de Sitter cosmology this contrast can be evaluated by mean of the exact solution for $\Omega > 1$ solution: the virial contrast density is $18\pi^2$. At that time the amplitude evaluated from the linear regime δ_v would be ~ 1.68 . This formula can be generalized to arbitrary cosmological models, in which case they are function of the redshift of collapse z_v . Oukbir & Blanchard (1997) provided exact formula for the growing rate, δ_v and Δ_v in open universes. Bryan and Norman (1998) provided useful fitting formula using a slightly different quantity: the contrast density compared to the critical density Δ_c in flat (and open) cases. Henry (2000, 2004) provided a comprehensive set of useful formula to deal with various other cases.

4. The mass function

An important work in the history of this field is the Press and Schechter paper (1974). They derived a formula for the mass function based on the spherical model assuming Gaussian fluctuations. It is only during the eighties that the accuracy of their formula to fit results from numerical simulations was noticed (Efstathiou et al., 1988). The physical motivation for the validity of the Press and Schechter approach has been discussed widely in the literature. Here I follow the argument developed in Blanchard et al. (1992). The fundamental assumption consists in specifying that an object of mass m or greater results from the collapse of a (homogeneous) sphere of radius R or greater which satisfies the non-linear criteria. The fraction of matter in objects with mass greater than some value m can be then written

$$F_{\text{NL}}(>m) = \frac{1}{\rho} \int_m^\infty n(m) m dm = \int_{\nu_S}^\infty F_R(\nu) d\nu \quad (4.1)$$

¹To my knowledge, G.Lemaître is the first one to have developed the calculation in detail (Lemaître, 1933). He is also a pioneer in this field as he was probably the first to try to connect cluster's properties with the cosmological framework, although in retrospect his arguments were not very convincing!

where the right side represents the probability that a piece of matter in the universe is included in some sphere with radius R or greater and which satisfies the non linear criteria $\delta > \delta_v = \nu_S \sigma(m)$ where $\sigma(m)$ is the rms matter fluctuation on the scale associated to m . The mass function can then be easily derived :

$$n(m) = -\frac{\bar{\rho}}{m} \frac{dF_{\text{NL}}(>m)}{dm}, \quad (4.2)$$

which leads to:

$$n(m) = -\delta_v \frac{\bar{\rho}}{m} \frac{1}{\sigma^2} \frac{d\sigma}{dm} F\left(-\frac{\delta_v}{\sigma}\right). \quad (4.3)$$

Press and Schechter assumed that the function F was a gaussian, changing the normalization constant to get the total mass of the universe to lie in non-linear objects (there is actually no proof that this result should hold). There has been considerable efforts to improve the derivation of the function F . However, it has also been realized that in order to provide accurate fitting formula one should define precisely what a cluster is. As we have pointed out, the definition which is used for modeling when the spherical model is assumed implies that a cluster is a spherical region with a contrast density Δ_v . White (2002) extensively compares mass function using different definitions of a cluster. Other values of the density contrast are sometime used, for instance $\Delta = 500$ is sometime preferred as providing tighter relations between mass and integrated properties of clusters. For the mass function, Jenkins et al. (2001) provide a fit which is used preferentially for objects defined at a fixed contrast density independent of the redshift. Intensive numerical simulations have allowed to provide accurate analytical fits to the mass function (Sheth, Mo & Tormen 2001, hereafter SMT; Jenkins et al. 2001; White 2002). In the following I use the expression from SMT:

$$\frac{dn}{dm} = \sqrt{\frac{2a}{\pi}} c \frac{\bar{\rho}}{m} \frac{d\nu}{dm} \left(1 + \frac{1}{(a\nu^2)^p}\right) \exp\left(-\frac{a\nu^2}{2}\right) \quad (4.4)$$

with $a = 0.707$, $c = 0.3222$ and $p = 0.3$ and $\nu = \frac{\delta}{\sigma(m)}$. In this formula, clusters are defined as spherical overdensities with contrast density $\Delta_v(z, \Omega_m, \dots)$. Further more this formula was checked for various cosmological frameworks. The SMT expression is almost identical to Jenkins' fit for a concordance model. Such fitting formula are describing the abundance of "clusters" in numerical simulations to an accuracy which is better than 20%.

5. Connection to the observations

The above expression of the mass function depends on the cosmological background and on the characteristics of the fluctuations, specified by their

their distribution function, their power spectrum and their amplitude. Historically, a power-law spectrum was assumed $P(k) \propto k^n$, although now only CDM-like spectrum are used, with a characteristic shape parameter Γ (current observations from large scale structure surveys favor a value of Γ in the range 0.1–0.3). It is furthermore assumed that clusters identified in numerical simulations essentially correspond to clusters in the sky. However the mass of clusters is difficult to evaluate properly. It seems therefore preferable to infer the relation between mass and some observable quantity (typically X-ray temperature or luminosity) and to study the distribution function of clusters in term of this quantity. In the following I concentrate on the temperature distribution function. The reason is that the temperature is believed to be tightly related to the mass of the system, for instance as indicated by the hydrostatic equation. From energy conservation argument, we can expect that the average thermal energy of a particle in the hot gas is related to the average kinetic energy of a particle in the gravitational potential of the cluster. μ being the average molecular weight² of the intracluster plasma one can write :

$$\frac{3}{2}kT = \frac{1}{2}\mu m_p \sigma_{3D}^2 \quad (4.5)$$

where σ_{3D} is the r.m.s. velocity of a particle. The (1D) velocity dispersion is defined as the velocity along the line of sight (as measured by observers): $\sigma^2 = 1/3\sigma_{3D}^2$. Therefore the average temperature of the gas in case of full thermalization can be written as:

$$T = \mu m_p \sigma^2 / k \quad (4.6)$$

For a singular isothermal sphere we have:

$$\sigma^2 = \frac{GM}{2R} \quad (4.7)$$

This relation can be applied at the virial radius, being defined as the radius enclosing a density contrast Δ_v :

$$R_v = \frac{1}{1+Z} \left(\frac{M}{4\pi/3\Delta_v \Omega_m \rho_c} \right)^{1/3}$$

We can then write a direct relation between mass and velocity dispersion:

$$\sigma = 896(\Omega_M \Delta(z_v, \Omega_M)/178)^{1/6} M_{15}^{1/3} (1+z)^{1/2} \text{km/s} \quad (4.8)$$

as well as between mass and temperature:

$$T = A_{TM}(\Omega_M \Delta(z_v, \Omega_M)/178)^{1/3} M_{15}^{2/3} (1+z) \quad (4.9)$$

² μ is generally taken to be equal to 0.59, corresponding to a fully ionized plasma with 24.5% in helium.

where $\Delta(z, \Omega_M)$ is the contrast density with respect to the density of the Universe for virialized objects (hereafter $h = 0.5$ when not specified). The subscript 15 means that masses are taken in unit of 10^{15} solar masses. The constant A_{TM} can be estimated from the above argument. However, there might be several reasons why the normalization is uncertain so we write it with an additional factor f_T :

$$A_{TM} = 4.92 f_T (h/0.5)^{2/3} \text{keV} \quad (4.10)$$

Mass estimates of X-ray clusters can be also obtained from the hydrostatic equation. This technique has been widely used in the past, although its validity and its efficiency has been questioned from time to time. One fundamental limitation is due to the fact that the mass cannot be estimated without a good knowledge of the temperature profile which is generally very poorly known. Even in this case, the hydrostatic method leads to uncertainties which are larger than naively expected (Balland & Blanchard, 1997). Hughes (1997) reached the same conclusions. Furthermore, the possible existence of turbulence (Norman & Bryan, 1999) in clusters might lead to additional pressure terms which are not taken into account in standard analysis, resulting in lower f_T . Indeed values of A_{TM} derived from numerical simulations are somewhat smaller than from the hydrostatic method (Markevitch, 1998; Roussel et al., 2000). There is a moderately good convergence between different sets of numerical simulations

$$f_T = 0.75 - 1. \quad (4.11)$$

the above range in the normalization represents the extreme values which have been published by different groups, the lower one corresponds to Bryan & Norman (1998), while the higher one corresponds to the value obtained by Evrard, Mathiessen & Navarro (1996). These normalizations lead to mass estimates which are larger than those derived from the hydro-static equilibrium method³. When converted in M/L ratio, they lead to (Roussel et al., 2000):

$$M/L \sim 640 - 800h \quad (4.12)$$

In order to infer the mass density of the universe, one has to make the assumption that the ratio of dark matter to light is the same everywhere in the universe. This is far from being obvious, as clusters represents only 10^{-5} of the total volume of the universe. Evidences for the presence of a such large quantity of dark matter are probably reasonable but far from being as robust as in clusters. A dramatic possibility would be that dark matter is present in large quantity only in clusters... (the amount of dark matter directly “seen” in galaxies from rotation curves is much smaller than in clusters). However, there are a

³From the hydrostatic method typical values are : $f_T \sim 1.3$

couple of evidences that dark matter around galaxies extends up to few 100 kpc from the pair wise velocity distribution (Bartlett & Blanchard, 1996) and up to a couple of Megaparsecs from weak lensing measurements (Van Waerbeke et al., 2000). The M/L ratio is transformed in term of Ω_0 by assuming that the ratio of matter to light is universal:

$$\rho_m = M/L \times \rho_l \quad (4.13)$$

where ρ_l is the light density of the universe (this quantity however may not be so well known, being underestimated for instance if galaxies are missed in present day survey). Then:

$$\Omega_0 = M/L \frac{8\pi G \rho_l}{3H_0^2} = \frac{M/L}{M/L|_c} \quad (4.14)$$

Using the determination of the luminosity function by Zucca et al. (1997), one finds $M/L|_c \sim 1250h$, the above M/L leading to $\Omega_0 \sim 0.5 - 0.65$, higher than values based on standard optical estimates. The main uncertainty on this method is due to the possibility that the distribution of light is not a fair representation of the dark matter distribution. For this reason, other methods of determination of the density parameter of the Universe are requested. Methods which do not rely on the assumption of the fairness of the light distribution can be qualified as global methods. Such global methods are rare. Clusters provide us with such global methods for which small errors bars have been obtained. These methods rely on our ability to model adequately properties of clusters and their behavior with mass, temperature and redshift accordingly to the various cosmological models.

6. Properties of Clusters and scaling relations

X-ray properties

X-ray imaging of clusters directly provide a map of the hot gas distribution, while the spectroscopy of this gas provide its temperature. Typical temperatures are found in the range between 1 and 15 keV. Bolometric luminosities can be very different from one cluster to an other and range from 10^{43} erg/s to nearly from 10^{46} erg/s. There is a well established correlation between luminosities and temperature, although it is widely dispersed. The following relation can be used for most applications:

$$L_{44} = L_{bol}/(10^{44}\text{erg/s/cm}^2) = 0.05T_{\text{keV}}^3 \quad (4.15)$$

Recent and detailed reanalysis can be found in Arnaud and Evrard (1999) and Markevitch (1998). The X-ray bremsstrahlung emission of clusters can generally be fitted by the emission of a gas distributed accordingly to the so called

β -model:

$$\rho_g(r) = \rho_0 \frac{1}{(1 + (r/r_c)^2)^{3\beta/2}} \quad (4.16)$$

where r_c is the core radius of the X-ray gas.

Scaling relations

The scaling argument holds on the idea that different clusters are geometrically identical. This means that clusters at a given epoch can be entirely described by only one quantity, their mass to which an unique scale is associated : the virial radius R_v . We have seen already how velocity dispersion and temperature can be related to the mass in the previous section⁴ A further assumption is that the gas fraction does not vary with mass neither with redshift. For instance, self similarity implies:

$$r_c \propto R_v \quad (4.17)$$

allowing to derive the luminosity–mass relation from the bremsstrahlung emissivity (assuming the Gaunt factor to be constant) :

$$L_x \propto n^2 T^{1/2} V \propto (\Omega_M \Delta(z_v, \Omega_M)/178)^{7/6} M^{4/3} (1+Z)^{3.5} \quad (4.18)$$

One can therefore predict the expected scaling of the luminosity–temperature relation, a well observed quantity:

$$L_x \propto T^2 (\Omega_M \Delta(z_v, \Omega_M)/178)^{1/2} (1+Z)^{3/2} \quad (4.19)$$

This conflicts with the observed relation. The origin of this discrepancy is not yet entirely understood. It is reasonable to believe that additional non-gravitational physics is necessary, which results in gas heating. Important scaling laws in clusters' properties have been found however from numerical simulations in the dark matter distribution. From several numerical simulations, Navarro, Frenk & White (1995) noticed that, at least for regular clusters, the dark matter profile follows a law commonly called the NFW profile:

$$\rho(r) = \frac{\delta_0}{cx(1+cx)^2} \quad (4.20)$$

the constant c being the concentration parameter and $x = r/R_v$. Notice that there is only one free parameter, as the average contrast density at the virial radius should be Δ_v . The mass profile can be estimated analytically. Currently,

⁴Scaling arguments allow to infer that $\sigma \propto M^{1/3}(\Omega_M \Delta(z_v, \Omega_M)/178)^{1/6} \cdot ((1+Z)^{0.5}$ and $T \propto M^{2/3}(\Omega_M \Delta(z_v, \Omega_M)/178)^{1/3} (1+Z)$ but not the constant of normalization.

$c = 5$ is used, although values obtained are in the range 5 to 15. The average value of c varies with the mass of clusters and with the cosmological model, low mass objects and low density universe giving higher c . In the outer part of clusters ($0.3 < x < 1.$) varying c from 5. to 10. makes at most 25% difference in mass, while at $x \sim 0.1$ (typical of the core radius) the difference is of the order of a factor of two.

7. Clusters abundance evolution.

The evolution of the abundance of clusters relative to the present day value is a direct test of Ω_0 which can be demonstrated almost like a mathematical theorem – see Blanchard and Bartlett (1998) and is almost insensitive to a non-zero cosmological constant. As X-ray clusters can be detected at high redshifts, they provide us with a global test of Ω_0 (Oukbir and Blanchard, 1992). In principle, it is relatively easy to apply, because the change in the abundance at redshift $\sim 1.$ is more than an order of magnitude in a critical universe, while it is almost constant in a low density universe. Therefore the measurement of the temperature distribution function (TDF) even at moderate redshift $z \sim 0.5$ should provide a robust answer. In principle, this test can be applied by using other mass estimates, like velocity dispersion, Sunyaev-Zeldovich, or weak lensing. However, mass estimations based on X-ray temperatures is up to now the only method which can be applied at low and high redshift with relatively low uncertainty.

The local temperature distribution function

The estimation of the local temperature distribution function of X-ray clusters can be achieved from a sample of X-ray selected clusters for which the selection function is known and for which temperatures are available. Until recently, the standard reference sample was the Henry and Arnaud sample (1991), based on 25 clusters selected in the 2 – 10 keV band. The ROSAT satellite has since provided better quality samples of X-ray clusters, like the RASS and the BCS sample, containing several hundred of clusters. Temperature information is still lacking for most of clusters in these samples and therefore do not yet allow to estimate the TDF in practice. In order to construct a sample of X-ray clusters, with temperatures measured for all clusters, one has to select clusters with a flux above 2.210^{-11} erg/s/cm² with $|b| > 20$. This leads to a sample comprising around 50 clusters. The inferred TDF is given in figure 1. This is in very good agreement with the TDF derived from the BCS luminosity function. The abundance of clusters is slightly higher than derived from the Henry and Arnaud sample as given by Eke et al. (1998) for instance. It is in good agreement with Markevitch (1998) for clusters with $T > 4$ keV, but is slightly higher for clusters with $T \sim 3$ keV. The power spectrum of fluctua-

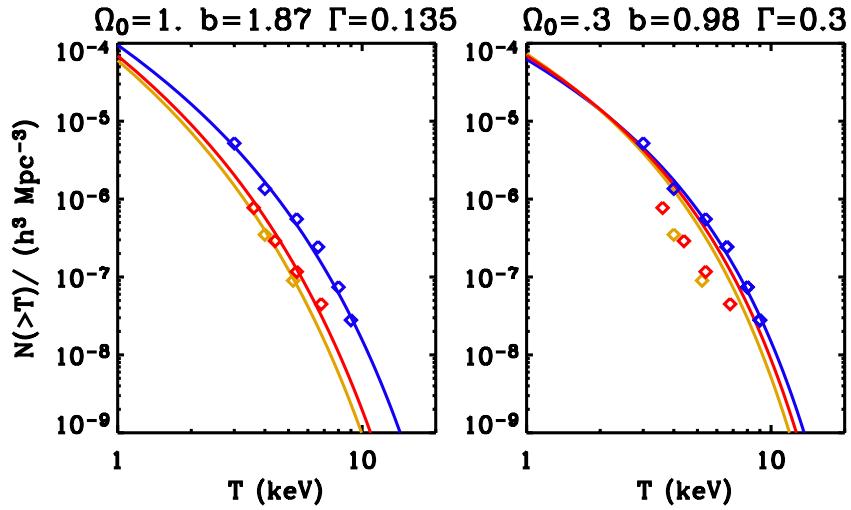


Figure 4.1. On these plots, the power of the cosmological test of the abundance of clusters is illustrated: the TDF normalized to present day abundances (dark lines) evolve much faster in a high density universe (left panel, $\Omega_0 = 1.$) than in a low density (flat) universe (right panel, $\Omega_0 = 0.3$): $z = 0.33$ (dark grey symbols and lines) the difference in the abundance of clusters is already of the order of 3 or larger. I also give an estimate of the local TDF (black symbols) as well as an estimate of the TDF at $z = 0.33$ (dark grey – symbols). Also are given for comparison data (Henry, 2000) and model at $z = 0.38$ (light grey symbols and lines). On the left panel, the best model is obtained by fitting simultaneously local clusters and clusters at $z = 0.33$ leading to a best value of Ω_0 of 1. The right panel illustrates the fact that a low density (flat) universe $\Omega_0 = 0.3$ which fits well local data does not fit the high redshift data properly at all.

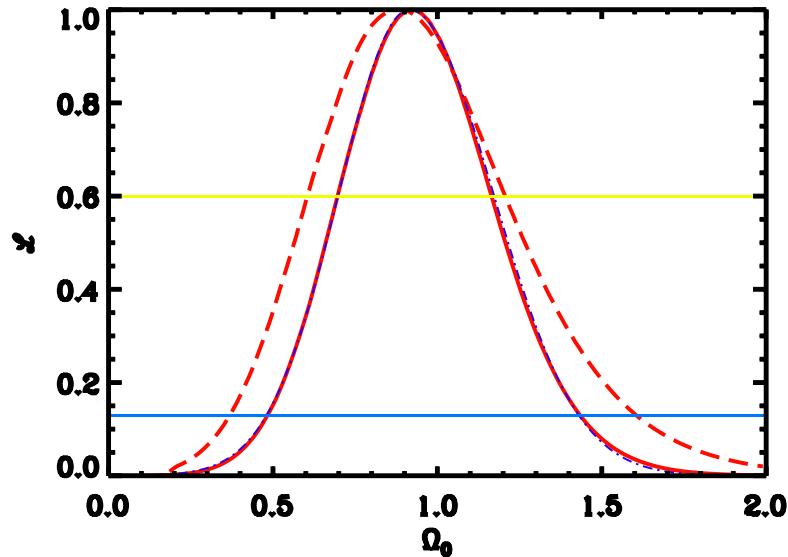


Figure 4.2. The comparison of the abundance of clusters at $z = 0.05$ with the abundance at $z = 0.33$ allows one to determine the likelihood of the mean density parameter of the universe. The continuous line corresponds to the open case, the dashed line corresponds to the flat case. In both cases a high value is preferred. The two horizontal lines allow to determine the 1 and 2 σ ranges for the parameter if the errors on the measured quantities are gaussian distributed (Blanchard et al., 2000). Revised version of this analysis with SMT mass function leads to slightly higher value (see Vauclair et al., 2003).

tions can be normalized from the abundance of clusters once the value of the constant A_{TM} is chosen. Using $A_{TM} = 6.35h^{2/3}$ leads to $\sigma_8 = \sigma_c = 0.6$ for $\Omega_0 = 1$ and to $\sigma_c = 0.7$ for $\Omega_0 = 0.3$ corresponding to $\sigma_8 = 0.96$ for a $n = -1.5$ power spectrum index (contrary to a common mistake the cluster abundance does not provide an unique normalization for σ_8 in low density models).

Application to the determination of Ω_0

The abundance of X-ray clusters at $z = 0.33$ can be determined from Henry's sample (1997) containing 9 clusters. Despite the limited number of clusters and the limited range of redshift for which the above cosmological test can be applied, interesting answer can already be obtained, demonstrating the power of this test. Comparison of the local TDF and the high redshift TDF clearly show that there is a significant evolution in the abundance of X-ray clusters (see figure 1), such an evolution is unambiguously detected because of the better

quality sample at $z \sim 0$. This evolution is consistent with the study of Donahue et al. (2000). Several groups have performed a likelihood analysis to estimate the mean density of the universe from the detected evolution between $z = 0.05$ and $z = 0.33$. The likelihood function is written in term of all the parameters entering in the problem: the power spectrum index and the amplitude of the fluctuations. The best parameters are estimated as those which maximized the likelihood function. Using ingredients as described above, Blanchard et al. (2000) obtains a high value for the preferred Ω_0 for the open and flat case, with a rather low error bars :

$$\Omega_0 = 0.92_{-0.22}^{+0.26} \quad (\text{open case}) \quad (4.21)$$

$$\Omega_0 = 0.86_{-0.25}^{+0.35} \quad (\text{flat case}) \quad (4.22)$$

the concordance model is therefore at 2σ . Interestingly, the best fitting model also reproduces the abundance of clusters (with $T \sim 6$ keV) at $z = 0.55$. The preferred spectrum is slightly different in each model: low density universe prefers $n \sim 1.7$, while high density universe prefers lower value $n \sim 1.9$, but with large uncertainties.

8. The baryon fraction

This method is based on the measurement of the baryonic fraction in clusters, consisting mainly of the hot gas seen in X-rays. The X-ray image of a cluster allows one to measure the mass of this X-ray gas. The knowledge of the X-ray temperature allows one to estimate the total mass M_t . It is possible therefore to estimate the baryon fraction in clusters (the contribution of stars, around 1% for $h = 0.5$ is often neglected to first order) *assuming that the remaining dark matter is non-baryonic*, which can be related to Ω_0 :

$$f_b = \frac{M_b}{M_t} = \Gamma \frac{\Omega_b}{\Omega_0}$$

the numerical factor Γ is introduced in order to correct for possible differences arising during cluster formation. Numerical simulations from various groups have shown that this factor is of the order of 0.90 in the outer part of clusters, although non-gravitational processes might alter this picture. Primordial nucleosynthesis allows the estimate of Ω_b , therefore the knowledge of f_b allows to infer Ω_0 . This method has been widely used since the pioneering work of White et al. (1993). Typical baryon fraction at the virial radius have been found in the range 15 to 25 % (for a Hubble constant of 50 km/Mpc/s).

Detailed studies of the baryon fraction in clusters have been conducted in recent years. There are some controversies about whether the baryon fraction varies with temperature or is constant. Roussel et al (2000) found that the baryon fraction in clusters follow a scaling law and found that the baryon

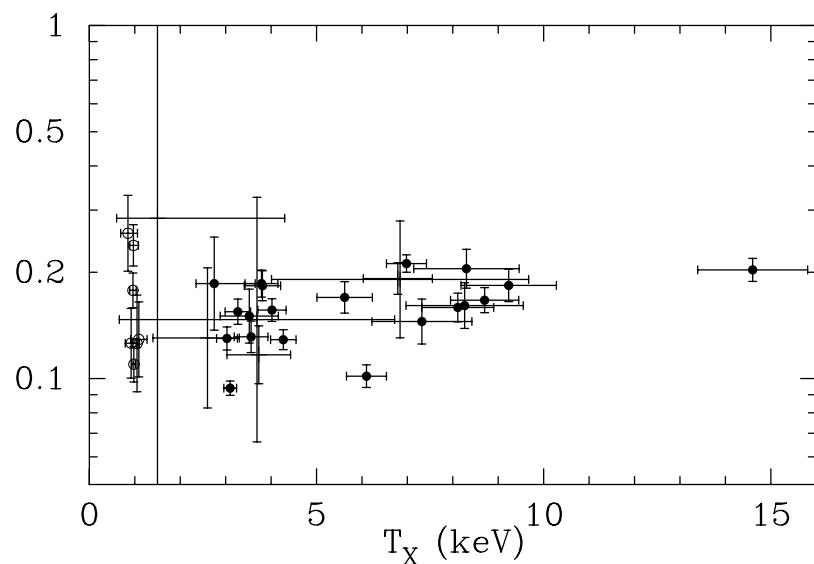


Figure 4.3. The baryon fraction at the virial radius versus temperature as derived by Roussel et al. (2000). No sign of a dependence with temperature is found.

fraction seems not to vary significantly with temperature in the range 1 to 15 keV (see Fig. 4.3). The mean baryon fraction (including stars) they obtained is typically 16% ($h = 0.5$), when the EMN mass estimator is used (higher baryons fraction can be found with the hydrostatic equation). An interesting consequence is about the baryon content of the universe which can be inferred from the same argument than for the total density of the universe.

$$\Omega_b = \frac{M_b/L}{M/L|_c} \quad (4.23)$$

Roussel et al (2000) found $M_b/L \approx 35h^{-1/2}$ leading to $\Omega_b \sim 0.027h^{-3/2}$. There has been some debate about the preferred value for nucleosynthesis, but there is now some convergence towards a high baryon content (Tytler et al., 2000), in full agreement with constraints obtained from WMAP (Spergel et al., 2003): $\Omega_0 \sim 0.02 - 0.024h^{-2}$, roughly consistent with the above value.

Using the above baryon fraction, one infers a value of Ω_0 in the range 0.3 – 0.5. It is of course vital to have a reliable estimates of f_b to apply this test. Recently, Sadat and Blanchard (2001, sb01) have challenged this question. They first noticed that Γ is a function of radius which behaves in numerical simulations with a specific pattern: from the very central part of clusters to the outer the baryon fraction first raises up and then tends to flatten in the outer part. However the apparent baryon fraction profile as inferred by observations does not behave like this, it rather raises up continuously from the central part to the outer one. If this trend is real it would mean that our understanding of cluster formation is very poor and probably very dramatic heating processes took place during the cluster formation. However, this is probably not the case because one would expect that the gas distributions in cluster would not exhibit any regularity in their shapes, while such regularity seems to be observed (Neumann & Arnaud, 1999; Ponman, 1999; Roussel et al., 2000). Different conclusions on the baryon fraction have been reached by sb01 : a) by using the most recent measurements of clusters properties in the outer part (Vikhlinin et al., 1999) b) by applying a correction for the clumping of the gas (accordingly to Mathiesen et al. (1999) the correction factor of the order of 1.16, probably an uncertain number), c) by using mass estimator from recent numerical simulations. They showed that the baryon fraction shape in clusters is in reasonable agreement with what is seen in numerical simulations and that the numerical value could be of the order of 10% ($h = 0.5$) or even smaller. In terms of Ω_0 this corresponds to values of the order of 0.8 – 1., consistent with what has been derived from clusters abundance evolution. Finally, Sasaki (1996) has proposed a different test based on the cluster gas fraction. The first basic assumption is that this gas fraction should be universal (identical in all clusters at any redshift). The idea of the test is that the gas fraction inferred on distant clusters, depends on the cosmological parameters which come

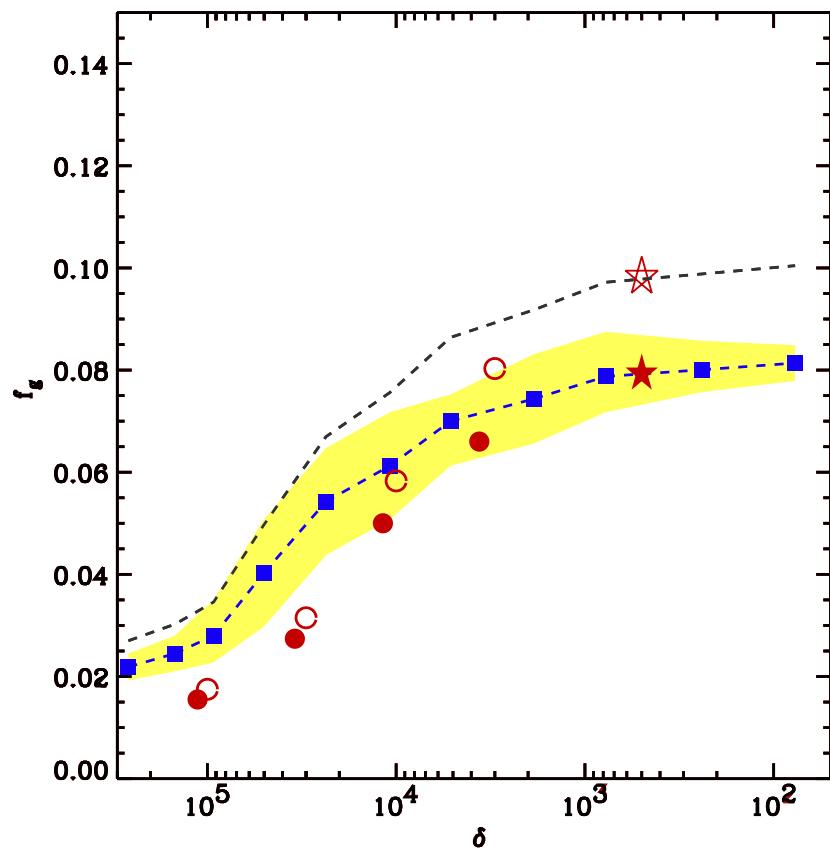


Figure 4.4. The comparison (Sadat & Blanchard, 2001) of the theoretical baryon fraction, as derived from numerical simulations, for two primordial gas fraction, $f_g = 0.11$ and $f_g = 0.09$ (dotted lines), with the observed gas fraction at different density contrasts and using two different mass estimators inferred from numerical simulations: filled symbols are values obtained with Bryan and Norman (1998) mass estimator, open symbols are obtained using Evrard et al (1996). Stars are derived from the data of Vikhlinin et al. (1999), corrected for the clumping factor (Mathiesen et al. (1999)).

in the angular distance. Therefore, by requesting that the apparent gas fraction does not evolve, one can obtain constraints on cosmological parameters.

9. Conclusion

The local TDF compared with Henry's sample at $z = 0.33$ clearly indicates that the TDF, inferred from EMSS, is evolving. This evolution is consistent with the evolution detected up to redshift $z = 0.55$ by Donahue et al. (1999). This indicates converging evidences for a high density universe, with a value of Ω_0 consistent with what Sadat et al. (1998) inferred previously from the full EMSS sample taking into account the observed evolution in the $L_x - T_x$ relation (which is moderately positive and consistent with no evolution). Low density universes with $\Omega_0 \leq 0.35$ are excluded at the two-sigma level. This conflicts with some of the previous analyses on the same high redshift sample (Henry, 1997, 2000; Eke et al., 1998; Donahue & Voit, 1999). However higher values were obtained by Viana and Liddle (1999), Sadat et al (1988) and Blanchard et al. (2000). The possible existence of high temperature clusters at high redshift like MS1054 (12 keV), cannot however be made consistent with this picture of a high density universe, unless their temperatures are overestimated by 30% or the primordial fluctuations are not gaussian. The baryon fraction in clusters is an other global test of Ω_0 , provided that a reliable value for Ω_b is obtained. However, the mean baryon fraction could have been overestimated in previous analysis, being closer to 10% (for $h = 0.5$) rather than to 15%-25%. Given that the preferred value for the primordial baryon value, this leads to high density for the universe.

Clusters provide us with the most important tests for the determination of the mean density of the Universe, which allow to suppress the degeneracies existing method based in CMB anisotropies. As we have seen, the cluster number evolution and the baryon fraction are powerfull potential tests, but which have not yet provide definitive answers and may indicate high value. Better understanding of systematic uncertainties in these methods as well as the scaling properties are necessary. For this purpose larger well controlled samples of high redshift X-ray clusters will be critical, a target that both Chandra and XMM can reach.

References

- Arnaud, M. & Evrard, A. E. 1999, MNRAS, 305, 631
- Balland, C. & Blanchard, A. 1997, ApJ, 487, 33
- Bartlett, J. & Blanchard, A. 1996, A&A, 307, 1.
- Blanchard, A., Valls-Gabaud, D. & Mamon, G 1992, A&A, 264, 365–378.
- Blanchard, A. & Bartlett, J. 1998, A&A, 314, 13

- Blanchard, A., Sadat, R., Bartlett, J. & Le Dour, M. 2000, A&A, 362, 809
 Bryan, G. L. & Norman, M. L. 1998, ApJ, 495, 80
 Donahue, M. & Voit, G. M. 1999, astro-ph/9907333, ApJL 523, L137
 Donahue, M., Voit, G. M., Scharf, C. A., Gioia, I., Mullis, C. P., Hughes, J. P.
 & Stocke, J. T. 2000, astro-ph/9906295, ApJ, 527, 525
 Efstathiou, G., Frenk, C. S., White, S. D. M. & Davis, M. 1998, MNRAS, 235,
 715
 Eke, V. R., Cole, S., Frenk, C. S. & Henry, P. J. 1998, MNRAS, 298, 1145
 Evrard, A. E, Metzler, C. A. & Navarro, J. F. 1996, ApJ, 469, 494
 Girardi, M., Borgani, S., Giuricin, G., Mardirossian, F. & Mezzetti, M. 1998,
 ApJ, 506, 45
 Henry, J. P & Arnaud, K. A. 1991, ApJ, 372, 410
 Henry, J. P. 1997, ApJL 489, L1
 Henry, J. P. 2000, ApJ 534, 565
 Hughes, J. P. 1997, astro-ph/9709272
 Jenkins, A., Frenk, C. S., White, S. D. M. et al. 2001, MNRAS, 321, 372
 Lemaître, G. 1933, Ann. Soc. Sci. Bruxelles A53, 51
 Mathiesen, B., Evrard, A. E & Mohr, J. J. 1999, ApJL 520, L21
 Markevitch, M. 1998, ApJ 503 77
 Navarro, J. F., Frenk, C.S. & White, S.D.M. 1995, MNRAS, 275, 720
 Neumann, D.M. & Arnaud, M. 1999, A&A 348, 711
 Norman, M. L. & Bryan, G. L. 1999, Lecture Notes in Physics, Berlin Springer
 Verlag, 530, 106, astro-ph/9802335
 Oukbir, J. & Blanchard, A. 1992, A&A 262, L21
 Oukbir, J. & Blanchard, A. 1997, A&A 317, 10
 Ponman, T. J., Cannon, D. B. & Navarro, J. F. 1999, Nature 397, 135
 Press, W. H. & Schechter, P. 1974, ApJ, 187, 425
 Roussel, H. Sadat, R. & Blanchard, A. 2000, A&A 361, 429
 Sadat, R., Blanchard, A. & Oukbir, J. 1998, A&A, 329, 21
 Sadat R. & Blanchard A., 2001, A&A, 371 ,19
 Sadat, R., Blanchard, A., Mendiboure, C. et al. 2004, astro-ph/0405457, A&A,
 in press.
 Sasaki, S. 1996, PASJ 48, 119
 Sheth, R. K., Mo, H. J., & Tormen, G. 2001, MNRAS, 323 1
 Spergel, D. N., Verde, L., Peiris, H. V. et al. (WMAP collab.) 2003, ApJS 148,
 175
 Tytler, D., O'Meara, J. M., Suzuki, N. & Lubin, D. 2000, astro-ph/0001318
 Physica Scripta T85, 12
 Van Waerbeke, L. et al., 2000, A&A 358, 30
 Viana, P. T. R. & Liddle, A. R., 1999, MNRAS, 303, 535
 Vikhlinin, A., Forman, W. & Jones, C. 1997, ApJ 525, 47
 White, M. 2002, ApJS 143, 241

- White, S. D. M., Navarro, J. F., Evrard, A. E. & Frenk, C. 1993, *Nature* 366,
429
- Zucca, E. et al., 1997, *A&A* 326, 477
- Zwicky, F. 1933, *Helv. Phys. Acta* 6 110

Chapter 5

ASTROPHYSICAL DETECTION OF DARK MATTER

S. Colafrancesco¹

*INAF Osservatorio Astronomico di Roma
Via Frascati 33, I-00040 Monteporzio
Italy
cola@mporzio.astro.it*

Abstract We discuss some of the astrophysical techniques to detect the presence and the nature of Dark Matter in large scale structures. We focus here on the cases of galaxies and galaxy clusters.

Keywords: Cosmology, Dark Matter, Galaxies, Galaxy Clusters.

1. Signals from the Dark universe

We live in a flat ($\Omega_0 \approx 1$), dark universe which is dominated by Dark Matter (DM) and Dark Energy (DE). Dark Matter provides a fraction $\Omega_m \approx 0.27 \pm 0.04$ of the overall matter-energy content (the rest being provided by Dark Energy with $\Omega_{DE} \approx 0.73 \pm 0.04$ with the baryonic contribution limited to $\Omega_b \approx 0.044 \pm 0.004$) and amounts to $\sim 86\%$ of the total mass content of the universe. The dark part of the universe sends us signals of the presence and of the nature of its constituents by several astrophysical probes. These probes are of *inference* and *physical* character. Inference probes are, e.g., the CMB anisotropy spectrum, the dynamics of galaxies, the hydrodynamics of the hot intra-cluster gas and the gravitational lensing distortion of background galaxies by the intervening potential wells of galaxy clusters. These probes tell us about the presence and the total amount of DM in the large scale structures from galaxies to clusters of galaxies ad to the overall universe. Physical probes tell us about the nature and the physical properties of the DM particles and come thru by the astrophysical effects of the DM interaction which can be directly or indirectly observed by astrophysical techniques.

2. Inference probes

The CMB can be considered as the largest detector for DM in the universe. In fact, the multipole structure of its anisotropy power spectrum depends on the overall amount of DM in the universe and the acoustic peaks of the CMB power spectrum are sensitive to the energy density ratio of DM to radiation. It is thanks to this fact that the BOOMERANG (first) and the WMAP (lastly) experiments were able to assess the overall amount of DM in the universe at the level of $\Omega_m \approx 0.27 \pm 0.04$ (Bennett et al. 2003).

Other estimates of the amount and of the distribution of DM in the universe come from the study of large scale structures at more recent epochs in the cosmic evolution. The reason why cosmic structures contain a record of the DM distribution in the universe is due to the fact that the evolution of the parent density perturbations was dominated by their DM content from early times on (see Peacock in these Proceedings). Thus the study of galaxies and galaxy clusters - the largest gravitationally bound structures in the universe whose potential wells are dominated by DM - provide information on both the amount of DM and on its density distribution.

Numerical simulations indicate that virialized DM halos on the scales of galaxies and galaxy clusters show a cuspy density profile $\rho(r) = \rho_0 g(r)$ where

$$g(r) = \left(\frac{r}{r_c}\right)^{-\eta} \left(1 + \frac{r}{r_c}\right)^{\eta-\xi} \quad (5.1)$$

with $\eta = 1$ and $\xi = 3$ reproducing the DM density profile that NFW (Navarro, Frenk & White 1997) claimed as "universal". Actually, the inner DM density profile found in N-body simulations can take slopes in a wider range, $r^{-0.5} - r^{-1.5}$, and it can be flattened or steepened by the presence of a baryonic feedback on DM (Gnedin & Primack 2003) or a DM-Dark Energy coupling (Macciò et al. 2003).

Observations of the inner parts of galaxies reveal that the DM density profile at scales $r \lesssim 1$ kpc from the galaxy center is much flatter than the NFW profile and that there is no evidence of a cusp at distances $r \lesssim 0.2$ kpc (see Fig.5.1). At $r \lesssim 2$ pc, the supermassive Black-Hole supposed to be at the galaxy centres dominates the galactic dynamics and its interaction with surrounding medium may lead to a complex galaxy density profile (see, e.g., discussion in Gnedin & Primack 2003). It should be noticed, on the other hand, that temperature profiles observed in Elliptical galaxies with Chandra from 0.7 to 35 kpc (as in NGC 4636) indicate that $M_{DM} \sim 50 - 80\% M_{total}$ and that the composite mass distribution has a steep slope with no core (Lowenstein & Mushotzky 2002). Based on the available evidence, we should conclude that there is not yet a definitive conclusion on the shape of the DM density profile in the galaxy centres.

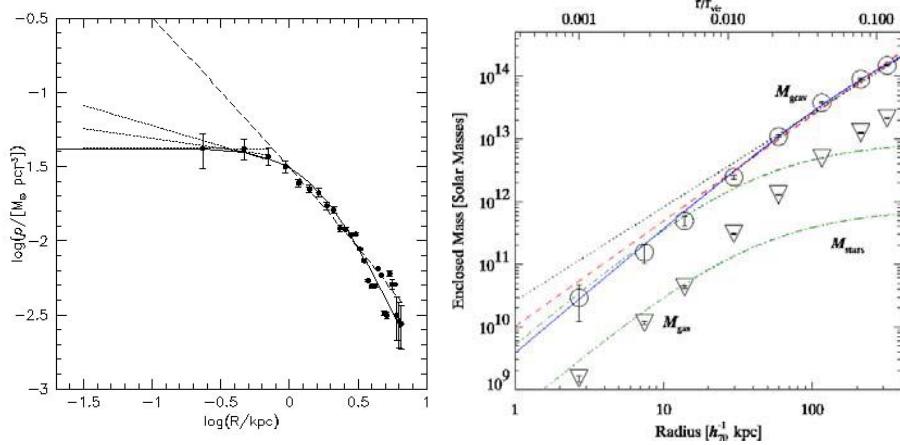


Figure 5.1. **Left.** The DM density profile in the inner parts of the galaxy NGC 6822 does not reveal a density cusp (Weldrake et al. 2003). The case of NGC2976 confirms that $\eta = 0.27$ at $r < 1.8$ Kpc (Simon et al. 2003). **Right.** The total gravitating mass (open circles) in the cluster A2029 (see the recent review by Buote 2003) overlaid with different models: NFW (solid), power-law (dashed) and Moore et al. (dotted). The gas mass (triangles) is also shown for comparison.

Observations of galaxy cluster do not help in this respect since the available data stop our understanding of the DM density profile at distances $r \lesssim 10$ kpc from their centres (see Fig.5.1). Down to these scales the NFW is still allowed and at smaller scales the inner slope remains quite uncertain even when the combined analysis of X-ray, gravitational lensing and galaxy dynamics data are taken into account (Dalal & Keeton 2003).

On the theoretical side, the uncertainty in the modeling of the DM distribution is nonetheless enhanced by at least two issues: i) the cusp problem may be alleviated by changes in the basic physics (e.g., self-interaction DM particles with large annihilation cross section, broken scale-invariance and modified gravity) or by the baryon-DM coupling and interaction (e.g., gas outflow during early stages of galaxy formation); ii) the cusp problem may be strengthened by the possible DM - Dark Energy coupling (e.g., modified particle dynamics).

The previous inference probes can be used to assess the presence, the amount and the spatial distribution of the DM in large scale structures but are not able to provide information on its nature.

3. Physical probes

The nature of the DM basic constituents is still unknown. Among the viable competitors for having a cosmologically relevant DM species (see, e.g., Gonçolo in these Proceedings), the leading candidate is the lightest supersymmetric

(SUSY) particle, plausibly the neutralino χ , with a mass M_χ in the range between a few GeV to a few hundreds of GeV. Physical probes on the nature of the DM building up the large-scale structure gravitational potential field can be obtained by studying the interaction of the DM particles (mainly their annihilation) through the relative signals of the interaction/annihilation in the galaxy (or galaxy cluster) atmospheres. These signals involve, in the case of a χ DM, emission of gamma-ray, neutrinos, synchrotron and bremsstrahlung radiation together with the Compton scattering of the CMB photons by the secondary electrons produced in the DM annihilation process.

The products of DM annihilation in large-scale structures

[Colafrancesco 2001]

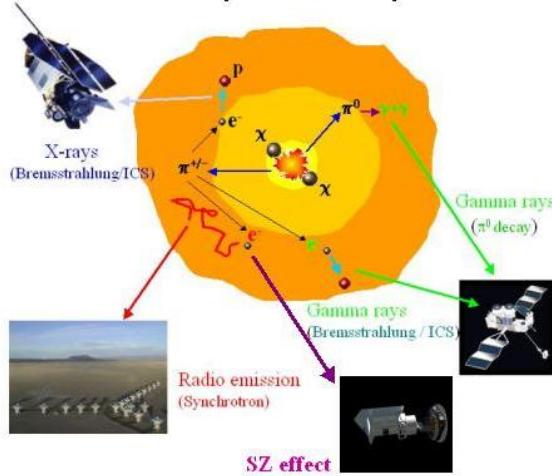


Figure 5.2. A simple model which shows the basic astrophysical mechanisms underlying the search for the nature of DM particles in large-scale structures (galaxies and galaxy clusters).

The χ annihilation rate in a DM halo is $R = n_\chi(r)\langle\sigma V\rangle_A$, where $n_\chi(r) = n_{\chi,0}g(r)$ is the neutralino number density and $\langle\sigma V\rangle_A$ is the $\chi\chi$ annihilation cross section averaged over a thermal velocity distribution at freeze-out temperature (see, e.g., Gondolo in these Proceedings). Although the $\chi\chi$ annihilation cross section is a non-trivial function of the mass and physical composition of the neutralino, to our purpose it suffices to recall that the χ relic density is approximately given by:

$$\Omega_\chi h^2 \simeq \frac{3 \times 10^{-27} \text{ cm}^3 \text{ s}^{-1}}{\langle\sigma V\rangle_A} \quad (5.2)$$

Hence, for the $\chi\chi$ annihilation, we shall assume a total cross section of $\langle\sigma V\rangle_A \approx 2.6 \cdot 10^{-26} \text{ cm}^3 \text{ s}^{-1}$ to be consistent with the value $\Omega_m h^2 \sim 0.116$ derived from WMAP (Bennett et al. 2003). For values of $\Omega_m h^2$ in the range $0.085 - 0.152$, the annihilation cross section is fixed to within a factor less than three. Detailed studies of the relic neutralino annihilation show that the above value is well inside the allowed range predicted in supersymmetric theories for a wide choice of masses and physical compositions of neutralinos that can be relevant as CDM candidates. Enhancing (suppressing) the χ annihilation rate will have on our results the simple effect of rescaling the final electron spectra by the same enhancement (suppression) factor. Neutralinos which annihilate inside a DM halo produce quarks, leptons, vector bosons and Higgs bosons, depending on their mass and physical composition. Electrons are then produced from the decay of the final heavy fermions and bosons (monochromatic electrons, with energy about M_χ , coming from the direct channel $\chi\chi \rightarrow ee$, are in general much suppressed, Turner & Wilczek 1990). The different composition of the $\chi\chi$ annihilation final state will in general affect the form of the final electron spectrum. Analytical expressions for the e^\pm spectrum has been given by Colafrancesco & Mele (2001) and we refer to this paper for further details.

The time evolution of the electron spectrum is given by the transport equation:

$$\frac{\partial n_e(E, r)}{\partial t} - \frac{\partial}{\partial E} \left[n_e(E, r) b(E) \right] = Q_e(E, r) \quad (5.3)$$

where $n_e(E, r)$ is the equilibrium spectrum at distance r from the cluster center for the electrons with energy E . The source electron spectrum, $Q_e(E, r)$, rapidly reaches its equilibrium configuration mainly due to synchrotron and Inverse Compton Scattering losses at energies $E \gtrsim 150 \text{ MeV}$ and to Coulomb losses at smaller energies (Colafrancesco & Mele 2001). Since these energy losses are efficient in the cluster atmosphere and DM annihilation continuously refills the electron spectrum, the population of high-energy electrons can be described by a stationary transport equation

$$-\frac{\partial}{\partial E} \left[n_e(E, r) b_e(E) \right] = Q_e(E, r) \quad (5.4)$$

from which the equilibrium spectrum can be calculated. Here, the function $b_e(E)$ gives the energy loss per unit time at energy E

$$b_e(E) = b_0(B_\mu) \left(\frac{E}{\text{GeV}} \right)^2 + b_{Coul}, \quad (5.5)$$

where $b_0(B_\mu) = (2.5 \cdot 10^{-17} + 2.54 \cdot 10^{-18} B_\mu^2)$ and $b_{Coul} = 7 \times 10^{-16} [n(r)/1\text{cm}^{-3}]$, if b_e is given in units of GeV/s. The source spectra we derived (see Fig.1 in Colafrancesco & Mele 2001 arise from an analytic approximation of the exact

shape of the electron spectrum that tries to cope with the details of the quarks and leptons decays and of the hadronization of the decay products. Detailed electron spectra can also be obtained by using state-of-the-art Monte Carlo simulations, although the analytical approximations used here can resume the relevant aspects of more detailed studies.

Gamma rays emission is predominantly due to the hadronization of the decay products of $\chi\chi$ with the production of a continuum gamma-ray spectrum due to the decay $\pi^0 \rightarrow \gamma + \gamma$ (Colafrancesco & Mele 2001), even though the direct neutralino annihilation results in a line emission feature at an energy $\sim M_\chi$. The continuum gamma-ray emission expected from $\chi\chi$ annihilation has a specific spectral shape (see Fig.5.3) which might be observable with the next generation gamma-ray experiments (both from space and ground-based). Gamma-ray emission is the most direct signal of DM annihilation and it could be revealed provided that we will have sufficient spectral and spatial resolution as well as a clear understanding of other competing gamma-ray emission mechanisms (like cosmic-ray acceleration) expected to be at work in DM halos (see Colafrancesco in these Proceedings). Gamma-ray emission

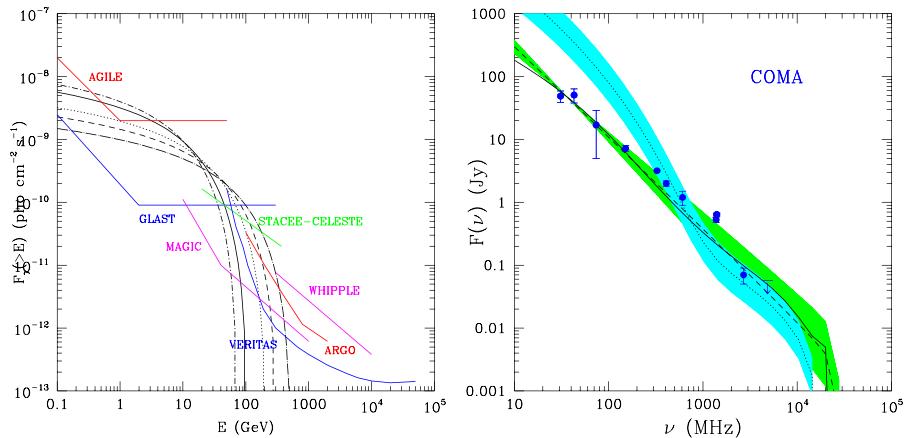


Figure 5.3. Left. The gamma-ray emission from $\chi\chi$ annihilation in a rich, Coma-like, nearby galaxy cluster is shown $M_\chi = 70 - 500$ GeV (from top down). The integral flux is compared to the sensitivity of ongoing and planned gamma-ray experiments, as labelled. *Right.* The diffuse synchrotron emission spectrum of secondary electrons produced in $\chi\chi$ annihilation is shown to fit the Coma radio-halo spectrum: the green area represent the prediction of a model in which the χ annihilates predominantly into fermions, while the blue area represent the gauge-boson dominated χ annihilation (from Colafrancesco & Mele 2001).

is also expected from secondary electrons produced by $\chi\chi$ annihilation via $\chi\chi \rightarrow \pi^\pm + X$, $\pi^\pm \rightarrow \mu^\pm \nu_\mu (\bar{\nu}_\mu)$, $\mu^\pm \rightarrow e^\pm + \bar{\nu}_\mu (\nu_\mu) + \nu_e (\bar{\nu}_e)$. These secondary electrons may produce gamma-rays through bremsstrahlung and Compton Scattering of CMB photons to high energies (Colafrancesco 2004a).

The same secondary electrons produced by $\chi\chi$ annihilation unavoidably produce synchrotron emission in the magnetized atmosphere of galaxy clusters and galaxies which can be observed at radio frequencies as a diffuse and extended radio-halo centered on the center of the DM halo. Radio observations of cluster radio-halo data are, in principle, very effective in putting constraints on the mass and composition of the neutralino (Colafrancesco & Mele 2001), under the hypothesis that DM annihilation provides a major contribution to the radio-halo flux (however, we remind the reader that other possibilities have been envisaged for the origin of radio halos in the context of cosmic ray acceleration models, see Colafrancesco in these Proceedings). If DM annihilation is the leading process for the production of secondary electrons with \gtrsim GeV energies in the cluster atmospheres, a pure energy requirement says that the neutralino mass is bound to be $M_\chi \geq 23.4 \text{GeV}(\nu/\text{GHz})^{1/2}(B/\mu\text{G})^{-1/2}$ to emit at frequencies $\nu \geq 1 \text{GHz}$, as observed in cluster radio halos (see Fig.5.3). The available data indicate that a neutralino responsible for the cluster radio halo emission should have $M_\chi \gtrsim 52 \text{GeV}(B/\mu\text{G})^{-1/2}$ and should annihilate predominantly into fermions, consistently with the available accelerator limits.

The production of secondary electrons in DM annihilation can also unavoidably provides an heating of the intracluster gas in the cluster cores by Coulomb scattering (Colafrancesco 2004a; Totani 2004) and also the Compton Scattering (CS) of CMB photons up to higher frequencies (Colafrancesco & Mele 2001; Colafrancesco 2004a). The relevance of these mechanisms to the observed quenching of cooling flows in the cluster cores (Colafrancesco, Dar & DeRujula 2004) and to the emission features of galaxy clusters at extreme UV and hard X-ray frequencies (see Colafrancesco in these Proceedings for a review) is still subject to a detailed exploration (Colafrancesco 2004a), but it will provide additional constraints to the neutralino physical properties.

It has been recently proposed (Colafrancesco 2004b) that the unavoidable Compton Scattering of CMB photons by the secondary electrons produced in $\chi\chi$ annihilation also provides an additional source of SZ effect, that we call here SZ_{DM} , with specific spectral and spatial features. According to these results, the DM induced spectral distortion can be written as

$$\Delta I_{DM}(x) = 2 \frac{(k_B T_0)^3}{(hc)^2} y_{DM} \tilde{g}(x) , \quad (5.6)$$

where T_0 is the CMB temperature and the Comptonization parameter y_{DM} is given by

$$y_{DM} = \frac{\sigma_T}{m_e c^2} \int P_{DM} d\ell , \quad (5.7)$$

in terms of the pressure P_{DM} contributed by the secondary electrons produced by neutralino annihilation. The quantity $y_{DM} \propto \langle \sigma V \rangle_A n_\chi^2$ and scales as $\propto \langle \sigma V \rangle_A M_\chi^{-2}$, providing an increasing pressure P_{DM} and optical depth

$\tau_{DM} = \sigma_T \int d\ell n_e$ for decreasing values of the neutralino mass M_χ . The function $\tilde{g}(x)$, with $x \equiv h\nu/k_B T_0$, for the SZ_{DM} effect contains all the spectral features of the effect and is interestingly different from the leading thermal SZ effect which is well studied in galaxy clusters (see fig.5.4). A major difference

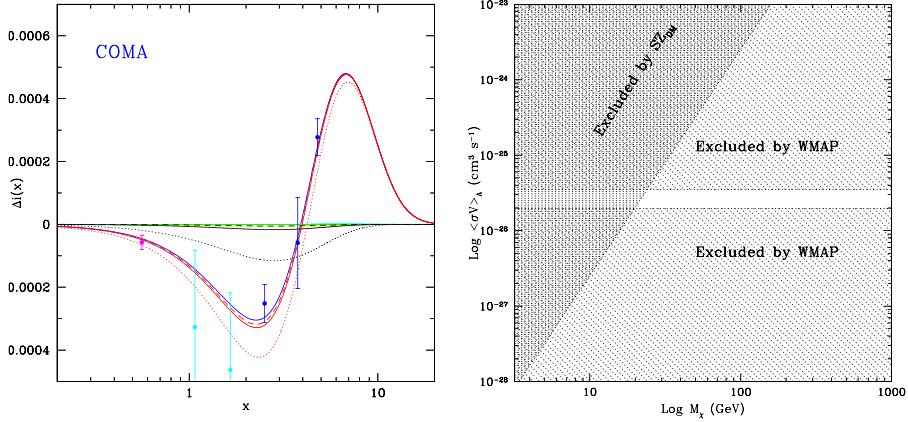


Figure 5.4. **Left.** The overall SZ effect in Coma produced by the combination of various electron populations: thermal hot gas with $k_B T = 8.2$ keV and $\tau = 4.9 \cdot 10^{-3}$ (solid blue curve) which best fits the available SZ data; relativistic electrons which best fit the radio-halo spectrum (yellow curve) provide a small additional SZ effect (Colafrancesco 2004a); warm gas with $k_B T \approx 0.1$ keV and $n \approx 10^{-3} \text{ cm}^{-3}$ (cyan curve) provides a small SZ effect due to its low pressure (Colafrancesco 2004c); DM produced secondary electrons with $M_\chi = 10$ (black dotted curve), 20 GeV (black solid curve) and 30 GeV (dashed solid curve). A pure-gaugino χ reference model is assumed in the computations. The relative overall SZ effect is shown as the dotted, solid and dashed red curves, respectively. A zero peculiar velocity of Coma is assumed consistently with the available limits. SZ data are from OVRO (magenta), WMAP (cyan) and MITO (blue). **Right.** The constraints on the $\langle \sigma V \rangle_A - M_\chi$ plane set by the SZ effect from Coma. The heavily dashed area is excluded by the analysis of the SZ_{DM} . The SZ constraints are combined with the WMAP constraint $0.085 \lesssim \Omega_m h^2 \lesssim 0.152$ which are translated on the quantity $\langle \sigma V \rangle_A$ (from Colafrancesco 2004b).

between the two SZ effects is the different position of the zero of the SZ effect which is moved to higher frequencies in the case of the DM produced electrons with respect to the case of the thermal distribution. As a consequence, the SZ_{DM} effect appears as a negative contribution to the overall SZ effect at all the frequencies which are relevant for the SZ experiments, $x \sim 0.5 - 10$. Note that the amplitude of the SZ_{DM} effect increases for decreasing values of the neutralino mass since the quantity $y_{DM} \propto \langle \sigma V \rangle_A n_\chi^2 \sim M_\chi^{-2}$. This property of the SZ_{DM} allows to set constraints on the neutralino mass. In fact, to be consistent with the SZ data for the Coma cluster, the SZ_{DM} effect cannot be due to neutralinos with $M_\chi \lesssim 20$ GeV for the value of $\langle \sigma V \rangle_A$ which sets the

neutralino relic abundance. Since SZ_{DM} depends on the quantity $\langle \sigma V \rangle_A n_\chi^2$, the available SZ data set actually an upper limit in the $\langle \sigma V \rangle_A - M_\chi$ plane (see Fig.5.4). Models with large values of $\langle \sigma V \rangle_A$ and low values of M_χ which are found in the shaded area are excluded by the excess SZ_{DM} effect in Coma. The limits on $\langle \sigma V \rangle_A$ set by WMAP (Bennett et al. 2003) restrict further on the available region of the $\langle \sigma V \rangle_A - M_\chi$ plane. Because the amplitude of the SZ_{DM} effect increases with decreasing values of M_χ , The high sensitivity planned for the future SZ experiments, especially at frequencies $x \approx 2.5$ and $x \gtrsim 3.8$, where the SZ_{DM} more clearly manifests itself, can provide much stringent limits to the additional SZ effect induced by DM annihilation. In this context, the next coming PLANCK-HFI experiment has enough sensitivity to probe in details the contributions of various SZ effects in the frequency range $x \approx 2 - 5$. An exciting possibility in this context could be offered by nearby systems which are gravitationally dominated by DM, which contain little or no gas (in either hot or warm forms) and show absence of non-thermal phenomena connected with the presence of cosmic rays. In such ideal DM systems, the major source of SZ effect would be just the one due to the annihilation of the WIMPs. Systems which could be assimilable to the ideal "pure" DM halos are dwarf spheroidal galaxies and/or low surface brightness galaxies. These systems seem to be ideal sites for studying the DM annihilation indirect signals which reveal themselves in a variety of astrophysical phenomena. In such a context, the possible detection of the DM induced SZ effect will provide an important complementary approach which can be studied by more traditional astronomical techniques.

4. Conclusion

The nature of Dark Matter, despite its overwhelming evidence, is still intangible. It is nonetheless appealing, in these respects, that some astrophysical features of galaxy clusters and of galaxies might give information on the fundamental properties of the DM particles.

References

- Bennett, C. L. et al. 2003, ApJS, 148, 175
- Buote, D. A. 2003, preprint astro-ph/0310579
- Colafrancesco, S. 2004a, Journal of Nuclear Physics, in press
- Colafrancesco, S. 2004b, A&A, in press (astro-ph/0405456)
- Colafrancesco, S. 2004c, in 'Soft X-ray excess emission from clusters of galaxies', R. Lieu & J. Mittaz Eds., p.137-146 and p.147-154 (astro-ph/0403404)
- Colafrancesco, S. & Mele, B. 2001, ApJ, 562, 24
- Colafrancesco, S., Dar, A. & De Rujula, A. 2004, A&A, 413, 441
- Dalal, N. & Keeton, C. R. 2003, preprint astro-ph/0312072

- Gnedin, O. Y. & Primack, J. 2003, preprint 0308385
Lowenstein & Mushotzky, R. F. 2002, preprint astro-ph/0208090
Macció, A. et al. 2003, preprint astro-ph/0309671
Navarro, J., Frenk, C. & White, S. D. M. 1997, ApJ, 490, 493
Simon, J. et al 2003, ApJ, 596, 957
Totani, T. 2004, Phys.Rev.Lett. in press, preprint astro-ph/0401140
Turner, M. S. & Wilczek, F. 1990, Phys.Rev.D, 42, 1001
Weldrake, D. T. F. et al. 2003, MNRAS, 340, 12

Chapter 6

NON-THERMAL AND RELATIVISTIC PROCESSES IN GALAXY CLUSTERS

S. Colafrancesco

*INAF Osservatorio Astronomico di Roma
Via Frascati 33, I-00040 Monteporzio
Italy
cola@mporzio.astro.it*

Abstract We present the observational evidence and the theoretical indications for the presence of relativistic particles (cosmic rays) in galaxy clusters. We discuss the basic ideas for their origin and explore the astrophysical techniques to unveil their nature.

Keywords: Cosmology, Galaxy Clusters, Cosmic Rays

Introduction

Clusters of galaxies are the largest gravitationally bound structures in the universe and, according to the viable cosmological scenario for structure formation, they are also the youngest structures in the universe. These structures are also the largest containers of cosmic material: *i*) they contain baryonic matter in condensed (stars, galaxies with a mass fraction $M_*/M \sim 0.013$) and diffuse (hot and warm intra-cluster medium - hereafter ICM - with a mass fraction $M_{ICM}/M \sim 0.13$) forms as probed by optical and X-ray emission, respectively; *ii*) the vast majority ($\sim 90\%$) of their matter content is constituted by Dark Matter (DM) which forms their gravitational potential wells probed by the gravitationally distorted images of lensed background galaxies; *iii*) a third component of the cluster atmospheres is constituted by a population of relativistic particles with a non-thermal energy distribution whose presence is mainly probed by the diffuse synchrotron radio-halo emission. As opposed to galaxies, clusters are essentially closed boxes and, as such, they retain invaluable information on the origin, evolution and the physical processes affecting

the baryonic and non-baryonic matter content of the universe. Clusters should be fair samples of the universe, and studies of their evolution can place strong constraints on structure formation scenarios and on cosmological parameters (see Blanchard at this Meeting). Since clusters are DM dominated, the detailed physics of cooling and star formation are much less important than in galaxies. However, while gravity is clearly the dominant interaction in these massive systems, several processes of non-gravitational origin are claimed to be responsible for other features like the excess entropy of the gas in low-mass systems, the heating at the cluster centers which counterbalance the onset of cooling flows, the origin of the extended radio halos, the nature of the emission excesses observed in the soft X-ray (and Extreme UV) and in the Hard X-Ray (HXR) energy ranges and the possible non-thermal and relativistic Sunyaev-Zel'dovich effect. Thus, the understanding of the role of non-thermal and relativistic processes in cluster evolution is, on one side required to extract the truly thermal features (X-rays, SZE) used for cosmological studies, and on the other side it is motivated by the mounting evidence for the complexity of physical phenomena (radio halos, EUV and HXR emission, cooling flows quenching) for which we do not have yet a clear explanation.

1. Non-thermal and relativistic phenomena in galaxy clusters

Relativistic particles do exist in the atmospheres of galaxy clusters. There are several evidences for their presence, energy spectra and spatial distribution. Many galaxy clusters (Giovannini et al. 2000; Kempner & Sarazin 2001) contain, in fact, large-scale diffuse radio sources (radio halos and relics) which have no obvious connection with the cluster population of radio galaxies, but are rather associated with the physical phenomena occurring in the cluster atmospheres (see, e.g., Feretti 2003 for a recent observational review). Diffuse synchrotron radio halo emission is observed in the frequency range $\nu \gtrsim 10$ MHz to ~ 30 GHz (see Fig.6.1) with a spectrum $J_\nu \propto n_{rel} B^{\alpha_r+1} \nu^{-\alpha_r}$, where n_{rel} is the relativistic electron density, B is the magnetic field and α_r is in the range $\sim 1 - 1.5$. To emit synchrotron radiation at these frequencies in a intra-cluster (IC) magnetic field of order $B_\mu \equiv (B/\mu G) \sim 1$, the electrons must have energies $E_e \approx 16.4 \text{ GeV } (\nu/\text{GHz})^{1/2} B_\mu^{-1/2}$ which fall in the range $\sim 3 - 50$ GeV (we adopt here the approximation that electrons radiate at a single frequency). The notable precision of the radio-halo spectra data (see Fig.6.1 allows to constrain the shape of the relativistic electron equilibrium spectrum, $n_{rel} \sim E_e^{-x}$, since the radio-halo spectral index α_r is directly related to the electron spectral index x by $\alpha_r = (x - 1)/2$ (e.g., Longair 1993). The observed values $\alpha_r \approx 1 - 1.5$ correspond to $x \approx 3 - 4$. Synchrotron emission from radio halos yields also a direct indication of the existence of

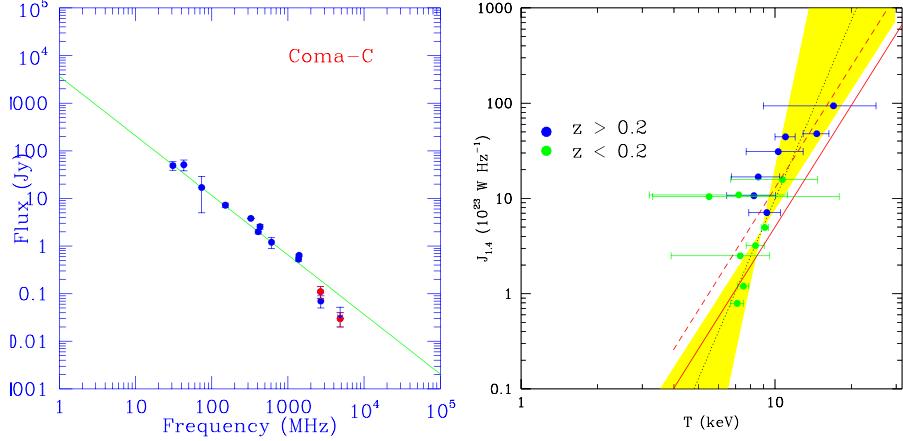


Figure 6.1. **Left.** The Coma radio-halo spectrum. The power-law fit to the data at $\nu \leq 1.4$ GHz corresponds to an electron spectral index $x = 3.5$ (see text for details). Data compilation is taken from Thiebarach et al. 2002. **Right.** The Radio luminosity $J_{1.4}$ – IC temperature T correlation for nearby radio-halo clusters (from Colafrancesco 1999; Colafrancesco & Mele 2001). We show the best fit and 1σ uncertainty region and the predictions of secondary models at $z < 0.2$ (solid line) and $z > 0.2$ (dashed line).

a magnetic field filling the cluster atmospheres. The IC magnetic field B is expected to have a decreasing radial profile after the cluster reaches its virial equilibrium (e.g., Concalves & Friaca 1999). A decreasing radial profile is also expected based on the correlation of the value of B with the IC gas density (Dolag et al. 2001). Estimates of its amplitude are quite uncertain and depend on the considered observational method (Carilli & Taylor 2002). Faraday Rotation measures local (central) values with amplitude $B_0 \sim 5 - 75 \mu\text{G}$ (Eilek 1999). The spatial distribution of radio-halo brightness seem to resemble the spatial distribution of the cluster X-ray brightness (Govoni et al. 2001) indicating a connection between non-thermal and thermal phenomena. However, the best indication of the connection between non-thermal and thermal phenomena in clusters is provided by the radio-halo luminosity J_ν – cluster temperature T correlation (Colafrancesco 1999; Liang et al. 2000; Colafrancesco & Mele 2001; see Fig. 6.1).

The cosmic ray electrons which are responsible for the radio-halo and relic synchrotron emission inevitably Compton scatter the CMB (as well as other local background) photons which will then gain energy and emit at higher energy $E \approx 2.7 \text{ keV} (E/\text{GeV})^2$. Electrons with $E \gtrsim$ a few GeV produce emission in the HXR range, while electrons with $E \lesssim 400 \text{ MeV}$ produce soft X-rays and UV emission. There is actually evidence for an excess of emission w.r.t. the thermal bremsstrahlung emission by the hot IC gas in about 20 nearby clusters

observed by BeppoSAX and RXTE (Nevalainen et al. 2004). The shape of the HXR emission observed towards the Coma cluster is still consistent with ICS of CMB photons off a non-thermal electron population with energies $E_e \gtrsim 3.3$ GeV whose spectrum should have a slope $x \sim 3.5 - 4$. Such ICS emission is expected to be spatially extended and diffuse, similarly to the parent radio-halo emission. The lack of spatial resolution of the BeppoSAX PDS and RXTE instruments do not allow, however, to test directly this hypothesis. Even though the most recent analyses tend to exclude a large contamination of the HXR flux by heavily absorbed Sy2 galaxies (Nevalainen et al. 2004) and by Blazars, the definite solution of this problem still waits for hard X-ray spectral imaging observations planned for the next future (NexT, Hexist). The ICS interpretation of the HXR excess in Coma also requires a quite large density of relativistic electrons, so large that the IC magnetic field required to accomodate for the simultaneous synchrotron radio emission is $B \lesssim 0.2 \mu\text{G}$. Such a low value of B contrasts with the lower limit $B \gtrsim 0.4 \mu\text{G}$ derived from the EGRET gamma-ray upper limit on Coma (see, e.g., Sreekumar et al. 1996, Colafrancesco & Mele 2001).

Several galaxy clusters show also an emission of extreme UV (Lieu et al. 1996, Durret et al. 2002) and soft X-ray (Bonamente et al. 2002, Kaastra et al. 2002) radiation in excess w.r.t. the thermal bremsstrahlung emission. This EUV emission excess may be consistent with both ICS of CMB photons off a non-thermal electron population (e.g., Lieu et al. 1999, Bowyer 2000) with $E_e = 608.5 \text{ MeV} (\hbar\nu/\text{keV})^{1/2} \gtrsim 149 \text{ MeV}$ for $\hbar\nu \gtrsim 60 \text{ eV}$, and with thermal emission from a warm gas at $k_B T_e \lesssim 1 \text{ keV}$ (Bonamente et al. 2002). In the case of Coma, the simple extrapolation of the ICS spectrum which fits the HXR excess down to energies $\lesssim 0.25 \text{ keV}$ does not fit the EUV excess measured in Coma because it is too steep and yields a too high flux compared to the measured flux by the EUV satellite in the $0.065 - 0.245 \text{ keV}$ band (Ensslin & Biermann 1998). Thus, under the assumption that the HXR and the EUV emission of Coma is produced by ICS of CMB photons, the minimal requirement is that a break in the electron spectrum should be present in the range $0.3 - 2.8 \text{ GeV}$ in order to avoid an excessive EUV contribution by the ICS emission and to be consistent with the radio halo spectrum.

Gamma-ray emission is also unavoidably associated to the presence of relativistic particles (cosmic ray - CR - electrons and protons) which are present in the cluster atmospheres. Primary CR electrons are able to produce gamma-ray emission either through direct bremsstrahlung radiation or through ICS of CMB (and other local background) photons. Primary CR protons can produce gamma-ray emission also through additional channels: the neutral pion decays into gamma-ray photons, $p + p \rightarrow \pi^0 + X$ with $\pi^0 \rightarrow \gamma + \gamma$, and the gamma-ray emission associated to the presence of secondarily produced electrons, $p + p \rightarrow \pi^\pm + X$, $\pi^\pm \rightarrow \mu^\pm \nu_\mu (\bar{\nu}_\mu)$, $\mu^\pm \rightarrow e^\pm + \bar{\nu}_\mu (\nu_\mu) + \nu_e (\bar{\nu}_e)$.

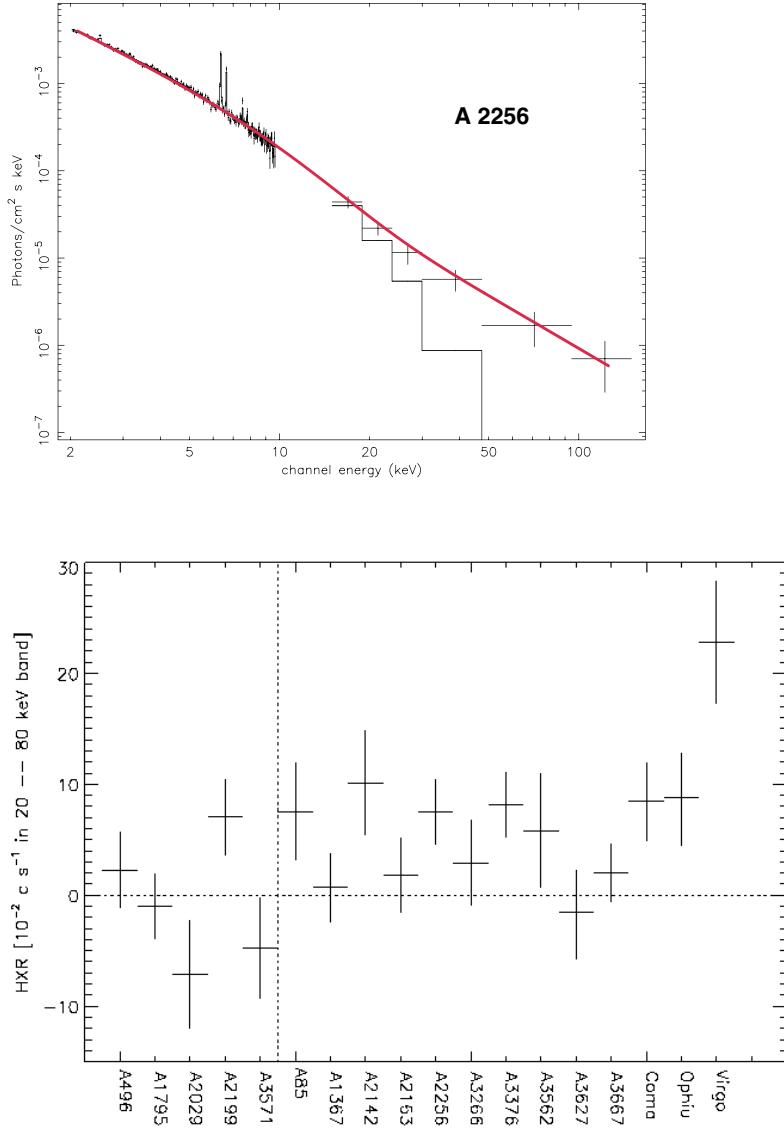


Figure 6.2. **Top.** The HXR emission from the cluster A2256 (data from Fusco-Femiano et al. 2000) together with the fit provided by the warming ray model of Colafrancesco et al. 2004 (thick line). **Bottom.** The 20-80 keV luminosity of the clusters which show evidence of HXR emission (from Nevalainen et al 2004).

This mechanism has an essential ingredient in the confinement of the CRs within clusters where they are produced (e.g., Colafrancesco & Blasi 1998).

A similar mechanism of gamma-ray emission can be also provided by the annihilation of Dark Matter particles (see Colafrancesco at this Meeting) which produce both direct gamma-ray emission and secondarily produced gamma-rays through $\pi^0 \rightarrow \gamma + \gamma$, as above.

There is not yet, however, a definite detection of diffuse gamma-ray emission from galaxy clusters. While there is a preliminary evidence of gamma-ray emission from a dozen bright, radio-active clusters which host powerful radio galaxies and Blazars and are associated to unidentified EGRET sources (Colafrancesco 2002), many of the quiet, X-ray selected clusters only have upper limits for their emission at $E > 100 \text{ MeV}$.

The presence of a substantial CR activity in the central regions of galaxy clusters has also been suggested to quench cooling flows (see, e.g., Colafrancesco, Dar & De Rujula 2004 and reference therein). Deep Chandra images of the cluster cooling region also show the emergence of extended bubbles of relativistic particles which are floating in the ICM, and of a substantial interaction of the radio-galaxy jets with the surrounding hot gaseous medium (e.g., McNamara 2003).

To summarize, the full picture of the cluster emission properties across the e.m. spectrum spans over more than 16 orders of magnitude in energy (see Fig.6.3). The hot thermal electron distribution with $T_e \sim 10^8 \text{ K}$ and $n_e \sim 10^{-3} \text{ cm}^{-3}$ provides the bulk of the thermal X-ray emission (see Forman at this Meeting). In addition, there is preliminary evidence also for a possible warm component with temperature $T_{\text{warm}} \sim 10^6 \text{ K}$ and much more uncertain density. A population of relativistic electrons with a non-thermal spectrum (whose origin is still uncertain) is expected in the energy range $\sim 0.2 - 30 \text{ GeV}$ and possibly extending to higher energies. The combination of these data provides overall constraints on the spectrum of the thermal (warm and hot) and relativistic particles (electrons and hadrons) which are present in the cluster atmospheres (see Fig.6.3).

2. The origin of cosmic rays in galaxy clusters

The available evidence suggests the simplified model for the cluster structure which is shown in Fig.6.4. According to this picture, the properties of CRs in the cluster atmospheres can be studied from multi-frequency observations. These observations will provide the ultimate answer to the origin of cosmic rays in galaxy clusters and in large scale structures. In the following, we will briefly introduce the basics of the available models for the origin of cosmic rays and for the related non-thermal phenomena observed in galaxy clusters.

Cosmic rays (electrons and protons) are actually expected to be produced or injected in the cluster atmospheres by different mechanisms. X-ray observations and N-body simulations both indicate the existence of shocks in the

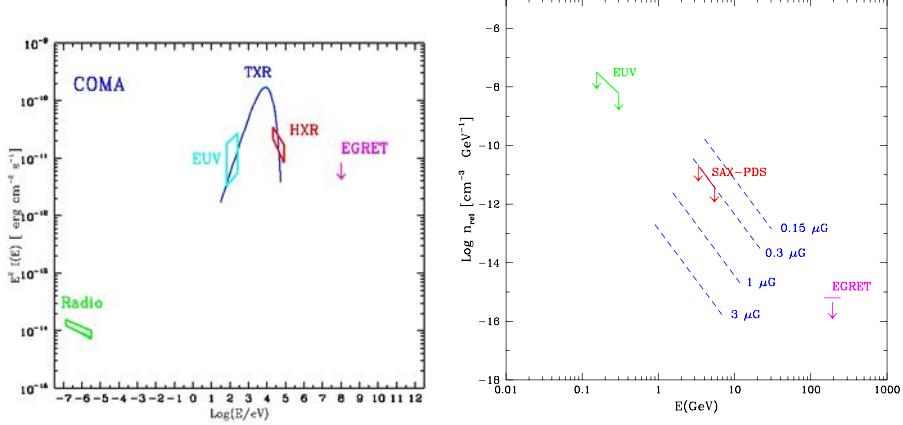


Figure 6.3. **Left.** The overall observed spectrum of the Coma cluster from radio to gamma-ray frequencies (see labels). **Right.** The constraints to the spectrum of relativistic electrons in Coma as obtained from different observation (see labels): the radio halo data (blue dashed lines) for different values of the IC magnetic field; the HXR data (red solid line with arrows), the EUV data (green solid line with arrows) and the EGRET upper limit (magenta arrow). The arrows indicate that the spectra should be considered as upper limits, because we cannot exclude that a fraction of the HXR and EUV flux is provided by active galaxies or warm gas, respectively.

ICM (see Fig.6.5) which can accelerate particles to relativistic energies. CRs can also be injected in the cluster atmosphere either by the energetic jets of active galaxies (e.g., Colafrancesco & Blasi 1998) or by the GRBs-cannonballs developing inside the many cluster galaxies (e.g., Colafrancesco et al. 2004). Finally, CRs with large enough energies (up to the parent DM particle mass) may also be produced as secondary decaying products in the annihilation of DM particles like neutralinos (e.g., Colafrancesco & Mele 2001).

The large size ($\sim Mpc$) of radio halos compared to the short paths traveled by high-E electrons - which rapidly loose their energy due to Compton and synchrotron losses (Longair 1993) - require, however, either an efficient and isotropic re-acceleration mechanism or an *in-situ*, stationary production.

Two distinct families of models for the CR origin of non-thermal phenomena in galaxy clusters have been proposed so far: *i*) the electronic and the *ii*) hadronic models.

Electronic models (first proposed by Jaffe 1977) deal with primarily accelerated electrons and rely on the existence of an efficient and *omnipresent* re-acceleration mechanism. These models also invoke low values of the volume averaged magnetic field $B_\mu \lesssim 0.2$ (Fusco-Femiano et al. 2004), require a rather high CR injection to fit the available HXR data, predict consequently a substantial amount of gamma-ray emission which is spatially concentrated,

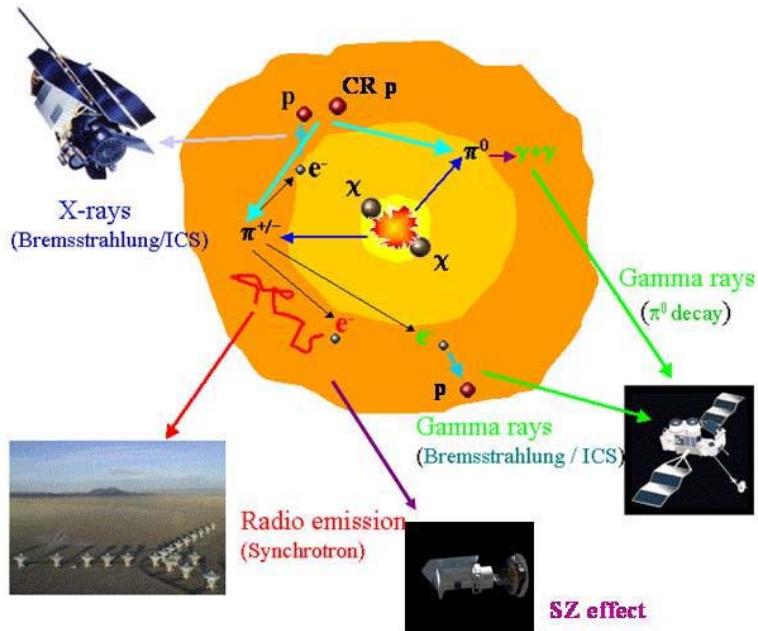


Figure 6.4. We show a simplified cluster model which describes the role of thermal and non-thermal particles (see text for details).

both in space and in time, around the shocks, and provide a modest amount of feedback on the IC gas (Colafrancesco 1999, Miniati et al. 2001). Since the main difficulty of this model is the need for an efficient and continuous re-acceleration of the initially accelerated electrons, two main mechanisms have been proposed so far: re-acceleration by shock waves produced by cluster merging and extended IC turbulence. Internal shocks produced from merging are small, fragmented and diffused. They have low Mach number and may provide $\sim 95\%$ of the gas thermalization efficiency as well as $\sim 90\%$ of the acceleration efficiency (Ryu et al. 2003). External-accretion shocks are formed at the outskirts of clusters as the result of the compression of infalling intergalactic material onto the virializing cluster and are large, coherent, and localized. These last shocks are energetically less important but can accelerate CRs to Ultra High Energies (Ryu et al. 2003). In principle, merger shocks can accelerate electrons to produce large scale synchrotron radio emission, but the radiative life-time of the electrons diffusing away from the shocks is so short that they would just be able to produce relics and not radio halos (Miniati et al. 2001). In addition, the low Mach number of the shocks produced even

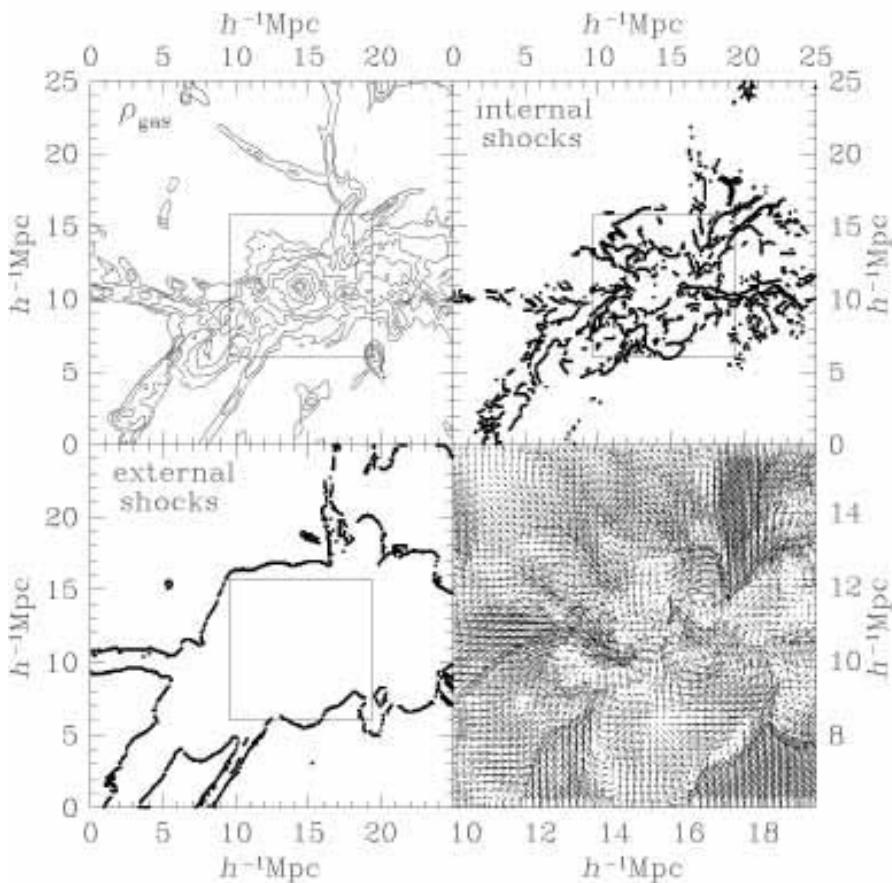


Figure 6.5. The 2-D slice of $(25h^{-1} \text{ Mpc})^2$ around a simulated $T \sim 3.3 \text{ keV}$ cluster showing gas density, shock distribution and velocity field (from Ryu et al 2003).

during the major mergers is too low to produce non-thermal emission with the observed spectra (Berrington & Dermer 2003). Re-acceleration of a pop-

ulation of relic electrons by turbulence powered by major mergers has been proposed to explain the radio halo emission features (Brunetti et al. 2001). The possibilities to channel the energy of the turbulence in particle acceleration include magneto-sonic waves, magnetic Landau damping, lower hybrid waves and Alfvén waves (see, e.g., Brunetti et al. 2003). The possibility for turbulent re-acceleration to work is, however, to strongly limit the energy density of the (equally accelerated) CR protons to \lesssim a few % which, otherwise, would induce a strong damping of waves leaving insufficient energy for the electron acceleration. Such diffusive acceleration mechanisms have also to contend with the Chandra X-ray observations of stable cold fronts which limit the amount of turbulence in the ICM (Mushotzky 2003).

Hadronic models deal with secondary electrons (first proposed by Dennison 1980) produced by the decay of collision/annihilation products (mainly $\pi^\pm \rightarrow e^\pm$) of p-p (Colafrancesco & Blasi 1998, Blasi & Colafrancesco 1999) or $\chi\bar{\chi}$ (e.g., Colafrancesco & Mele 2001) interactions. The secondary electrons are produced *in situ* (thus avoiding to invoke re-acceleration) and require values $B_\mu \sim 1 - 10$, found to be more consistent with Faraday Rotation data. These models predict a substantial p/e⁻ ratio (like that observed in our Galaxy and in SNe remnants) and an extended modest gamma-ray emission with both hadronic and electronic signatures. The *in-situ* character of these models produce an extended emission both in space and in time (it is actually stationary) and a quite strong feedback on the ICM (Colafrancesco 1999, Miniati et al. 2001).

In this framework it has been recently suggested that cosmic rays can also be accelerated in the cluster atmosphere by GRBs cannonballs originating in the cluster galaxies (see Colafrancesco et al. 2004). This model has also the advantage to solve other puzzling problems in clusters like the cooling flow quenching, the origin of magnetic fields, and the lensing to X-ray mass mismatch.

A last possibility to produce relativistic electrons in the cluster atmospheres is due to the interaction of high-E gamma-rays with low energy (CMB/IR) photons, with the subsequent production of e^\pm pairs (Timokhin et al. 2003). The main uncertainty of this last model is the unknown origin of the very high-E gamma-rays in clusters.

In conclusion, the basic difference between these two classes of models is a difference between episodic (the primary) and stationary (the *in-situ*) models.

3. The astrophysics of cosmic rays in galaxy clusters

Cosmic rays residing in galaxy clusters produce several astrophysical signatures among which there are diffuse synchrotron radio emission, ICS of CMB (and other background) photons which are then moved to higher frequencies

(EUV, X-ray and gamma-ray), bremsstrahlung radiation at nearly the energy of the relativistic electrons, gamma-ray emission by $\pi^0 \rightarrow \gamma + \gamma$ in hadronic collisions and DM annihilations, non-thermal SZ effect, Coulomb heating. Thus, we can use, in principle, the available observations of non-thermal phenomena in cluster to probe the origin and the leading production mechanisms of the CRs in these structures. We will refer in the following, for the sake of illustration, to the specific case of the Coma cluster which is the system with the widest observational coverage so far.

The radio halo spectrum of Coma can be fitted out to $\nu_r \sim 1.4$ GHz with both primary and secondary models (e.g., Sarazin 1999, Blasi & Colafrancesco 1999, Colafrancesco & Mele 2001, Brunetti et al. 2001, Thierbach et al. 2003) and it does not provide a strong test bed. The claimed steepening of the spectrum at higher frequency can be attributed to the ageing of primarily accelerated electrons but can also reproduced in specific in-situ models (Thierbach et al. 2003).

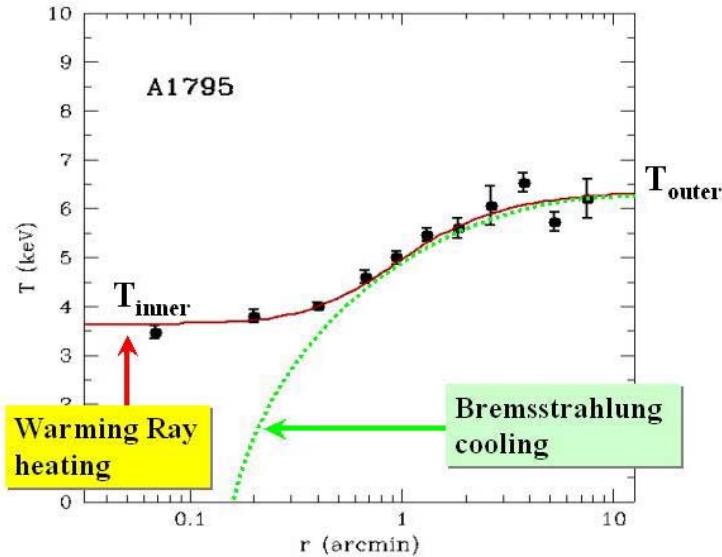


Figure 6.6. The radial temperature distribution of the cluster A1795 as fitted by the warming rays model (see Colafrancesco, Dar & DeRujula 2004 for details). The schematic representation of the T distribution due to bremsstrahlung cooling is shown as the dotted curve.

The HXR emission observed in Coma may be fitted with ICS of CMB photons off a non-thermal electron population with energies $E_e \gtrsim 3.3$ GeV whose spectrum should have a slope $x \sim 3.5 - 4$, but requires a large amount of relativistic electrons which overproduce gamma-rays by bremsstrahlung and exceeds the EGRET upper limit at $E > 100$ MeV (Colafrancesco 2004). Alternative explanations of the HXR excess in terms of bremsstrahlung emission from a suprathermal electron population (Sarazin & Kempner 2000, Blasi et al. 2000, Ensslin et al. 1999) seem to be excluded due to large energy injection required by the associated non-thermal bremsstrahlung emission (Petrosian 2001). A model in which the HXR emission in Coma is provided by synchrotron emission of high energy electrons produced in the interaction of very high-E gamma-rays (Timokhin et al. 2003) relies on the (unclear) possibility to have a specific source of very high-E gamma-rays in clusters. The HXR emission from clusters can indeed be fitted by a model in which CRs are injected in the IC plasma by GRBs-cannonballs acting in the many cluster galaxies (Colafrancesco, Dar & De Rujula 2003). In this model, the electrons knocked-on by the cannonball recoil with large kinetic energy and then, the scattered electrons cool and thermalize mainly via bremsstrahlung at high-E and by Coulomb collision at low-E. This model provides both a simple solution for warming up the cluster cores in cooling-flow clusters and for producing HXR tails in non cooling-flow clusters, as observed.

Both primary and secondary electron models (Atoyan & Voelk 2000, Brunetti et al. 2001, Blasi & Colafrancesco 1999, Miniati et al. 2001) have been analyzed to reproduce the spectral and spatial features of the EUV excess in Coma without a definite solution. Additional experimental information has been recently added to the complexity of the problem: in particular, the EUV intensity distribution seems to be highly correlated with the thermal X-ray intensity and produce a constant ratio between the azimuthally averaged EUV and X-ray intensities (Bowyer et al. 2004). Specific secondary models seem, at present, one of the few viable possibilities to reproduce the EUV emission features of Coma.

The next generation gamma-ray observatories (GLAST, VERITAS, MAGIC) have sensitivity and spectral resolution to disentangle, or at least constrain, the CR origin in galaxy clusters (see Fig.6.7). Gamma-ray observations offer, in fact, the most direct look at both the basic mechanism and at the site for the origin of CRs in galaxy clusters. Gamma-rays are, in fact, expected to be emitted from clusters due to a variety of mechanisms: $\pi^0 \rightarrow \gamma + \gamma$ due to either proton-proton collisions or to DM particle (neutralino) annihilation; bremsstrahlung and ICS emission from both primary and secondary electrons accelerated and/or injected in the ICM. Interestingly, the emission spectra of these models are quite different at gamma-ray energies and this could allow to

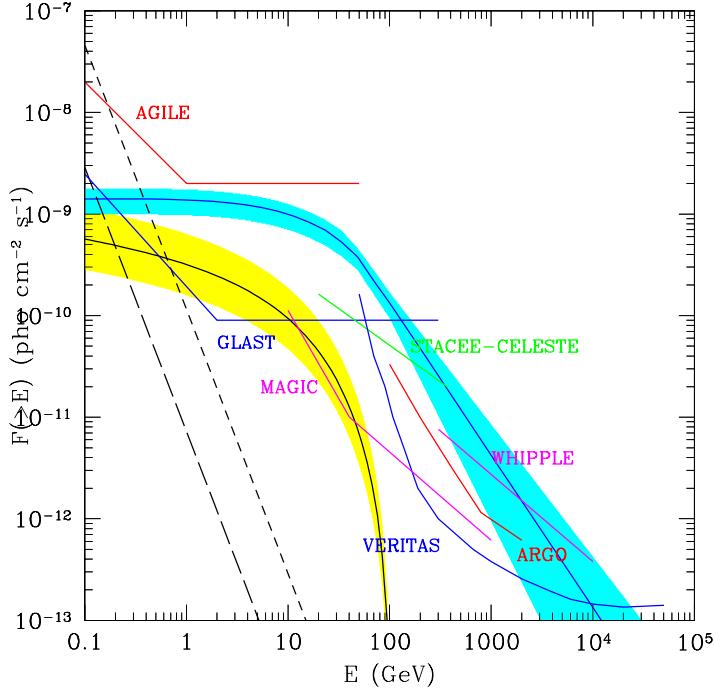


Figure 6.7. Predictions of the diffuse gamma-ray emission of Coma as expected in different models for the CR origin: bremsstrahlung for $B_\mu = 0.3$ (short dashes) and 1 (long dashes); neutralino annihilation for $M_\chi = 100$ GeV (yellow area; an enhancement factor ~ 3 has been chosen) and p-p collision (blue area) (from ?).

shed a light onto the CR origin in Large Scale Structures (see Colafrancesco 2004 for a review).

In this context, it has been recently proposed (Colafrancesco et al. 2003) to use the non-thermal SZ effect in clusters (i.e., the Compton scattering of CMB photons off high-E, non-thermal electrons), which depends on the total pressure of the electronic population, to set constraints on the overall spectrum of the clusters CRs. The data on the radio-halo cluster A2163, for which wide frequency SZ observations are available, show that the relativistic particle pressure is $P_{rel} \lesssim 0.3 P_{thermal}$. As a consequence, the energy spectrum of the relativistic electrons should have a minimum-E break at ~ 50 MeV or flatten substantially at $E \lesssim 1$ GeV (see Colafrancesco et al. 2003 for details). In the case of Coma one finds that the pressure of the relativistic electrons cannot exceed $\sim 4\%$ of the thermal pressure. The study of non-thermal processes via the associated SZ effect is a promising technique in the light of the next gener-

ation, high-sensitivity radio and sub-mm experiments like, e.g., PLANCK and ALMA.

4. Conclusions

The nature and the origin of cosmic rays in galaxy clusters are not yet known. In addition, the origin of the (possibly associated) non-thermal processes is still a matter of debate. However, the evidence for the presence of CRs in the atmospheres of many galaxy clusters stimulate a large wealth of theoretical and observational activity. In this context, multi-wavelength astrophysics is the key to piece together the puzzle offered by the various astrophysical phenomena occurring in galaxy clusters. It is appealing, in these respects, to expect that the astrophysical features of galaxy clusters might give information on the fundamental properties of the relativistic particles diffusing in the cluster atmosphere.

References

- Atoyan, A. & Voelk, H. 2000, ApJ, 535, 45
- Berrington, R.C. & Dermer, C.D. 2003, ApJ, 594, 709
- Blasi, P. et al. 2000, ApJ, 535, L71
- Bonamente, M. et al. 2002, ApJ, 576, 688
- Bowyer, S. 2000, AAS, 32, 1707
- Bowyer, S. et al. 2004, ApJ submitted
- Brunetti, G. et al. 2001, MNRAS, 320, 365
- Brunetti, G. et al. 2003, astro-ph/0312482
- Carilli, C.L. & Taylor, G.B. 2002, ARA&A, 40, 319
- Colafrancesco, S. 1999, in "Diffuse Thermal and Relativistic Plasma in Galaxy Clusters", H. Boehringer, L. Feretti, P. Schuecker Eds., p.269 (astro-ph/9907329)
- Colafrancesco, S. & Mele, B. 2001, ApJ, 562, 24
- Colafrancesco, S. 2002, A&A, 396, 31
- Colafrancesco, S. et al. 2003a, A&A, 397, 27
- Colafrancesco, S., Dar, A. & De Rujula, A. 2004, A&A, 413, 441
- Colafrancesco, S. 2004, Journal of Nuclear Physics G, in press
- Dennison, B. 1980, ApJ, 239, L93
- Dolag, K. et al. 2001, A&A, 378, 777
- Durret, F. et al. 2002, A&A, 390, 397
- Eilek, J. 1999, in "Diffuse Thermal and Relativistic Plasma in Galaxy Clusters", H. Boehringer, L. Feretti, P. Schuecker Eds., p.
- Ensslin, T. & Biermann, P. 1998, A&A, 330, 90
- Ensslin, T. et al. 1999, A&A, 344, 409
- Feretti, L. 2003, preprint astro-ph/0301576
- Fusco-Femiano, R. et al. 2000, ApJ, 543, L7

- Fusco-Femiano, R. et al. 2004, ApJ, in press (astro-ph/0312625)
Giovannini, G. et al. 1999, New Astronomy, 4, 141
Gonçalves, D.R. & Friaça, A.C.S. 1999, MNRAS, 309, 651
Govoni, F. et al. 2001, A&A, 369, 441
Jaffe, W.J. 1977, ApJ, 212, 1
Kaastra, J. et al. 2002, A&A, 397, 445
Kempner, J. & Sarazin, C.L. 2001, ApJ, 548, 639
Liang, H. et al. 2000, ApJ, 544, 686
Lieu, R. et al. 1996, Science, 274, 1335
Lieu, R. et al. 1999, ApJ, 510, L25
Longair, M. 1993, ‘High-Energy Astrophysics’, (Cambridge Univ. Press: Cambridge)
Miniati, F. et al. 2001, ApJ, 526, 233
McNamara, B. 2003, astro-ph/0310708
Mushotzky, R. 2003, astro-ph/0311105
Nevalainen, J. et al. 2004, ApJ, in press
Petrosian, V. 2001, ApJ, 557, 560
Ryu, Dongsu et al. 2003, ApJ, 594, 627
Sarazin, C.L. & Kempner, J. 2000, ApJ, 533, 73
Sreekumar, P. et al. 1996, ApJ, 464, 628
Thierbach, M. et al. 2003, A&A, 397, 53
Timokhin, et al. 2003, astro-ph/0305149

Chapter 8

AN INTRODUCTION TO QUINTESSENCE

Alain Riazuelo

*Institut d'Astrophysique de Paris, 98bis boulevard Arago, 75014 Paris, France
riazuelo@iap.fr*

Introduction

A large number of observations seem to indicate that in addition to baryonic and dark matter, a large fraction of the universe energy density is under the form of some unknown form of matter with negative pressure. This form of matter, usually called “dark energy” may be a cosmological constant, but this explanation is not satisfactory for several reasons. Moreover, the exact nature of this dark energy is unknown: what observation suggest is that (i) dark energy is likely to be the dominant form of matter in the universe¹ ($\Omega_{\text{DE}} \sim 0.7$), (ii) the expansion of the universe is probably accelerating today, so that the pressure-to-density ratio w_{DE} is smaller than $-\frac{1}{3}$, and (iii) it is unclustered. The idea of quintessence is to explain this dark energy with a scalar field which would behave closely to a cosmological constant today, just as the inflaton behaved as a cosmological constant at very early times. After a short overview of the difficulties to explain the dark energy by a cosmological constant (Sec. 8.1), we shall explain the main characteristic we need for a scalar field to behave as dark energy today (Secs. 8.2-8.3). We shall then give a few considerations about this model from the point of view of high energy physics (Sec. 8.4). Then the cosmological consequences of such a scalar field will be briefly discussed in Sec. 8.5, and we shall conclude by giving the observational status and prospect of dark energy (Sec. 8.6).

1. The two cosmological constant problems

There are two problems which appear when one considers the possibility of a small but non zero cosmological constant. The first one has to do with the

¹ See however Alain Blanchard's contribution for another point of view.

smallness of the cosmological constant: if we assume that $\Omega_{\text{DE}} \sim 0.7$ and that dark energy is a cosmological constant (the exact value of Ω_Λ does not matter here), then, in term of Planck units, one has

$$\rho_\Lambda \sim 10^{-122} \rho_{\text{Pl}}. \quad (8.1)$$

The biggest mystery related to the cosmological constant lies there: indeed, from very general quantum field theory considerations, any particle with momentum \mathbf{k} should have a contribution $E \sim \hbar\omega$ to the total energy even in its lowest energy state, so that one expects that any species should contribute to the vacuum energy density according to

$$\rho_{\text{vac}} \sim \int \frac{1}{2} \hbar\omega d\mathbf{k}. \quad (8.2)$$

This calculation is in principle valid till the breakdown of quantum field theory itself, i.e. till Planck scale, so that one can put an ultraviolet cutoff at $k_{\text{Pl}} \sim 1/L_{\text{Pl}}$ in the above equation and the result is

$$\rho_{\text{vac}} \sim k_{\text{Pl}}^4 = \rho_{\text{Pl}}. \quad (8.3)$$

As stated above, this is 122 orders of magnitude above what is observed. We therefore face the necessity to find some mechanisms which cancel almost all these contributions (in one part in 10^{122} ...). Naively, this can be achieved in the framework of supersymmetry, which assumes that there is an extra symmetry between known particles and their “superpartners”, because in this case, the signs of the contribution of the ordinary particles and their superpartners are opposite and their respective contributions cancel. However, no supersymmetric has ever been observed so far, so that one concludes that supersymmetry must be broken at some energy scale above the highest energies that can be reached in accelerators (a few 100 GeV). In this case, the cancellation mechanism should work only at energies higher than the supersymmetry breaking scale, so that we are left with a contribution of order

$$\rho_{\text{vac}} \sim k_{\text{SB}}^4 > 10^8 \text{ GeV}^4 > 10^{54} \rho_\Lambda. \quad (8.4)$$

Although this represents a very significant improvement with respect to the initial problem, one is still very far from the observed order of magnitude of vacuum energy. So far, no satisfactory solution to this problem is known, so that one is left with the assumption that there is some unknown mechanism that exactly cancels any contribution to the energy density.

The second cosmological problem now comes from the observed existence of some form of energy with negative pressure. If we assume that most of the contributions to the vacuum energy cancel each other, we still face the problem of why this cancellation is so efficient but not perfect, and, incidentally,

why it is such that it roughly corresponds to today's energy density. Since no known energy scale seems to be related to the observed dark energy density, it is difficult to find a natural explanation for this specific amount. The aim of quintessence is to address this second problem (only), by supposing that there is a dynamical origin for the dark energy, today's energy density being a consequence of the evolution of some fields.

2. A scalar field as dark energy

As we have seen during the course on inflation, a scalar field can behave as a cosmological constant when its kinetic term becomes negligible in front of its potential term. However, the features of the scalar field we are interested in differ significantly from an inflationary scalar field: in the former case, we want a field that is negligible at early times and which dominates afterwards, whereas in the latter case, it is the contrary. Historically, the first scalar field dark energy model was aimed to address the possibility to have some components with a constant equation of state parameter w other than 0 (matter), $1/3$ (radiation), $-1/3$ (curvature) and -1 (cosmological constant) (Ratra & Peebles 1988).

By definition of a homogeneous scalar field, one has the following relation between the kinetic and the potential terms:

$$\frac{1}{2}\dot{\phi}^2 = \frac{1+w_Q}{1-w_Q}a^2V. \quad (8.5)$$

Differentiating gives, in the hypothesis of constant w_Q ,

$$H\dot{\phi} \propto \frac{1+w_Q}{1-w_Q}a^2V'. \quad (8.6)$$

Differentiating again gives

$$f(H^2, \dot{H}) \propto \frac{1+w_Q}{1-w_Q}a^2V''. \quad (8.7)$$

Assuming that one is in an era where the background equation of state parameter w_B is constant, one can derive a relation between \dot{H} and H^2 with the help of the Friedmann equations, so that the three relations above can be translated into

$$\frac{VV''}{V'^2} = \text{Const.} \quad (8.8)$$

This equation admits the simple solutions

$$V \propto \phi^\alpha, \quad 1+w_Q = \frac{\alpha}{\alpha+2}(1+w_B), \quad (8.9)$$

$$V \propto \exp(-\lambda\phi), \quad w_Q = w_B, \quad (8.10)$$

the latter case being the limit of the former one when $\alpha \rightarrow \infty$. This study is not complete. In order this solution to be interesting, we have to study its stability with respect to small departures from this solution. This is to this subject that we now turn.

3. Stability of the $w_Q = \text{Const}$ regime

Let us study the stability of the solution $w_Q = \text{Const}$ found above. To do so, we perturb the Klein-Gordon equation around this solution. Here, we need only to consider the case $\phi = \phi_{w_Q=\text{Const}} + \delta\phi$, where $\delta\phi$ is a function of time only. After simple manipulations, we find

$$\ddot{\delta\phi} + 3H\dot{\delta\phi} + a^2V''\delta\phi = 0, \quad (8.11)$$

which can be rewritten expliciting the function f in Eq. (8.7),

$$\ddot{\delta\phi} + 3H\dot{\delta\phi} + \frac{1-w_Q}{1+w_Q}\alpha(\alpha+1)\frac{1}{2\eta^2}\left(\frac{6(1+w_B)}{(\alpha+2)(1+3w_B)}\right)^2\delta\phi = 0. \quad (8.12)$$

Stability is insured when the exponents of power law solutions to this equation all have a negative real part. This is the case as soon as

$$w_B \geq -\frac{1}{3}, \quad (8.13)$$

$$\alpha > 0. \quad (8.14)$$

Therefore, as soon as the exponent of the potential is positive, the solution is stable when the matter content is dominated by “normal” forms of matter (baryons, CDM, radiation and possibly curvature). When the field has reached its w_Q solution it is said to be in the “tracking” regime. All the parameters of the field are then fixed. The stability of this tracking regime ensures that regardless of the initial conditions of the field, it will someday reach the tracking solution, so that we do not have to care about the initial state of the field. This represents one of the most significant improvement of quintessence models with respect to models with a cosmological constant.

From Eq. (8.9), it is clear that $w_Q < w_B$, so that for the special solution $w_Q = \text{Const}$, the field starts with an negligible energy density which subsequently decays less fast than that of the background, so that ultimately the quintessence field will dominate the energy density of the universe. Since we have today $\Omega_{\text{DE}} \sim \Omega_{\text{mat}}$, today’s epoch represents the transition between the matter dominated era and the quintessence dominated era (as observations favor an Ω_Q only slightly larger than Ω_{mat}).

So far, we studied only the case where the quintessence field was subdominant. The subsequent evolution of the scalar field (when it becomes the main form of energy) is slightly more complicated to compute, but one can show

that as the energy density of the scalar field reaches that of the background, the field equation of state parameter slowly decays toward -1 so that as the field dominates it behaves more and more closely to a cosmological constant (even though the field never stops because there is no minimum to the potential, the kinetic term decays faster than the potential term so that w_Q reaches -1).

An example of the temporal evolution of a quintessence field is given in Fig.8.1.

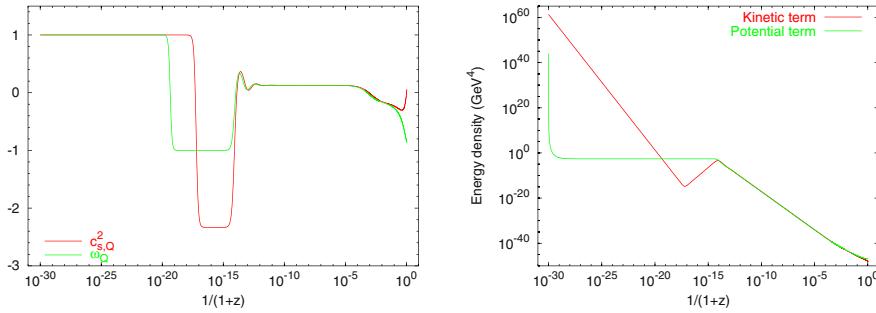


Figure 8.1. Evolution of a quintessence field between $z = 10^{32}$ and now. Left figure shows the equation of state parameter w_Q as a function of the redshift, whereas right figure shows the evolution of the kinetic and potential term. The tracking solution is reached at $z \sim 10^{14}$ and the parameter w_Q is then constant. It varies in fact around $z \sim 10^3$ as one enters into the matter dominated era, whose corresponding w_Q differs from that of the radiation era. The initial energy density of the field has been chosen so that it is larger than that of the tracking solution. Because of this, the field tries to “rejoin” the tracking solution by decreasing its energy density as fast as possible (with a kinetic-dominated regime with $w_Q = 1$), whereas its potential energy remains constant after a very fast initial decrease (because of the large kinetic term). When the kinetic term drops below the potential term, w_Q goes from 1 to -1 till the still constant potential energy reaches the value of the tracking solution, which happens at $z \sim 10^{14}$ here. The tracking regime ends when the field energy density reaches that of the background, which happens at late times here $z \sim 1$, and where the field subsequently behaves as a cosmological constant with $w_Q \rightarrow -1$. The exact way w_Q goes toward -1 is the main way one can distinguish between quintessence models.

4. Model building

Focusing on the inverse power law potential $V = M^{4+\alpha}/\phi^\alpha$, let us try to determine the values of the two parameters M and α . At the epoch of transition between matter and quintessence, Eq. (8.7) is still marginally valid, so that the normalization M of the potential is such that

$$\frac{M^{4+\alpha}}{\phi_0^{\alpha+2}} \sim H_0^2, \quad (8.15)$$

or

$$V_0'' \sim H_0^2 \sim (10^{-61} M_{\text{Pl}})^2. \quad (8.16)$$

(The subscript 0 meaning that we consider the corresponding quantities today.) On the other hand, the first Friedmann equation $H^2 = \kappa\rho$ leads to

$$H_0^2 \sim \kappa \frac{M^{4+\alpha}}{\phi^\alpha} \sim \frac{M^{4+\alpha}}{M_{\text{Pl}}^2 \phi^\alpha}, \quad (8.17)$$

from which we deduce that the transition between the matter era and the quintessence era occurs when

$$\phi_0 \sim M_{\text{Pl}} \quad (8.18)$$

(as for inflation...), from which one deduces the energy scale M

$$M \sim \left(M_{\text{Pl}}^\alpha \frac{H_0^2}{\kappa} \right)^{\frac{1}{4+\alpha}} \sim (M_{\text{Pl}}^\alpha \rho_c)^{\frac{1}{4+\alpha}} \sim M_{\text{Pl}} \times 10^{\frac{-122}{4+\alpha}}. \quad (8.19)$$

This allows to have a relation between the energy scale M and the exponent α . If we want to have an energy scale larger than 10 TeV, then we have to choose $\alpha \geq 4$. With larger exponents, the energy scale is even larger.

A second relation involving α can in principle be derived when we know the current value of the dark energy equation of state parameter w_{DE} . Indeed, since the potential does not possess a local minimum, the field never stops, so that it never behaves exactly as a cosmological constant. Moreover, even if its equation of state parameter w decays (without ever reaching) toward -1 , the rate at which this transition occurs depends on the steepness of the potential: the steeper the potential, the slowest w goes toward -1 . In particular, one finds, at the epoch $\Omega_Q \sim 0.7$,

$$\alpha = 11 \Rightarrow w_Q^0 = -0.29, \quad (8.20)$$

$$\alpha = 2 \Rightarrow w_Q^0 = -0.63. \quad (8.21)$$

Therefore, it is difficult to reconcile a realistic model (with high M and therefore large α), with a situation where the field would mimic closely a cosmological constant. Therefore (anticipating on the fact that observation seem to favor a low value of w_{DE}), either we need a low exponent α and a fine tuning on the mass scale M , or we have to modify the shape of the potential. An example of such a modification is motivated by supergravity considerations which are far beyond the scope of this introduction to quintessence. Its predicts that as the field reaches Planck scale, and regardless the amplitude of the potential $V(\phi)$, supergravity corrections modify the shape of the potential into

$$V(\phi) = \frac{M^{4+\alpha}}{\phi^\alpha} \exp \frac{\kappa\phi^2}{2}. \quad (8.22)$$

In this case, regardless of the value of α , the value of the equation of state parameter w is always around -0.8 when $\Omega_Q = 0.7$. Let us note that this is far from the end of the story. In order to build a realistic model, we have in particular to (i) find a framework which can “naturally” account for an inverse power law potential, (ii) explain why the particle associated to this field have never been observed: since V'' is extremely low today, one expects that the associated particles are very light, so that if they are not detected experimentally, then they must be extremely weakly coupled to ordinary matter, a situation which may necessitate some new fine tuning in the model. For a much deeper discussion about all this, see for examples Refs. Brax & Martin 1999; Brax et al. 2000; Brax et al. 2001 and references therein.

5. Dark energy and structure formation

As compared to a cosmological constant, quintessence modifies the late time evolution of the expansion rate of the universe. It can therefore affect the luminosity distance of supernovae as well as the angular distance of CMB patterns. However the most dramatic difference between a quintessence models and the now standard Λ CDM scenario comes when one considers structure formation.

The reason is as follows: if we suppose that quintessence had a constant w_Q , then the redshift at which the transition occurred is given by

$$z_Q^{\text{mat}} = -1 + \left(\frac{\Omega_Q^0}{\Omega_{\text{mat}}^0} \right)^{\frac{-1}{3w_Q}}. \quad (8.23)$$

This redshift can be quite high if w_Q is close to 0: it varies between 0.33 and 3.1 when w_Q goes from -1 to -0.2 . After the matter-to-quintessence transition, the Jeans instability is no longer efficient. Therefore, structure can no longer grow and the power spectrum is reduced by a factor

$$\left(\frac{1 + z_Q^{\text{mat}}}{1 + z_\Lambda^{\text{mat}}} \right)^2 = \left(\frac{\Omega_Q^0}{\Omega_{\text{mat}}^0} \right)^{\frac{2}{3}\left(1 + \frac{1}{w_Q}\right)}, \quad (8.24)$$

as compared to a Λ CDM model.

In the hypothesis of an exponential potential, the situation is even worse because Jeans instability is always reduced by the presence of quintessence (i.e., Ω_{mat}^0 is never equal to 1). More detailed calculation give the following value of the matter power spectrum normalization σ_8 as a function of w_Q in Fig. 8.2.

6. Observational status

The most direct probe of quintessence lies in supernovae data. The reason, as we said, is that supernovae data (in the hypothesis that supernovae are

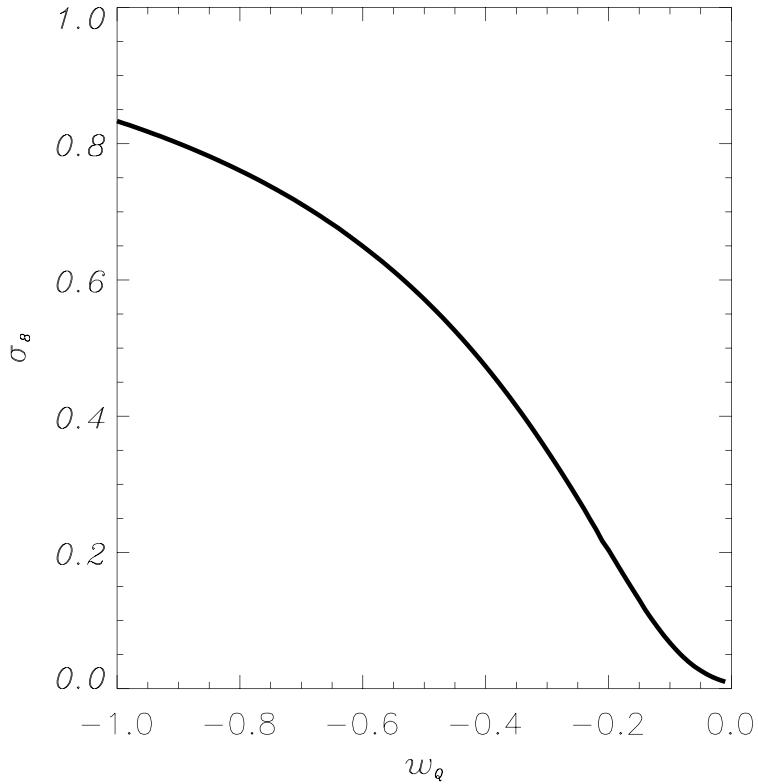


Figure 8.2. σ_8 as a function of w_Q in a fiducial QCDM model. Cosmological perturbations are normalized to COBE data, so that the initial amplitude of cosmological perturbations is almost constant in all these models. σ_8 therefore reflects the fact the low $|w_Q|$ dramatically reduce the total growth of cosmological perturbations.

truly standard candles) directly probe the expansion history of the universe $H(z)$, which, through the Friedmann equations, probe $\rho_Q(z)$. Although very powerful in principle, this method is at present not very sensitive to the possible variations of $w_Q(z)$ because of systematic uncertainties in supernovae measurements, as well as a yet small number of observed objects. On the contrary, structure formation is presently more sensitive to quintessence models, although this situation might change with high precision supernovae measurements such as with the SNAP satellite experiment. As often in cosmology the most interesting constraints do not come from a single observational test, but from the combination of several. This is summed up in Fig. 8.3 below. Present data (pre-WMAP and including WMAP) can give narrow constraints

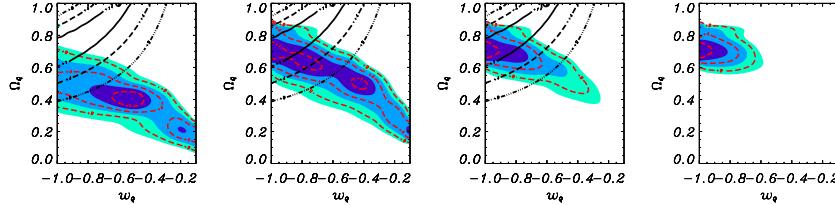


Figure 8.3. Constraints on Ω_Q and w_Q in quintessence models with constant w_Q using various astrophysical tests. Constraints coming from supernovae alone are the black contours. The two leftmost figures show confidence contours combining CMB and various constraints on σ_8 found in the literature (?; Pierpaoli et al. 2001 and Hoekstra et al. 2002; Seljak 2001; Viana et al. 2002; Reiprich & Böhringer et al. 2002). The first constraint (with high σ_8) appears to be in disagreement with supernovae data (the former favors low Λ / quintessence, whereas the latter favors high Λ / quintessence). The third figure combines CMB, low σ_8 constraints and constraints on H_0 . Finally, the last one adds the constraints coming from supernovae and give yet very stringent constraints on quintessence scenarios. Taken from Ref. [Douspis et al. 2003].

on quintessence models. The most striking features is that combining all the available datasets favors models with w_Q smaller than -0.7 . Although value smaller than -1 are even more favored they would correspond to very unusual models (for example with negative valued potentials or non standard kinetic terms). As we already said, this significantly narrows the space of available quintessence models, and it is likely that a pure power law potential is not the most relevant answer to dark energy issues. At present it is however still difficult that data favor without ambiguity the Λ CDM model or a quintessence model with $w_Q > -1$. In particular, more precise data on σ_8 are crucial in determining a precise value of w_Q as the contours coming from CMB + σ_8 and those from supernovae are almost orthogonal.

References

- A. Blanchard, R. Sadat, J. Bartlett & M. Le Dour 2000, astro-ph/9908037, A&A, 362, 809
- B. Ratra & P.J.E. Peebles 1988, Phys. Rev. D 37, 3406.
- P. Brax & J. Martin 1999, Phys. Lett. B468, 40
- P. Brax, J. Martin & A. Riazuelo 2000, Phys. Rev. D 62 103505
- P. Brax, J. Martin & A. Riazuelo 2001, Phys. Rev. D 64 083505.
- E. Pierpaoli, D. Scott and M. White 2001, MNRAS 325, 77
- H. Hoekstra, H. Yee & M. Gladders 2002, ApJ 577, 595
- U. Seljak 2001, MNRAS 337, 769
- P. T. P. Viana, R. C. Nichol & A. R. Liddle 2002, ApJ 569, L75
- T. H. Reiprich & H. Böhringer 2002, ApJ 567, 716

M. Douspis, A. Riazuelo, Y. Zolnierowski & A. Blanchard 2003, A&A 405,
409

Chapter 7

AN INTRODUCTORY OVERVIEW ABOUT COSMOLOGICAL INFLATION

Alain Riazuelo

Institut d’Astrophysique de Paris, 98bis boulevard Arago, 75014 Paris, France

riazuelo@iap.fr

1. Introduction

In the late seventies, the history of the early Universe was described with the help of the hot Big-Bang scenario: the universe originated from an initial singularity and had then expanded, being filled by radiation and subsequently by non relativistic matter (baryon and Dark Matter).

This scenario provided a convincing explanation for the history of the Universe: it explained the abundance of light element synthesized during nucleosynthesis, and had predicted the existence of the cosmic microwave background. However, a number of questions remained unanswered, and it was unanimously accepted that in its present form, the hot Big-Bang scenario faced a number of very serious problems: the horizon problem (why is the universe so homogeneous and isotropic on large scales?), the flatness problem (why is its curvature so small?), the monopole problem (why aren’t there any heavy relic particles?) and the structure formation problem (which phenomenon is at the origin of the density fluctuations which gave birth to the structures we see today?).

In a sense, all these problems had to do with some choice of “initial conditions” for the Universe. In particular, the Universe had to be supposed “initially” homogeneous and isotropic. These features had to be put by hand in order the cosmological scenario to work properly, a situation which was of course very uncomfortable.

In its original formulation, inflation provided an explanation for the homogeneity and isotropy of the Universe (Guth 1981). It also explained why no heavy relics were observed in the Universe. Moreover, it was soon realized that for some unexpected reason, it could seed the density perturbations which gave birth to all the structures we observe in the Universe (Vilenkin 1982; Linde

1982; Starobinsky 1982). Therefore, even though there is no explicit proof for such scenario, inflation has rapidly become a very popular scenario¹

So far, inflation represent the simplest (and maybe the most natural) way to solve all above mentioned problems. More importantly, it also proposes a framework which allows to address them and is at present the most widely accepted scenario for the history of the early Universe. Of course, this does not mean that this represents the ultimate step toward our description of the early universe, but most of the alternative to inflation fit in a similar framework: even though some alternative scenarios rely on very different physical processes, in practice, the techniques one uses to compute the predictions from all these scenarios do not differ much from each others. Therefore, the following sections must be seen as a general overview of the current framework as well as a set of recipes which are extremely useful to address some issues about the early Universe, rather than an unambiguously well-established presentation about how things really happened.

We shall first review the (well-known) problems of the hot Big-Bang scenario in the next section. Then we shall do a presentation of the inflationary mechanism, where we shall also introduce some important quantities: the slow roll parameters (Section 7.3). In order to understand properly how inflation can seed density perturbations in the Universe, we shall then make an introduction to the problem of density perturbation in cosmology (Section 7.4). We shall then adapt this formalism to the inflationary situation where the Universe experiences a quasi-exponential expansion under the influence of a single scalar field (Section 7.5). The seeds for the cosmological perturbations (i.e. what we have to take as initial conditions when solving the perturbation equations are in fact the quantum fluctuations of this scalar field. We shall make a very brief introduction to this subject in Section 7.6. With all these tools we shall then compute the final spectrum (i.e., long after inflation) of the cosmological perturbations in Section 7.7.

We have found useful to give here a complete derivation of the most important results (the production of density perturbations). A knowledge of general relativity and quantum field theory is of course welcome, but not completely necessary. There is a large number of review articles about inflation which should be useful for the readers who would like to study this topic further. We recommend (non exhaustively) the book by Linde (focused on the high-energy physics side) (Linde 1990), and the book by Liddle and Lyth (more focused on CMB in general) (Liddle & Lyth) as well as the numerous references therein.

¹When talking about inflation, we prefer the use of the word “scenario” to the word “theory” as its aim is at present to provide a *description* of some part of the history of the Universe. Contrarily to other far better established physical theories, there is no experimental tests of inflation, although high precision measurements of the cosmic microwave background anisotropies might in a near future give a very convincing confirmation of the scenario.

French speaking readers might also benefit from the PhD thesis by J. Lesgourges (Lesgourges 1998) and D. Langlois (Langlois 1994), the latter being somewhat more technical than the former.

2. The hot Big-Bang scenario and its problems

Definitions and expansion dynamics

Observation suggests that we live in a homogeneous and isotropic patch of the Universe. The hypothesis of homogeneity and isotropy of the Universe translate into the fact that the line element can be written under the form

$$ds^2 = c^2 dt^2 - a^2(t) (\gamma_{ij} dx^i dx^j), \quad (7.1)$$

where an implicit summation on the indices i and j is assumed. Here, we have considered that the spacelike sections $t = \text{Const}$ are maximally symmetric spaces, i.e., that their metric γ_{ij} is either Euclidean, spherical or hyperbolic. This means, for example, that one can choose a set of (spherical) coordinates where one has

$$\gamma_{ij} dx^i dx^j = d\chi^2 + s_K^2(\chi) (d\theta^2 + \sin^2 \theta d\varphi^2), \quad (7.2)$$

with

$$s_K(\chi) = \chi \quad \text{if } K = 0 \quad (7.3)$$

$$s_K(\chi) = \frac{\sin(\sqrt{K}\chi)}{\sqrt{K}} \quad \text{if } K > 0 \quad (7.4)$$

$$s_K(\chi) = \frac{\sinh(\sqrt{|K|}\chi)}{\sqrt{|K|}} \quad \text{if } K < 0 \quad (7.5)$$

depending on the value of the curvature K .

The expansion of the Universe is encoded into the time dependence of the scale factor $a(t)$. The quantity t corresponds to the cosmic time. One can easily show that observers situated at $x^i = \text{Const}$ follow geodesics, therefore cosmic time corresponds to the proper time of such observers. It is assumed that any matter element is at rest or almost at rest with respect to the spatial part of the coordinate system. For this reason, the x^i are called comoving coordinates, and any distance measured using line element $\gamma_{ij} dx^i dx^j$ is called comoving distance.

It is very convenient to introduce the conformal time, η , defined as

$$ad\eta \equiv cdt. \quad (7.6)$$

(From now on, we will set $c = 1$.) The line element can then be rewritten

$$ds^2 = a^2(t) (d\eta^2 - \gamma_{ij} dx^i dx^j). \quad (7.7)$$

The interpretation of this quantity is the following: consider a photon, or any massless particle, which travels along the χ direction. By definition the line element for such a particle is null, therefore one has $|d\eta| = |d\chi|$. Since there is a correspondence between η and t through Eq. (7.6), $|\Delta\eta|$ can therefore be seen as the maximal distance that can be traveled by any particle between time $t_1(\eta)$ and $t_2(\eta + \Delta\eta)$. When η varies from $-\infty$ to some finite value today, any observer has been in causal contact with any observer in the past. On the contrary, if η has varied between two finite values, then only a fine region of the Universe has been in contact with the observer. Equivalently if asymptotically η goes to infinity, an observer can send signal to any observer, whereas if η reaches a finite value in the future, only a finite set of observer can receive signals from an observer. In particular, if $a \propto t^\alpha$, then one has $a \propto \eta^{\frac{\alpha}{1-\alpha}}$, and if $\alpha < 1$ (decelerating expansion), η varies between 0 and $+\infty$, if $\alpha > 1$ (accelerating expansion), η varies between $-\infty$ and 0.

The matter content of the Universe is described as an ensemble of perfect fluids of density ρ_f and pressure P_f . An important quantity is the “equation of state parameter” w_f defined as

$$w_f \equiv \frac{P_f}{\rho_f}. \quad (7.8)$$

For non relativistic matter, one has $w_{\text{mat}} = 0$, for radiation one has $w_{\text{rad}} = \frac{1}{3}$.

The evolution of the matter density is computed through the so-called “conservation equations”: the stress-energy tensor of the matter species have components (in the (t, x^i) coordinates):

$$T_{00} = \rho_f, \quad (7.9)$$

$$T_{0i} = 0, \quad (7.10)$$

$$T_{ij} = a^2 P \gamma_{ij}. \quad (7.11)$$

The “conservation equation” is written²

$$D_\mu T^{\mu\nu} = 0, \quad (7.12)$$

where D_μ stands for the covariant derivative. This translates into:

$$\partial_t \rho_f + 3H(P_f + \rho_f) = 0, \quad (7.13)$$

where we have introduced the Hubble parameter H , which is the expansion rate of the scale factor:

$$H \equiv \frac{da}{adt}. \quad (7.14)$$

²Although in Minkowski space, this corresponds to energy and momentum conservation, this is no longer the case in an expanding universe.

The conservation equation can equally be written using the conformal time η :

$$\dot{\rho}_f + 3H(P_f + \rho_f) = 0, \quad (7.15)$$

where a dot denotes a derivative with respect to the conformal time and where we have introduced the “conformal Hubble parameter” H :

$$H \equiv \frac{da}{a d\eta} = \frac{aH}{c}. \quad (7.16)$$

These two equations can be rewritten as

$$\rho_f \propto a^{-3(1+w_f)}. \quad (7.17)$$

For non relativistic species, the density decreases as a^{-3} , just because of the increase of a physical volume as a^3 due to the expansion. For radiation, the density decreases as a^{-4} because in addition to the decrease of the particle density, each particle experiences a redshift due to the expansion (its wavelength grows as a), and therefore the energy of an individual particle decreases as a^{-1} .

The evolution of the scale factor is then computed using the Einstein equations, $G_{\mu\nu} = 8\pi GT_{\mu\nu}$, where $G_{\mu\nu}$ is the Einstein tensor and G is Newton’s constant. These equations write

$$3 \left(H^2 + \frac{K}{a^2} \right) = 8\pi G \sum_f \rho_f, \quad (7.18)$$

$$- \left(2\partial_t H + 4H^2 + \frac{K}{a^2} \right) = 8\pi G \sum_f P_f. \quad (7.19)$$

In term of the conformal time (which we will use from now), this gives

$$3 \left(H^2 + K \right) = 8\pi G a^2 \sum_f \rho_f, \quad (7.20)$$

$$- \left(2\dot{H} + H^2 + K \right) = 8\pi G a^2 \sum_f P_f. \quad (7.21)$$

Defining (formally) some density and pressure associated to the curvature by

$$\rho_K \equiv \frac{-3K}{8\pi G a^2}, \quad (7.22)$$

$$P_K \equiv -\frac{1}{3}\rho_K, \quad (7.23)$$

we see that curvature can formally be considered as a fluid of equation of state parameter $w_K = -\frac{1}{3}$. Using the first Friedmann equation, we obtain

$$a \propto t^{\frac{2}{3+3w}} \propto \eta^{\frac{2}{1+3w}}, \quad (7.24)$$

where w represent the average equation of state parameter $\sum_f P_f / \sum_f \rho_f$. In the standard case, matter is composed of non relativistic matter, radiation and possibly curvature. Therefore, one has $-\frac{1}{3} \leq w \leq \frac{1}{3}$. This means than that the scale factor grows as some positive power of the conformal time η . This also means that $\partial_t^2 a < 0$, or that the expansion is decelerating. This is quite intuitive: if one considers a ball of expanding fluid in a Newtonian framework, one expects that the expansion of the fluid will slow down with time because of the influence of gravity³.

Horizon problem

The horizon problem can be re-expressed as follows: from Eq. (7.24), one has

$$\eta \propto t^{\frac{1+3w}{3+3w}}. \quad (7.25)$$

As long as one is in a situation where $\frac{1+3w}{3+3w}$ is positive, the initial singularity happens at time $t = 0$, $\eta = 0$, and the distance a photon can travel since the Big-Bang (this distance is called the horizon for obvious reasons) is simply

$$d \propto t^{\frac{1+3w}{3+3w}}. \quad (7.26)$$

This means that an observer who would begin to receive signal from time $t_i = 0$ could receive signal from further and further region as t increases. As we already said, it is convenient to use comoving coordinates because the fact that d increases expresses the fact that the number of (comoving) objects we can communicate with increases with time. Equivalently, this means that one *discovers* a Universe one larger and larger scales. This can be rephrased in the following way in term of physical rather than comoving quantities: a *physical* length $D = ad$ grows as the scale factor a , whereas the size of the observable universe grows as the inverse of the Hubble parameter $H^{-1} \propto t$. Therefore, as the expansion is decelerating, any physical length is initially “outside” the Hubble radius ($H^{-1} < ad$) and then enters into the Hubble radius ($ad < H^{-1}$). This is exactly what would happen in a purely Minkowskian space if one would allow observers to exchange signal only after some given time. As a consequence, two objects separated by a given distance $\Delta\eta$ were not in causal contact before some time given by Eq. (7.26). The horizon problem simply states this fact: we see that the Universe is homogeneous and isotropic

³A word of caution is in order here: in a Newtonian framework, pressure acts as a repulsive force. In general relativity, this is not the case. Pressure represents some form of (microscopic) kinetic energy, and therefore is an attractive force as any form of energy density; this is the reason why, for example, a high angular velocity cannot prevent a neutron star from collapsing when it is spinning very fast. When performing a careful analysis, one can show that in the Newtonian interpretation of the Friedmann equations, the quantity that matters is not ρ , but $\rho + 3P$. This is why as long as $w > -\frac{1}{3}$, gravity acts as an attractive force which slows down the expansion.

on large scales, but if the equation of state parameter w was always larger than $-\frac{1}{3}$ there is no causal process which can explain this fact.

Flatness problem

The flatness problems comes from similar considerations. The Einstein equations cannot of course predict the value of the curvature K , or of the curvature radius $R_c = a/\sqrt{|K|}$. What is known observationally is that the curvature radius is very large: $H^{-1} > a/\sqrt{|K|}$. Using Eq. (7.24), this can be translated at any epoch (subscript 0 denotes today):

$$\left(\frac{t}{t_0}\right)^{\frac{1+3w}{3+3w}} \gg R_c^0 H_0. \quad (7.27)$$

In other word, as any physical length, the curvature radius grows as a , that is more slowly than the Hubble radius. In order for the curvature radius to be larger than the Hubble radius today it has to be immensely larger than the Hubble radius at earlier times. Again, this is unexpected. One would for example prefer that at Planck epoch, both the curvature radius and the Hubble radius were of same order of magnitude.

Unwanted relic problem

Contrarily to the flatness and horizon problem which have a very similar origin, the unwanted relic problem comes from a very different cause. During the expansion, the Universe becomes colder because of the redshift that radiation experiences. It can happen that during this cooling phase, some phase transition occur, during which some stable objects called topological defects are created.

The simplest example can be given using a real field with a potential like

$$V(\phi) = \frac{1}{4}\lambda|\phi|^4 + \frac{1}{2}\mu(T^2 - T_c^2)|\phi|^2. \quad (7.28)$$

At high temperature ($T > T_c$), the potential has only one local minimum at $\phi = 0$. However, at low temperature, the potential has two local minima located at $\phi = \pm\sqrt{\frac{\mu}{\lambda}(T_c^2 - T^2)}$. When the temperature drops below T_c , the field will reach one of its two minima. If one considers two causally disconnected regions, it is clear that there is no reason that the field takes the same value in these two regions. As a consequence, one expects that there will be different regions where the field takes distinct values (since there are always causally disconnected region in a Universe that was always radiation dominated, see above). Between such regions, the field will reach 0 and have a higher (non zero) energy density. One therefore obtains a planar object known

as a domain wall. Such wall can be finite (like a bubble) or can have a very complicated infinite shape such as the surface of an infinite sponge. In the latter case, one has a topologically stable configuration. For this reason, such objects are called topological defects. They are created during phase transitions, which here amounts to say that the value of the field in each point of space is no longer invariant under some symmetry: as long as the potential had one minimum, the field was at $\phi = 0$ everywhere, and was therefore globally invariant under the transformation $\phi \rightarrow -\phi$, but this is no longer the case below T_c . The type of topological defects that can be created depends on the type of symmetry which is broken. Here one breaks a Z_2 (discrete) symmetry, which creates domain walls. If one had started from a complex field with the same potential, the field would initially have been invariant under any transformation like $\phi \rightarrow \phi e^{i\theta}$, that is under a U(1) transformation. After the phase transition, the field would try to take values of fixed module $|\phi| = \sqrt{\frac{\mu}{\lambda}(T_c^2 - T^2)}$ and arbitrary phase. If one considers a contour on which the phase of the field varies by 2π , then it is easy to see that there is at least one point in any surface bounded by this contour where the field has the value $\phi = 0$. One therefore obtains a linelike defect called a cosmic string. As for the domain walls, one can have finite closed loops, and infinite strings. Again, the latter are topologically stable and can contribute to the total energy density of the Universe since the energy density is usually negligible outside the string but is non zero inside.

If the field has three components, then it is easy to see that the objects that will be formed will be pointlike. Such objects are called monopoles. Although there are some (stable) monopole and anti-monopole configurations which can annihilate, this process is not very efficient as one does not expect these objects to have long range interactions (an annihilation between a monopole and an anti-monopole is in general extremely unlikely). The net result of this is that at some epoch such a phase transition will convert some amount of energy into these monopoles and anti-monopoles which then behave as non relativistic matter.

However, any estimation of the cosmological parameters gives some constraints on the ratio between matter and radiation. One has something like

$$\frac{\Omega_{\text{mat}}}{\Omega_{\text{rad}}} \sim 10^4. \quad (7.29)$$

Since $\rho_{\text{mat}} \propto a^{-3}$ and $\rho_{\text{rad}} \propto a^{-4}$, this means that no significant amount of monopole can have been produced before a redshift of $z \sim 10^4$, which corresponds to an epoch where the temperature was only a few electronvolts. On the other hand, any high energy physics theory predicts that the Universe must undergo a series of phase transitions at very high temperature, in particular at

GUT scale ($\sim 10^{15}$ eV). So again, the standard hot Big-Bang scenario fails to explain why these unwanted relics do not exist⁴.

Structure formation problem

Maybe the most annoying problem of the hot Big-Bang scenario is that it does not provide any explanation for the existence of structures in the Universe. It is well-known that structures can form through the Jeans instability only in a matter dominated era. However since the matter domination occurred quite “recently” in the history of the universe (around $z \sim 10^4$, see Eq. (7.29)), one is forced to suppose that small density fluctuations already existed before that epoch. Since no efficient process is known to form density perturbations in a radiation-dominated universe, so one has to suppose that the seeds for the astrophysical objects we observe were part of the initial condition of the whole scenario. As we shall now see, the biggest success of inflation is to provide a simple explanation for the presence of such density perturbations, in addition to solving quite naturally the other problems.

3. Inflation and inflationary dynamics

As we have seen above, the horizon and flatness problems can be solved if we suppose that at some epoch, the quantity $\frac{1+3w}{3+3w}$ has become negative. This is in particular the case if $-1 < w < -\frac{1}{3}$. The reason is that during such phase, the physical lengths still grow as the scale factor, but this time, they grow *faster* than the Hubble radius.

There is very simple form of matter which can exhibit such an equation of state parameter: a scalar field. The idea of inflation is based on the hypothesis that the Universe has been dominated at some early epoch by a scalar field whose equation of state parameter remained close to -1 for a while.

Scalar field

A scalar field is defined from the Lagrangian density L and the action $S = \int L d^4x$

$$S = \int \left(\frac{1}{2} D_\mu \phi D^\mu \phi - V(\phi) \right) \sqrt{|g|} d^4x. \quad (7.30)$$

⁴Note that we do not need to detect directly these monopoles to state this result. Should these monopoles exist, then their density should be very large and would have been detected. Therefore, we are sure that they are not present even if in practice a direct detection of these objects could be difficult.

The stress-energy tensor of such a field is computed by varying the action with respect to the metric,

$$\delta S = \int T_{\mu\nu} \delta g^{\mu\nu} \sqrt{|g|} d^4x. \quad (7.31)$$

In a FRW metric, one obtains (in (η, x^i) coordinates)

$$T_{00} = a^2 \rho_\phi = \frac{1}{2} \dot{\phi}^2 + a^2 V, \quad (7.32)$$

$$T_{0i} = 0, \quad (7.33)$$

$$T_{ij} = a^2 P_\phi \gamma_{ij} = \left(\frac{1}{2} \dot{\phi}^2 - a^2 V \right) \gamma_{ij}, \quad (7.34)$$

where we have supposed here that the field was homogeneous, so that the spatial gradients of the field are negligible.

The equation of state parameter w_ϕ is therefore given by

$$w_\phi = \frac{\frac{1}{2} \frac{1}{a^2} \dot{\phi}^2 - V}{\frac{1}{2} \frac{1}{a^2} \dot{\phi}^2 + V}. \quad (7.35)$$

We see that as long as the kinetic term $\frac{1}{2} \frac{1}{a^2} \dot{\phi}^2$ is negligible with respect to the potential term V , the field can behave like a cosmological constant and produce an accelerated phase of expansion.

Neglecting the gradient terms, the conservation equation writes

$$\ddot{\phi} + 2H\dot{\phi} + a^2 V' = 0, \quad (7.36)$$

where a prime denotes a derivative with respect to the field ϕ . This equation is also called the Klein-Gordon equation, and can be directly derived by varying the Lagrangian with respect to the field ϕ ($\delta S / \delta \phi = 0$).

Let us suppose that the field behaves like a cosmological constant for some amount of time. In this case, Eq.(7.24) tells us that the expansion is almost exponential ($H \sim \text{Const}$, $a \propto e^{Ht}$). Moreover, the relation between a and η is

$$\eta \propto \frac{1}{a}. \quad (7.37)$$

This means that the maximal comoving distance from which one can receive signal sent at some given time *decreases* with time. Intuitively, this can be understood by the fact that since the expansion is accelerating, the apparent recession velocity of other observers will increase with time and ultimately go beyond c , so that no signal can be received from them. In special relativity, this means in fact that a signal will become infinitely redshifted before being able to reach the observer.

In practice, the outline of the scenario is the following: at some stage, a region of the Universe becomes dominated by a scalar field (the inflaton) which is dominated by its potential term. If one has a small, causally connected, homogeneous and isotropic region, then inflation can start in this region which starts expanding exponentially (one can show that it suffices that the region is bigger than the Hubble radius for inflation to start there). This region can then become very large, much larger than the size of the observable universe. We are therefore left with a homogeneous region which is now much bigger than the Hubble radius. This explains why when the expansion will start decelerating one will discover a homogeneous universe on larger and larger scales. Note that during an almost exponential expansion epoch, the physical size of the observable universe $a|\eta|$ is constant, but the comoving size $|\eta|$ decreases. Therefore the number of objects one can see decreases with time (just because, as we said, the apparent recession velocity of other observers has become larger than c). At some stage, the inflationary phase stops. The Universe is then dominated by only one species, the scalar field which produced inflation, because its energy density remained almost constant during that phase, whereas the energy density of radiation (say) decreased as a^{-4} . One therefore has an “empty” and “cold” universe, in the sense that the particle density and the temperature of any other species than the inflaton has decreased dramatically. But the Universe still has a large energy density in the form of the scalar field. Usually, at the end of inflation, this inflaton decays into various particles, and its energy density is transferred into various particles which then thermalize. These complicated stages are known as the preheating and reheating. After that, one recovers the usual radiation-dominated Universe of the standard hot Big-Bang scenario. However, even though objects outside our observable surroundings are no longer causally connected with us, the fact that this formerly causally connected region was homogeneous survives to the inflationary phase. Therefore, the homogeneous region is now much larger than the size of the observable Universe, which naturally solves the horizon problem. Equivalently, the curvature radius grows enormously during inflation, so that it is natural that the curvature has become very small, even long after the end of inflation. Finally, since the density of any species has dramatically decreased during inflation, the monopole problem has a natural solution (provided of course that no phase transition producing monopoles or other unwanted defects occurs afterward).

The most useful quantity to describe the amount of inflation is the e -fold number N . It is defined as

$$N \equiv \ln \left(\frac{a_f}{a_i} \right), \quad (7.38)$$

where a_i and a_f represent the scale factor at the beginning and the end of inflation. If the expansion is exactly exponential, then N is the number of Hubble times that inflation lasts.

What is the necessary number of e -folds to solve the horizon and flatness problems? For the horizon problem, we want that the observable universe today was inside the Hubble radius at the beginning of inflation. Let us assume for definiteness than the Universe was matter dominated since $z_{\text{eq}} \sim 10^4$ till now then radiation dominated before. The relation between the comoving horizon today η_0 and the comoving horizon η_f at some early epoch deep in the radiation era is given by Eq. (7.24):

$$\eta_0 = \eta_f \sqrt{1 + z_{\text{eq}}} \frac{1 + z_f}{1 + z_{\text{eq}}}. \quad (7.39)$$

Let us suppose this early epoch η_f corresponds to a GUT-scale temperature era, $T_f \sim 10^{16}$ GeV, or $z_f \sim 10^{28}$. This gives $\eta_0 \sim 10^{26} \eta_f$: the size of a causally connected region is of course much larger than the comoving size of a causally connected region in the past. In order to solve the horizon problem, one wants that the our observable universe was inside a causally connected region in the past.

During inflation, the physical size of the observable universe is almost constant, $a_i \eta_i \sim a_f \eta_f$. However, between the beginning and the end of inflation, the physical distance of two comoving objects has grown of a factor a_f/a_i . Therefore the size of a region which was causally connected at the beginning of inflation is $\eta_i \frac{a_f}{a_i}$. In order to solve the horizon problem, one therefore need to impose that

$$\eta_0 < \eta_i \frac{a_f}{a_i}. \quad (7.40)$$

If we suppose (as it is reasonable from high-energy physics considerations) that inflation occurred as some GUT-scale epoch $T_f \sim 10^{16}$ GeV, then the minimum number of e -folds is given by

$$N > \ln \left(\frac{\eta_0}{\eta_i} \right) \sim 60. \quad (7.41)$$

This fairly large number explains why one wants inflation to be quasi-exponential: should it not be the case, then for typical potentials (like $V \propto \phi^2$ or $V \propto \phi^4$) it would not last long enough. Therefore, for inflation to proceed, we are now going to impose the conditions which are necessary (i) for inflation to occur (i.e., $w \sim -1$) and (ii) for it to last sufficiently long: $\dot{w}/(w+1) \ll H$. We will then study whether or not such conditions are likely or not.

Before that, let us note that it seems more appropriate that inflation does not end earlier than at GUT epoch because of the monopole problem (we prefer that inflation ends when the temperature is below that of any phase transition

that could produce a large amount of topological defects). It can however start at much higher energies: even though the equation of state parameter w_ϕ is almost equal to -1 , the energy density slowly decreases during inflation and if inflation lasts a sufficient amount of time (possibly much larger than 60 e -folds), then the ratio between the energy density at the beginning and at the end of inflation can reach several orders of magnitude.

Slow-roll conditions

In order inflation to proceed for a sufficiently large amount of time, we want the scalar field to behave like a cosmological constant. We therefore first impose that the is much closer $w = -1$ than $w = -\frac{1}{3}$ and that it is necessarily smaller than $-\frac{1}{3}$

$$1 + w \ll \frac{2}{3}, \quad (7.42)$$

which can be rewritten

$$\frac{\dot{\phi}^2}{a^2} \ll V. \quad (7.43)$$

Such a configuration can happen if we put for example a field on the slope of a $V \propto \phi^{2n}$ potential and that the field is initially at rest, so that the kinetic term is zero. However, the field will subsequently roll toward the minimum of its potential, and the kinetic term will increase as the potential term will increase. In order for inflation to last long, we also need to impose the the field accelerates slowly. This ensures that it will remain for long almost at rest on the slope of the potential. One way to characterize this constraint is to say that the variation of the kinetic term is not much larger than the variation of the potential term, or that the rate of change of $1 + w$ is small as compared to the Hubble parameter. This gives

$$\frac{\partial_\eta (1 + w)}{1 + w} \ll H, \quad (7.44)$$

so that, approximating $1 + w$ by $\frac{\dot{\phi}^2}{a^2 V}$,

$$2 \frac{\ddot{\phi}}{\dot{\phi}} - 2H - \frac{\dot{\phi}V'}{V} \ll H. \quad (7.45)$$

Now, the Klein-Gordon equation can be rewritten

$$\partial_\eta \left(\frac{1}{2} \frac{\dot{\phi}^2}{a^2} \right) + \dot{V} + 3H \frac{\dot{\phi}^2}{a^2}. \quad (7.46)$$

Using these two equations, one has

$$\frac{\partial_\eta \frac{\dot{\phi}^2}{a^2}}{\frac{\dot{\phi}^2}{a^2}} \ll H, \quad (7.47)$$

so that one can drop the $\ddot{\phi} - H\dot{\phi}$ term in the Klein-Gordon equation which is modified as

$$3H\dot{\phi} \sim -a^2 V', \quad (7.48)$$

or

$$9H^2 \frac{\dot{\phi}^2}{a^2} \sim a^2 V'^2. \quad (7.49)$$

Using again the first condition, one obtains

$$a^2 V'^2 \ll 9H^2 V. \quad (7.50)$$

Using the first Friedmann equation (recalling that we neglect the curvature K and the kinetic term in the energy density) gives

$$V'^2 \ll 24\pi G V^2. \quad (7.51)$$

We therefore define the *first slow-roll parameter* ϵ

$$\epsilon \equiv \frac{V'^2}{16\pi G V^2} \ll 1. \quad (7.52)$$

Using the modified Klein-Gordon equation, one can also express ϵ as

$$\epsilon = \frac{3}{2} \frac{\dot{\phi}^2}{a^2 V}. \quad (7.53)$$

The advantage to of taking 16 instead of 24 in the denominator of the first definition is that one can rewrite ϵ as

$$\epsilon = -\frac{\partial_t H}{H^2} = 1 - \frac{\dot{H}}{H^2}, \quad (7.54)$$

with the drawback is that although $\epsilon \rightarrow 0$ corresponds to the de Sitter case, it is $\epsilon = \frac{3}{2}$ that corresponds to the case $a \propto t$ case (i.e., the limit between accelerating and decelerating cases). This parameter expresses that the rate of change of the Hubble parameter H is small: $|\partial_t H/H| = \epsilon H \ll H$. The first slow roll parameter is in fact insufficient to describe the two conditions we have imposed because we also want ϵ to remain small: we also want $\dot{\epsilon}/\epsilon \ll H$. Starting from (7.52), we have

$$\frac{\dot{\epsilon}}{\epsilon} = \left(2 \frac{V''}{V'} - \frac{V'}{V} \right) \dot{\phi}. \quad (7.55)$$

Using the modified Klein-Gordon equation and then the first Friedmann equation, one gets

$$\dot{\epsilon} = 2H\epsilon \left(\epsilon - \frac{V''}{8\pi GV} \right). \quad (7.56)$$

One can then define the *second slow-roll parameter* δ (not to be mistaken with the density contrast δ , see below...) by

$$\delta \equiv \frac{V''}{8\pi GV}, \quad |\delta| \ll 1. \quad (7.57)$$

One can continue this procedure by imposing that δ varies slowly and define a whole hierarchy of slow-roll parameters, which involve higher and higher derivatives of the potential. Unless the potential has some discontinuous high-order derivative, imposing that the first two slow roll parameters are small is generally enough⁵.

An example: chaotic inflation

Before continuing the study of the dynamics of the inflationary phase, let us focus on one specific example of inflationary scenario: chaotic inflation. Historically, this was not the first model that was proposed but we think it was the first to provide a satisfactory scenario. The main difficulty with inflation is to have the slow roll conditions to be satisfied at some epoch. Indeed, as we saw, one need to put the field away from the minimum of its potential for the inflaton to behave like a cosmological constant. The first inflationary model (Guth 1981) supposed a potential like that of Eq. (7.28) where the field slowly moved away from its minimum because of a phase transition. However, this led to a number of difficulties, see for example Ref. (Liddle & Lyth). Fortunately, it was soon realized that it was not necessary to have a time dependent potential for inflation to proceed. Linde (Linde 1985) noticed that inflation could start as soon as the Universe would exit the Planck era. The idea was that it is reasonable to suppose that at the end of the Planck era (when $\rho_\phi \sim \rho_{\text{Pl}}$), no large-scale correlation could be expected in the scalar field, so that one could expect very “irregular” (hence, chaotic) initial condition with

$$\frac{1}{2} \frac{1}{a^2} \dot{\phi}^2 \sim O(M_{\text{Pl}}^4), \quad \frac{1}{2} \frac{1}{a^2} (\nabla\phi)^2 \sim O(M_{\text{Pl}}^4), \quad V(\phi) \sim O(M_{\text{Pl}}^4) \quad (7.58)$$

at any point. It is therefore possible that (in the sense that it is not forbidden that at some point one has)

$$V \gg \frac{1}{a^2} \dot{\phi}^2, \quad \frac{1}{a^2} (\nabla\phi)^2. \quad (7.59)$$

⁵Note that although the definition of the first slow-roll parameter is widely accepted, the way the second slow-roll parameter varies among the authors, see, e.g. Ref. Schwartz et al 2001.

When this is the case, the conditions for inflation to start are satisfied. Whether or not such how likely such a configuration is on a sufficiently large region requires more involved calculations, but at least if one supposes that the Universe is very irregular at the exit of the Planck era, then the probability that this happens is certainly non zero. Since the volume of the inflationary regions grow much faster than that of the other regions, one can hope that after some time most of the volume of the universe is contained within inflationary regions. The general pictures is therefore that the Universe is filled by a number of (large) patches which all still undergo or have undergone some inflationary phase. The duration of this phase varies form patch to patch: some may still be inflating whereas our patch has stopped from inflating long ago (and in this case would be much smaller). Therefore even if within an inflated region the Universe is very homogeneous and isotropic, one expects that the very large scale structure of the universe is still very irregular as a reminiscence of its chaotic state at the exit of the Planck era.

Which potential can lead to inflation within this context? If one considers power law potentials like $V = \frac{1}{2}m^2\phi^2$ or $V = \frac{1}{4}\lambda\phi^4$, then the slow roll condition impose that

$$\phi \gg M_{\text{Pl}}, \quad (7.60)$$

but do not say anything about the value of the parameters m or λ . However, the condition that inflation starts at the exit of the Planck era imposes that $\rho_\phi \sim M_{\text{Pl}}^4$, which translates into

$$m \ll M_{\text{Pl}}, \quad (7.61)$$

$$\lambda \ll 1. \quad (7.62)$$

Whether or not it is natural to have “small” parameters in the potential of some field at the exit of Planck era is at present unknown. However it is generally considered as unlikely and might be a problem for the whole scenario. We will not enter into the (technical) debate about this, but will keep in mind that in its simplest form these last constraints might be unsatisfactory and require more involved scenarios. However, and for definiteness, we will suppose from now on that inflation occurs with a field with such a potential.

When does inflation end?

From the modified Klein-Gordon equation and the first Friedmann equation (and assuming $V \propto \frac{X}{n}\phi^n$), one has

$$\frac{a}{a_i} = \exp \left(-\frac{8\pi G}{2n} (\phi^2 - \phi_i^2) \right), \quad (7.63)$$

where the subscript i stands for quantities evaluated at the beginning of the inflationary phase. If $V = \frac{1}{4}\lambda\phi^4$, the time evolution of the field is given by

$$\phi = \phi_i \exp \left(-\sqrt{\frac{4\lambda}{24\pi G}}(t - t_i) \right), \quad (7.64)$$

and when $n \neq 4$, one has

$$\phi^{\frac{4-n}{2}} - \phi_i^{\frac{4-n}{2}} = -\frac{2}{4-n} \sqrt{\frac{nX}{24\pi G}}(t - t_i). \quad (7.65)$$

The end of the accelerated phase occurs when

$$(\partial_t \phi)^2 > V, \quad (7.66)$$

which here happens when

$$\phi^2 = \frac{n}{24\pi G} \sim M_{\text{Pl}}^2. \quad (7.67)$$

(Unless exponent n is abnormally large or small in front of 1.) As a consequence, the number of e -folds is given by

$$N = \ln \left(\frac{a_f}{a_i} \right) = \frac{8\pi G}{2n} (\phi_i^2 - \phi_f^2) \sim \left(\frac{\phi_i}{M_{\text{Pl}}} \right)^2. \quad (7.68)$$

Since one has initially $\phi_i \gg M_{\text{Pl}}$, one expects the number of e -folds to be very large. This means that in chaotic inflation there is absolutely no difficulty to reach the minimum number of e -fold of 60. This also mean that there is probably no hope to see the very large scale structure of the universe: in typical models, N can easily reach 10^5 or 10^6 : the size of the inflationary region we are in is something like e^{10^5} or e^{10^6} times larger than the size of our observable universe.

4. Basics of cosmological perturbations

So far, we have been able to build a scenario which solves the horizon, flatness and monopole problems. As we already said, this scenario also explains the existence of an almost scale invariant spectrum in the cosmological perturbations. However, the derivation of this crucial result is significantly more involved than the previous one. First because we need to do a careful study of the cosmological perturbation in the context of general relativity and in an expanding universe. Second because we then need to solve these equations in the specific case of inflation. Third, because perturbation theory only tell of the evolution of cosmological perturbations, so that we need to specify the

initial conditions for the perturbations. These initial conditions will be given by the quantum field theory: although space time and the unperturbed part of the inflaton can be described classically, it exhibits quantum fluctuations. The miracle of the inflationary scenario is that (i) these quantum fluctuations do survive to the accelerated expansion phase and (ii) they subsequently behave as classical perturbations. Before going into the study of the quantum fluctuations, we are first going to give an overview of the classical perturbations in general relativity. Although long and somewhat tedious, this part seems to us necessary in order to understand correctly the subtle difference between the general relativistic study and the Newtonian analysis of cosmological perturbations (however, this part can be skipped, and only the results of paragraph 7.4.0 will be used later).

Gauge issues

The main difference between perturbations in a Newtonian framework and in general relativity comes from the definition of the coordinate system. In Newtonian theory, there is a fixed space time, in which there exist two configurations, the “unperturbed” and the “perturbed” one. It is easy to make comparison between any quantities because one has simply to compare the value of any field in the two configurations at the same point. In general relativity, the situation is more complicated. Because of the influence of gravity, the perturbed and unperturbed configurations correspond to *different* spacetimes (or manifolds in the mathematicians language). In order to make comparisons between two configurations, it is therefore necessary to define some mappings i and i' between \mathbf{R}^4 and each configurations. In some situations, in particular in the “unperturbed one”, there may be some natural mapping, for example such that the mapping allows to define a conformally Minkowskian metric. But in the perturbed case, there is in general no natural straightforward way to define a mapping between \mathbf{R}^4 and the perturbed manifold.

The first step is therefore to understand how to define unambiguously a mapping⁶. Let us first define two mappings between \mathbf{R}^4 and the two (unperturbed

⁶This problem is in fact related with the well-known gauge problems in electromagnetism. From the Maxwell equations $\nabla \cdot \mathbf{B} = 0$ and $\nabla \wedge \mathbf{E} = -\partial_t \mathbf{B}$, one can naturally replace the electric field \mathbf{E} and the magnetic field \mathbf{B} by the four-vector (V, \mathbf{A}) with $\mathbf{B} = \nabla \wedge \mathbf{A}$ and $\mathbf{E} = -\nabla V - \partial_t \mathbf{A}$. However, it is easy to see that if (V, \mathbf{A}) is a solution, so is four-vector $(V - \partial_t \phi, \mathbf{A} + \nabla \phi)$. Therefore in order to solve the equations, one has to impose one condition on the four-vector (because there is one degree of freedom — ϕ — in the solution), such as $V = 0$ or $\nabla \cdot \mathbf{A} = 0$. Such a condition is called a gauge condition. It is unphysical in the sense this is just a technical (or numerical) constraint one imposes to solve the equations, and the final observable quantities \mathbf{E} and \mathbf{B} do not depend on it. Still, it is absolutely necessary to remove this ambiguity in the four vector before solving the equations. We face the same situation in general relativity, expect that what has to be fixed is not some field, but the coordinate system itself.

and perturbed) spacetimes M and M' :

$$i : \mathbf{R}^4 \rightarrow M \quad (7.69)$$

$$x^\mu \rightarrow P, \quad (7.70)$$

and

$$i' : \mathbf{R}^4 \rightarrow M' \quad (7.71)$$

$$x^\mu \rightarrow P', \quad (7.72)$$

We then define a “perturbation” of some quantity X by

$$\delta X(x^\mu) \equiv X(P'(x^\mu)) - X(P(x^\mu)), \quad (7.73)$$

so that we evaluate the same quantity X at points of the two different spacetimes which have the same coordinate (label). We will state that one spacetime can be considered as a small perturbation of the other if any perturbation δX is sufficiently small at any point (more precisely, if there exist some sets of mappings which allow to have small perturbations everywhere). As we easily see, the perturbation defined that way explicitly depend on the (arbitrary) choice of the mapping $i' \circ i^{-1}$. Therefore, we should note explicitly this choice of mapping:

$$\delta_{i,i'} X(x^\mu) = X(i'(x^\mu)) - X(i(x^\mu)). \quad (7.74)$$

Let us now consider a new mapping between \mathbf{R}^4 and M' ,

$$i'' : \mathbf{R}^4 \rightarrow M' \quad (7.75)$$

$$x^\mu \rightarrow P'', \quad (7.76)$$

such that the difference between the two coordinate systems defined by i' and i'' is small: for all P , one has

$$d(P, i'' \circ i'^{-1}(P)) \ll \left| \frac{\nabla X}{\nabla X} \right|_P. \quad (7.77)$$

In this case, the new quantity

$$\delta_{i,i''} X(x^\mu) = X(i''(x^\mu)) - X(i(x^\mu)) = \delta_{i,i'} X(x^\mu) + d(i'(x^\mu), i''(x^\mu)) \nabla X|_{i'(x^\mu)} \quad (7.78)$$

is also small and can equally define the “perturbation” of the quantity X . We see that to proceed, we are first going to have to define (or, in fact, to choose) one specific mapping in order to have an unambiguous definition of the perturbation. The coordinate choice we shall do is somehow arbitrary, and does not necessarily have a clear physical meaning. This explains why there is a number of different choices which are made in the literature, some being more relevant than others for specific purpose.

Fixing the coordinate system

We study cosmological perturbations in a four dimensional spacetime. Since the choice of the coordinate system is arbitrary, by changing the coordinate system, one introduces four unphysical degrees of freedom (as many degrees of freedom as coordinates). The manifold is entirely described by a metric, which is a 4×4 symmetric tensor, which therefore has ten independent components. Since we can introduce four unphysical degrees of freedom through coordinate choice, we see that there are in fact $10 - 4 = 6$ true degrees of freedom in the metric⁷.

Let us therefore study how the metric components change under the coordinate change $i''^{-1} \circ i'$. Let us define

$$i''^{-1} \circ i' : \mathbf{R}^4 \rightarrow \mathbf{R}^4 \quad (7.79)$$

$$x^\mu \rightarrow x'^\mu = x^\mu + \xi^\mu, \quad (7.80)$$

so that a point P which has coordinates x^μ with the i' mapping has coordinates $x^\mu + \xi^\mu$ with the i'' mapping. From the conditions (7.77, 7.78), we shall perform a first order expansion in term of the small coordinate change ξ^μ and neglect any term of order $O(\xi^{\mu 2})$. Differential geometry tells us that the metric components at some point P which has coordinates x^μ in mapping i' and coordinates x'^μ in mapping i'' are related through

$$g'_{\alpha\beta}(P) = g'_{\alpha\beta}(x'^\mu) = \frac{\partial x^\mu}{\partial x'^\alpha} \frac{\partial x^\nu}{\partial x'^\beta} g_{\mu\nu}(x^\mu), \quad (7.81)$$

where the superscript ' above g means that it is evaluated through the mapping i'' . From Eq. (7.79), one has

$$\frac{\partial x'^\alpha}{\partial x^\beta} = \delta_\beta^\alpha + \frac{\partial \xi^\alpha}{\partial x^\beta}, \quad (7.82)$$

which at lowest order can be inverted into

$$\frac{\partial x^\alpha}{\partial x'^\beta} = \delta_\beta^\alpha - \frac{\partial \xi^\alpha}{\partial x^\beta}, \quad (7.83)$$

Therefore, one obtains

$$g'_{\alpha\beta}(P) = g_{\alpha\beta}(P) - g_{\mu\beta} \frac{\partial \xi^\mu}{\partial x^\alpha} - g_{\alpha\nu} \frac{\partial \xi^\nu}{\partial x^\beta}. \quad (7.84)$$

⁷Usually it is said that four-dimensional general relativity has only two degrees of freedom. This is because when one imposes that one is in the vacuum, only two out of the six above mentioned degrees of freedom survive (these are the gravitational wave, which can propagate even in the absence of matter). But in general, in the presence of matter, the gravitational field depends on six numbers.

However, as we already explained, in order to define a perturbation, we want to compare the values of $g_{\alpha\beta}$ at points which have the same coordinates. Here, point P has coordinates x^μ with mapping i' and coordinates $x^\mu + \xi^\mu$ with mapping i'' . Let us consider point P' which has coordinates $x - \xi^\mu$ with mapping i' and therefore coordinates x^μ with mapping i'' . What we are interested in is

$$\delta_{i',i''} g = g'_{\alpha\beta}(P') - g_{\alpha\beta}(P). \quad (7.85)$$

We obtain

$$\delta_{i',i''} g = -\xi^\rho \partial_\rho g_{\alpha\beta} - g_{\mu\beta} \frac{\partial \xi^\mu}{\partial x^\alpha} - g_{\alpha\nu} \frac{\partial \xi^\nu}{\partial x^\beta}, \quad (7.86)$$

which can be rewritten using covariant derivatives

$$\delta_{i',i''} g = -\xi^\rho D_\rho g_{\alpha\beta} - g_{\mu\beta} D_\alpha \xi^\mu - g_{\alpha\nu} D_\beta \xi^\nu = -D_\alpha \xi_\beta - D_\beta \xi_\alpha. \quad (7.87)$$

Scalar-vector-tensor decomposition

We start here from an FRW metric, that is an unperturbed metric such that

$$g_{00} = a^2, \quad (7.88)$$

$$g_{0i} = 0, \quad (7.89)$$

$$g_{ij} = -a^2 \gamma_{ij}. \quad (7.90)$$

The perturbed metric (i.e., the components $\delta_{i,i'} g_{\alpha\beta}$ that we will simply note $\delta g_{\alpha\beta}$) can be defined without loss of generality as

$$\delta g_{00} = 2a^2 A, \quad (7.91)$$

$$\delta g_{0i} = a^2 B_i, \quad (7.92)$$

$$g_{ij} = -a^2 (2C \gamma_{ij} + 2E_{ij}), \quad (7.93)$$

where the indices of vector B_i and of tensor E_{ij} are raised and lowered with metric γ_{ij} . Still without loss of generality, we can decompose B_i into a gradient and a curl part:

$$B_i = \nabla_i B + \bar{B}_i, \quad (7.94)$$

where the ∇_i represents the covariant derivative with respect to metric γ_{ij} ⁸, and imposing

$$\nabla_i \bar{B}^i = 0. \quad (7.95)$$

⁸For the reader that would not know about covariant derivative, suppose that one is in the Euclidean case ($K = 0$) and that the spatial part of the unperturbed manifold has Cartesian coordinates, so that γ_{ij} is simply a Kronecker symbol, and the ∇_i are simply ordinary derivative with respect to the coordinate x^i .

Equivalently, we can decompose E_{ij} into a divergence-free and trace-free tensor $\bar{\bar{E}}_{ij}$, a divergence-free vector \bar{E}_i and a scalar E :

$$E_{ij} = \nabla_{ij}E + \frac{1}{2}(\nabla_i\bar{E}_j + \nabla_j\bar{E}_i) + \bar{\bar{E}}_{ij}, \quad (7.96)$$

$$\gamma_{ij}\bar{\bar{E}}^{ij} = 0, \quad (7.97)$$

$$\nabla_i\bar{\bar{E}}^{ij} = 0, \quad (7.98)$$

$$\nabla_i\bar{E}^i = 0. \quad (7.99)$$

The aim of this decomposition is that, as we shall see, the sets of “scalar” quantities A, B, C, E , “vector” quantities \bar{B}_i, \bar{E}_i and “tensor” quantities $\bar{\bar{E}}_{ij}$ evolve independently from each others. Note that we are left with four scalars (A, B, C, E), two vectors (\bar{B}_i, \bar{E}_i) which both have three components but which obey divergenceless constraint so that we have four independent components, and one tensor ($\bar{\bar{E}}_{ij}$) which is a 3×3 symmetric matrix with one traceless constraint and three divergenceless constraints, which therefore has only two independent components. As expected, we are still left with ten independent components for the metric perturbations. As we said, four of these perturbations are in fact unphysical. Let us decompose the infinitesimal coordinate change ξ^μ into

$$\xi^\mu = (T, L^i), \quad (7.100)$$

and let use decompose L^i into a scalar and a vector part:

$$L_i = \gamma_{ij}L^j, \quad (7.101)$$

$$L_i = \nabla_iL + \bar{L}_i, \quad (7.102)$$

$$\nabla_i\bar{L}^i = 0. \quad (7.103)$$

This suggests that two out of the four scalar degrees of freedom are unphysical and one out of the two two-component vectors is also unphysical.

Using (7.87) it is now possible to see how the quantities A, B , etc transform under coordinate change ξ^μ . For example, one has

$$\delta_{i,i''}g_{00} = \delta_{i,i'}g_{00} - 2a^2(TH + \dot{T}), \quad (7.104)$$

a notation that we will shorten into

$$\delta g_{00} \rightarrow \delta g_{00} - 2a^2(TH + \dot{T}). \quad (7.105)$$

In a similar way, it is easy to show that

$$A \rightarrow A - HT - \dot{T}, \quad (7.106)$$

$$B \rightarrow B + \dot{L} - T, \quad (7.107)$$

$$C \rightarrow C - HT, \quad (7.108)$$

$$E \rightarrow E - L, \quad (7.109)$$

$$\bar{B}_i \rightarrow \bar{B}_i + \dot{\bar{L}}_i, \quad (7.110)$$

$$\bar{E}_i \rightarrow \bar{E}_i - \bar{L}_i, \quad (7.111)$$

$$\bar{\bar{E}}_{ij} \rightarrow \bar{\bar{E}}_{ij}. \quad (7.112)$$

The presence of the above mentioned unphysical degrees of freedom translates into the fact that (i) a given component can in general be put to any value through the relevant coordinate transform and (ii) there must exist combination of the metric perturbations which remain unchanged under these coordinate transforms. The number of these quantities (called for obvious reasons “gauge invariant quantities”) is precisely equal to the true number of physical degrees of freedom.

From the four scalar quantities, it is possible to define two independent gauge invariant quantities, usually called the Bardeen potentials Φ and Ψ :

$$\Phi \equiv -C + H(B + \dot{E}), \quad (7.113)$$

$$\Psi \equiv A - H(B + \dot{E}) - (\dot{B} + \ddot{E}). \quad (7.114)$$

(Note that there are some variations in the literature about the definition of these quantities; sometimes the definitions of Φ and Ψ are swapped, and their sign is also sometimes different. Here we choose the convention that Φ and Ψ are equal in the absence of anisotropic stress (see below), and that Φ is the quantity that appears in the Laplacian term of the 00 part of the Einstein equations (the general relativistic analog of the Poisson equation), thus following the Newtonian convention to note the gravitational potential by Φ .) It is of course possible to define other scalar gauge invariant quantities. For example one can define

$$X \equiv A - C - \partial_\eta \left(\frac{C}{H} \right), \quad (7.115)$$

but this quantity can be expressed in term of the two gravitational potentials:

$$X = \Psi + \Phi + \partial_\eta \left(\frac{\Phi}{H} \right). \quad (7.116)$$

For the vector quantities, there exists only one gauge invariant divergenceless vector \bar{V}_i :

$$\bar{V}_i \equiv \bar{B}_i + \dot{\bar{E}}_i. \quad (7.117)$$

Finally, let us note that the tensor part \bar{E}_{ij} is itself gauge invariant. As we said, one has to choose a gauge (i.e., a coordinate system) in order to solve the equations. It is well-known that the vector modes are generally irrelevant for cosmology, so that most of the coordinate systems which are used only deal with scalar perturbations. Let us review some of the most used coordinate systems.

Some gauges

Choosing a gauge corresponds to choosing a coordinate system. The coordinate is well defined when there exists only one mapping which allows to go from some arbitrary coordinate system (where the metric perturbations have no preferred form) to one specific coordinate system (which obeys some constraints). Note that not all the coordinate systems are possible. For example, one cannot impose simultaneously $A = C = 0$ as the quantity X is in general not 0. Also, imposing some constraints on the coordinate system does not necessarily fix it completely (as is the case for the synchronous gauge). We strongly advocate the use of a well-defined coordinate system such as the Newtonian or the flat-slicing gauges defined below.

Synchronous gauge. The synchronous gauge is defined by the constraint $\delta g_{0\mu} = 0$. It translates into imposing a mapping i'' such that $A = B = \bar{B}_i = 0$. We strongly discourage the use of this gauge, for the reason that it is not completely fixed. Indeed, the coordinate change

$$T = \frac{f(x^i)}{a}, \quad (7.118)$$

$$L = \int_{g(x^i)}^{\eta} \frac{f(x^i)}{a(\eta')} d\eta, \quad (7.119)$$

$$L_i = h_i(x^i), \quad (7.120)$$

can modify C and E without changing A and B_i . This means that imposing the above condition does not completely fix the coordinate system. The reason is that this gauge is physically defined by a set of free falling observers. One chooses a spacelike hypersurface which defines the observers and the other spacelike hypersurfaces are defined by the trajectory and the proper time of each observer. The ambiguity is that neither the coordinate system of the hypersurface nor the way the clocks of different observers are set are specified and can be arbitrarily modified. This ambiguity in the definition of the coordinate system leads to the apparition of “gauge modes”, that is behavior in the evolution of the perturbation which do not have any physical meaning. If these gauge modes happen to become much larger than the physically relevant quantities, then a number of numerical stability problems can occur. The only

way to solve then is to tune extremely carefully the initial conditions for the perturbations so that the unwanted gauge modes do not appear. But this is of course complicated, as well as useless since well-defined coordinate systems easily allow to avoid this problem⁹.

Longitudinal or Newtonian gauge. A far better defined gauge is the Newtonian gauge (also called longitudinal gauge). It is defined as the gauge in which the (scalar) metric perturbations are diagonal: $B = E = 0$. From the definition of the Bardeen potentials, this means that the scalar part of the metric perturbations are

$$\delta g_{00} = 2a^2\Psi, \quad (7.121)$$

$$\delta g_{ij} = 2a^2\Phi\gamma_{ij}. \quad (7.122)$$

Despite its name, we do not think that this gauge corresponds to the most “Newtonian” coordinate system, that is, the coordinate system in which the perturbation equation behave most closely to their Newtonian analog (see below), as metric perturbations arise in the mass conservation equation.

Flat-slicing gauge. Another very useful coordinate system is the so-called flat-slicing gauge. It is defined as the system in which the spatial part of metric is unperturbed (for the scalar quantities): $C = E = 0$. This gives

$$\delta g_{00} = X, \quad (7.123)$$

$$\delta g_{0i} = \nabla_i \left(\frac{\Phi}{H} \right). \quad (7.124)$$

This is the coordinate system in which we will write the cosmological perturbation equations.

Gauge invariant matter perturbations

The perturbed stress-energy tensor of a perfect fluid can be written

$$T_{\mu\nu} + \delta T_{\mu\nu} = (P + \delta P + \rho + \delta\rho)(u_\mu + \delta u_\mu)(u_\nu + \delta u_\nu) - (P + \delta P)(g_{\mu\nu} + \delta g_{\mu\nu}) + \Pi_{\mu\nu}, \quad (7.125)$$

where $\rho + \delta\rho$ represent the density, $P + \delta P$ represents the pressure, $u_\mu + \delta u_\mu$ represent the bulk four-velocity (by definition u_μ is a unit timelike vector, $u_\mu u^\mu = 1$), and $\Pi_{\mu\nu}$ is the anisotropic stress of the fluid. This tensor is orthogonal to the four-velocity u_μ .

⁹Still, for some reasons we do not understand this gauge remains very popular in the cosmological literature...

Using the same derivation as for the metric components, it is easy to check that the perturbed part δX of a homogeneous quantity X transforms as

$$\delta X \rightarrow \delta X - T\dot{X}, \quad (7.126)$$

so that one can define two gauge invariant quantities δX^\sharp and δX^\flat as

$$\delta X^\sharp \equiv \delta X - \dot{X}(B + \dot{E}), \quad (7.127)$$

$$\delta X^\flat \equiv \delta X - \dot{X}\frac{C}{H}. \quad (7.128)$$

These two quantities are related to each other by $\delta X^\sharp = \delta X^\flat - \dot{X}\Phi$. They reduce to the perturbation that an observer would see if he imposed the conditions $B = E = 0$ or $C = 0$ in his coordinate system, respectively, that is δX^\sharp represents the perturbation of X in Newtonian gauge, whereas δX^\flat represents the perturbation of X in the flat-slicing gauge.

For the four-velocity $u^\mu + \delta u^\mu = dx^\mu/d\tau$, recalling that it is a unit vector (i.e., that $(u^\mu + \delta u^\mu)(u^\nu + \delta u^\nu)(g_{\mu\nu} + \delta g_{\mu\nu}) = 1$, one has

$$u^\mu + \delta u^\mu = \left(\frac{1-A}{a}, \frac{v^i}{a} \right), \quad (7.129)$$

where again we use metric γ_{ij} to raise and lower the indices of three-vector v^i that we shall also decompose into a scalar part v and a divergenceless part \bar{v}_i

$$v_i = \nabla_i v + \bar{v}_i, \quad (7.130)$$

with $\nabla_i \bar{v}^i = 0$. Again, it is easy to check that v_i transforms into

$$v_i \rightarrow v_i - \dot{L}_i, \quad (7.131)$$

so that one can easily define the following gauge invariant quantities (which are the same in Newtonian and longitudinal gauge):

$$v^\sharp = v^\flat = v - \dot{E}, \quad (7.132)$$

$$\bar{v}_i^\sharp = \bar{v}_i^\flat = \bar{v}_i - \dot{\bar{E}}_i. \quad (7.133)$$

Finally, the anisotropic stress $\Pi_{\mu\nu}$ is already gauge invariant (because it has non unperturbed counterpart), and we also decompose it into

$$\Pi_{00} = 0, \quad (7.134)$$

$$\Pi_{0i} = 0, \quad (7.135)$$

$$\Pi_{ij} = P\pi_{ij}, \quad (7.136)$$

$$\pi_{ij} = \left(\nabla_{ij} - \frac{1}{3}\Delta \right) \pi + \frac{1}{2} (\nabla_i \bar{\pi}_j + \nabla_j \bar{\pi}_i) + \bar{\pi}_{ij}. \quad (7.137)$$

Einstein and conservation equations

After this rather long but necessary introduction, we can now compute the perturbed Einstein equations. Using the quantities defined above, they read

$$2\Delta\Phi - 6H^2X = 8\pi Ga^2\delta\rho^\flat, \quad (7.138)$$

$$2HX - 2H\Phi + 2\frac{\dot{H}}{H}\Phi = -8\pi Ga^2(P + \rho)v^\flat, \quad (7.139)$$

$$2H\dot{X} + 4\dot{H}X + 2H^2X = 8\pi Ga^2\left(\delta P^\flat + \frac{2}{3}P\Delta\pi\right), \quad (7.140)$$

$$\Phi - \Psi = 8\pi Ga^2P\pi. \quad (7.141)$$

If we neglect the expansion term, we see that the first equation reduces to the usual Poisson equation. The last equation insures that the two Bardeen potentials are similar since in general the anisotropic stresses are small (they are negligible for non relativistic matter as well as for a scalar field). Note that in the absence of any form of matter all the scalar metric perturbations are 0. In addition to the scalar perturbations, there exists one equation for the tensor modes:

$$\ddot{\bar{E}}_{ij} + 2H\dot{\bar{E}}_{ij} + (2K - \Delta)\bar{E}_{ij} = 8\pi Ga^2P\bar{\pi}_{ij}. \quad (7.142)$$

This corresponds to a damped oscillator, which was expected since the tensor modes correspond to gravitational waves. These can propagate even in the absence of matter. In addition to these equations, one also has to specify the equations governing the evolution of matter, that is the matter conservation and the Euler equations. In the absence of gravitational interactions, they read

$$\dot{\delta}^\flat = -(1+w)\Delta v^\flat - 3H\left(\frac{\delta P^\flat}{\rho} - w\delta^\flat\right), \quad (7.143)$$

$$\dot{v}^\flat = -(1-3c_s^2)Hv^\flat - \Psi - 3c_s^2\Phi - \frac{\delta P^\flat}{P+\rho} - \frac{2}{3}\frac{w}{1+w}(\Delta + 3K)\pi \quad (7.144)$$

The first equation corresponds to the mass conservation equation $\partial_t\rho + \nabla(\rho\mathbf{v}) = 0$. The second equation corresponds to the Euler equation: the first term of the right handside is the damping due to the expansion, the second and third terms correspond to the usual gravitational force, the fourth term corresponds to the pressure gradient term, and the last term is the viscosity term. Note that the exact form of these equations depend on the choice of the coordinate system we did. For example, should we have chosen the Newtonian gauge, then some metric perturbation would have appeared in the mass conservation equation (because of the relation between the density contrasts in each gauge). This simply reflects the fact that when one is not in the flat slicing gauge, there

is sometimes some ‘‘shrinking’’ of the coordinate system of the spacelike sections which make the density artificially increase: it is not the true density that increases, just the coordinate density because of some local shrinking of the coordinate grid.

5. Inflationary perturbations

With the help of the preceding section, we are now in position of writing the equations corresponding to the evolution of the cosmological perturbations during inflation.

Perturbed scalar field

As we said earlier, the stress-energy tensor of a scalar field is given by

$$T_{\mu\nu} = D_\mu \phi D_\nu \phi - g_{\mu\nu} \left(\frac{1}{2} D_\rho \phi D^\rho \phi - V \right). \quad (7.145)$$

The density is defined as the eigenvalue of unit timelike eigenvectors U_μ . This eigenvector is given by

$$U_\mu = \pm \frac{D_\mu \phi}{\sqrt{D_\rho \phi D^\rho \phi}}, \quad (7.146)$$

where the sign is chosen so that $U_0 > 0$. The corresponding eigenvalue is therefore

$$\rho = \frac{1}{2} D_\rho \phi D^\rho \phi + V. \quad (7.147)$$

Equivalently, the pressure is defined as the eigenvalue associated to the space-like unit eigenvectors. One easily finds

$$P = \frac{1}{2} D_\rho \phi D^\rho \phi - V. \quad (7.148)$$

With these definitions, it is possible to compute the gauge invariant perturbed component of the stress energy tensor:

$$\delta\rho^b = \frac{1}{a^2} \dot{\phi} \delta\dot{\phi}^b - \frac{1}{a^2} \dot{\phi}^2 X + V' \delta\phi^b, \quad (7.149)$$

$$\delta P^b = \frac{1}{a^2} \dot{\phi} \delta\dot{\phi}^b - \frac{1}{a^2} \dot{\phi}^2 X - V' \delta\phi^b, \quad (7.150)$$

$$v_i^b = -\frac{\nabla_i \delta\phi^b}{\dot{\phi}} + \frac{\nabla_i \Phi}{H}, \quad (7.151)$$

$$\pi_{ij} = 0. \quad (7.152)$$

Note that they involve metric perturbations as the metric is present in the kinetic term ($D_\mu\phi D^\mu\phi = g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi$). The Einstein equations then give

$$2\Delta\Phi - 6H^2X = 8\pi G \left(\dot{\phi}\delta\phi^\flat - \dot{\phi}^2X + a^2V'\delta\phi^\flat \right), \quad (7.153)$$

$$2HX - 2H\Phi + 2\frac{\dot{H}}{H}\Phi = 8\pi G \left(\dot{\phi}\delta\phi^\flat - \dot{\phi}^2\frac{\Phi}{H} \right), \quad (7.154)$$

$$2H\dot{X} + 4\dot{H}X + 2H^2X = 8\pi G \left(\dot{\phi}\delta\phi^\flat - \dot{\phi}^2X - a^2V'\delta\phi^\flat \right), \quad (7.155)$$

$$\Phi - \Psi = 0. \quad (7.156)$$

In particular, combining these equations, we have

$$X = 2\Phi + \partial_\eta \left(\frac{\Phi}{H} \right), \quad (7.157)$$

$$2HX + 2\frac{K}{H}X = 8\pi G\dot{\phi}\delta\phi^\flat, \quad (7.158)$$

$$\Delta\Phi - H\dot{X} - (4H^2 + 2\dot{H})X = 8\pi Ga^2V'\delta\phi^\flat. \quad (7.159)$$

In particular, when the curvature is negligible (which we shall suppose now), the second equation gives a proportionality relation between X and $\delta\phi^\flat$.

For the evolution of the scalar field, the most convenient is to use directly the perturbed version of the Klein-Gordon equation (which is in fact equivalent to the Euler equations), $D_\mu D^\mu\phi + \partial_\phi V = 0$. This gives

$$\ddot{\delta\phi}^\flat + 2H\dot{\delta\phi}^\flat + (a^2V'' - \Delta)\delta\phi^\flat = \dot{\phi} \left(\dot{X} - \Delta\frac{\Phi}{H} \right) + (2\ddot{\phi} + 4H\dot{\phi})X. \quad (7.160)$$

We recognize as we could have guessed that this is a wave equation with a source term and a damping term. With the help of Eqns (7.159) and then (7.158), it is possible to get rid of the gravitational perturbations in the right handside of the above equations and obtain a closed equation in term of $\delta\phi^\flat$. After heavy manipulations, one gets¹⁰

$$\ddot{\delta\phi}^\flat + 2H\dot{\delta\phi}^\flat - \Delta\delta\phi^\flat = \left(\frac{\partial_\eta^2(a\dot{\phi}/H)}{a\dot{\phi}/H} - \frac{\ddot{a}}{a} \right) \delta\phi^\flat. \quad (7.161)$$

If we set

$$u_S \equiv a\delta\phi^\flat, \quad (7.162)$$

$$z_S = a\frac{\dot{\phi}}{H}, \quad (7.163)$$

¹⁰The tenacious reader is strongly encouraged to derive this equation at least once...

we see that this equation can finally be put into the beautifully compact form

$$\ddot{u}_S + \left(k^2 - \frac{\ddot{z}_S}{z_S} \right) u_S = 0. \quad (7.164)$$

Going from the field to the metric perturbation is now easy: instead of using (7.158) it is more convenient to define

$$V_S \equiv \frac{2}{8\pi G} \frac{a^2}{H} \Phi, \quad (7.165)$$

so that one has

$$u_S = \frac{1}{z_S} \dot{V}_S, \quad (7.166)$$

$$V_S = -\frac{1}{k^2} z_S^2 \partial_\eta \left(\frac{u_S}{z_S} \right). \quad (7.167)$$

Using the gravitational wave equation (7.142), it is also possible to get a similar expression. With the definitions

$$u_T \equiv \frac{a}{\sqrt{64\pi G}} \bar{\bar{E}}_{ij}, \quad (7.168)$$

$$z_T = a, \quad (7.169)$$

one obtains¹¹

$$\ddot{u}_T + \left(k^2 - \frac{\ddot{z}_T}{z_T} \right) u_T = 0. \quad (7.170)$$

At this stage, we have a unified way of describing the scalar and tensor perturbations. What remains to do is first to find a solution for these equations (and in particular their asymptotic form), and then the initial condition for the perturbations. When this is done, we will compute the power spectrum of these perturbations.

Large scale solution

Let us first consider the asymptotic solution for this equation. During an accelerated phase of expansion, any physical length exits the Hubble radius at some stage: during inflation the ratio k/H decreases with time. Therefore at late times, the evolution equation of the scalar perturbations becomes

$$\ddot{u}_S - \frac{\ddot{z}_S}{z_S} u_S = 0. \quad (7.171)$$

¹¹The justification for the presence of the factor $1/\sqrt{64\pi G}$ will be given later.

The solution of this equation is

$$u_S = z_S \left(C_1 + C_2 \int_{-\infty}^{\eta} \frac{d\eta'}{z_S^2} \right), \quad (7.172)$$

where the lower bound on the integral in front of C_2 is chosen such that any constant term that would arise is absorbed by constant C_1 . Given the shape of the integral, it is easy to see that the term proportional to C_1 is a growing term, and the term proportional to C_2 is a decaying term. Since we are interested in the late time behavior of the solution, we shall focus on the C_1 part of the solution. Using Eq. (7.166), we have

$$V_S = C_1 \int_{-\infty}^{\eta} z_S^2 d\eta'. \quad (7.173)$$

(Note that using Eq. (7.167) is not possible as we have neglected terms proportional to k^2 in solution (7.172).) This can in fact more cleverly be rewritten

$$\Phi = C_1 \frac{H}{a^2} \int_{-\infty}^{\eta} a^2 \left(1 - \frac{\dot{H}}{H^2} \right) d\eta. \quad (7.174)$$

Although this solution has been derived in the case of a scalar field, one can in fact show that it is valid as long as there are no large scale anisotropic stresses. If we suppose the Universe undergoes a series of eras of constant w , then the above equation can be solved in the hypothesis where the last era of constant w lasts much longer than the previous ones. One obtains

$$\Phi = C_1 \frac{3+3w}{5+3w}, \quad (7.175)$$

where we have taken the w of the last and longest era. In particular, during the matter era, one has $\Phi = \frac{3}{5}C_1$. Note that Φ is much larger in the matter era than during inflation because of the $3+3w$ term which is then very small. Strictly speaking, this does not mean that there is any “amplification” of the cosmological perturbations between inflation and the matter era: had we considered the density contrasts instead of the gravitational field, we would have seen that they remain constant on large scales. This is therefore simply the fact that the relation between the density contrast and the gravitational field is complicated on large scales (and significantly different from what happens in Newtonian physics¹²).

¹²Again, it is always possible to define a gauge where the 00 part of the Einstein equations reduces exactly to the Poisson equation, but at the cost of drastically modifying the mass conservation and the Euler equations.

6. Basics of quantum field theory

So far, we have been able to compute the large scale amplitude of the cosmological perturbations as a function of some quantity C_1 which now has to be fixed by the initial conditions, which in our context are given by quantum field theory.

As a warm-up, let us recall the formalism to describe a particle evolving in a potential V . One describes the particle through the action S and the Lagrangian L :

$$S = \int L dt, \quad (7.176)$$

$$L = \frac{1}{2}m\dot{\mathbf{x}}^2 - V(\mathbf{x}). \quad (7.177)$$

The equation of motion of the particle are deduced by varying the action with respect to the position \mathbf{x} :

$$\delta S = \int \left(m\dot{\mathbf{x}} \cdot \delta \dot{\mathbf{x}} - \nabla V \cdot \delta \mathbf{x} \right) dt \quad (7.178)$$

$$= - \int \delta \mathbf{x} dt (m\ddot{\mathbf{x}} + \nabla V), \quad (7.179)$$

after some integration by parts. The equation of motion are found by imposing that the action is minimized when varying $\delta \mathbf{x}$, which gives the expected $m\ddot{\mathbf{x}} = -\nabla V$.

From the Lagrangian, one can define the conjugate momentum \mathbf{p} and the Hamiltonian through the Legendre transform

$$\mathbf{p} = \frac{\partial L}{\partial \dot{\mathbf{x}}}, \quad (7.180)$$

$$H = \mathbf{p} \cdot \dot{\mathbf{x}} - L = \frac{1}{2} \frac{\mathbf{p}^2}{m} + V. \quad (7.181)$$

When quantizing the theory, the variable \mathbf{x} and its conjugate momentum \mathbf{p} become operator obeying the commutation rule

$$[\hat{\mathbf{x}}, \hat{\mathbf{p}}] = i, \quad (7.182)$$

and their evolution is given by

$$\dot{\hat{\mathbf{x}}} = -i[\hat{\mathbf{x}}, \hat{\mathbf{H}}], \quad (7.183)$$

$$\dot{\hat{\mathbf{p}}} = -i[\hat{\mathbf{p}}, \hat{\mathbf{H}}]. \quad (7.184)$$

One can do the same thing for a scalar field, except that since the field is defined everywhere in space, one uses the Lagrangian density L ,

$$S = \int L d^4x, \quad (7.185)$$

$$L = \left(\frac{1}{2} D_\mu \phi D^\mu \phi - V \right) \sqrt{|g|}. \quad (7.186)$$

In the present case, one can start from the equation of motion of the field (that we express in term of the variable u_S if it is the scalar field or in term of the variable u_T if it is the gravitational waves)

$$\ddot{u} + \left(k^2 - \frac{\ddot{z}}{z} \right) u = 0. \quad (7.187)$$

Note that since we start from the Fourier components of the field, a few differences arise as compared to the study of a field in real space. The Lagrangian density is given by

$$L = \frac{1}{2} \dot{u} \dot{u}^* - \left(k^2 - \frac{\ddot{z}}{z} \right) u u^*. \quad (7.188)$$

Because of the complex nature of the variable u , the conjugate momentum v is given by¹³

$$v = \left(\frac{\partial L}{\partial \dot{u}} \right)^*. \quad (7.189)$$

Also, since the field is complex, any operator must follow the constraint

$$\hat{A}_k = \hat{A}_{-k}^+. \quad (7.190)$$

When quantizing the theory, the commutation relations for the complex field \hat{u} become

$$[\hat{u}_k, \hat{v}_{k'}] = i\delta(\mathbf{k} + \mathbf{k}'), \quad (7.191)$$

$$[\hat{u}_k, \hat{v}_{k'}^+] = i\delta(\mathbf{k} - \mathbf{k}'), \quad (7.192)$$

When quantizing the theory, one can decompose the (time dependent) operator \hat{u}_k into a sum of time independent operators and time dependent functions:

$$\hat{u}_k = \hat{a}_k u_k(\eta) + \hat{b}_k u_k^*(\eta), \quad (7.193)$$

where \hat{a}_k and \hat{b}_k are time independent operators, and where u_k is a complex function. With these conventions, one can easily see that $\hat{b}_k = \hat{a}_{-k}^+$. One can

¹³In order to derive this result, it suffices to notice that the theory can be expressed into the form of two real fields u^R and u^I such that $u = \frac{1}{\sqrt{2}}(u^R + iu^I)$ and to rewrite the corresponding action.

also show that the function u_k follows the classical equation of motion (and hence depends only on the modulus of \mathbf{k}). If, without loss of generality we impose that the operator \hat{a}_k follows the usual commutation rule

$$[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}^+] = \delta(\mathbf{k} - \mathbf{k}'), \quad (7.194)$$

then the amplitude of the function u_k is given by

$$u_k \dot{u}_k^* - u_k^* \dot{u}_k = i. \quad (7.195)$$

Moreover, when the term $\ddot{\frac{z}{z}}$ is negligible, the solution for u_k is under the form

$$u_k = A(k)e^{ik\eta} + B(k)e^{-ik\eta}. \quad (7.196)$$

Quantum field theory tells that one can only use solution with positive energy, defined as

$$i \frac{\partial}{\partial \eta} u_k = E u_k, \quad (7.197)$$

so that $A(k) = 0$. With the normalization (7.196), this gives

$$u_k = \frac{1}{\sqrt{2k}} e^{-ik\eta} e^{i\varphi}, \quad (7.198)$$

Where φ is an arbitrary and unimportant phase. Thus, we see that the normalization of the mode is given by the commutator (7.182) and the normalization of the Lagrangian. The justification of the prefactor of Eq. (7.168) comes from the fact that the Lagrangians for u_S and u_T can be found by varying at second order the Lagrangians for the scalar field and of the gravitational field, respectively. The meaning of the amplitude u_k here is that it corresponds to the variance of the quantities u_S, u_T .

7. Perturbation spectrum

In order to finish the calculation, let us assume we are in the almost de Sitter case, that is, that $w \sim -1$. In this case, one has by definition

$$\frac{\ddot{z}_T}{z_T} = \frac{2}{\eta^2}. \quad (7.199)$$

Moreover, when $w \sim -1$, one also has $H \sim \text{Const}$, and $\partial_t \phi \sim \text{Const}$ (because $H^2 \propto \rho_\phi \sim V$ and $w \sim -1 \sim \text{Const}$ implies that the kinetic term is also constant, being proportional to the potential term). Therefore we have also

$$\frac{\ddot{z}_S}{z_S} = \frac{2}{\eta^2}. \quad (7.200)$$

Thus we have

$$\ddot{u}_X + \left(k^2 - \frac{2}{\eta^2} \right) u_X = 0. \quad (7.201)$$

There exists an exact solution to this equation that can be given in term of Haenkel functions, which here can be expressed analytically

$$u_X = -A(k)\sqrt{\frac{2}{\pi k}} \left(1 - \frac{i}{k\eta} \right) e^{-ik\eta} - B(k)\sqrt{\frac{2}{\pi k}} \left(1 + \frac{i}{k\eta} \right) e^{ik\eta}. \quad (7.202)$$

In order to fix the integration constants $A(k)$ and $B(k)$, we shall impose that the usual Minkowskian solution is valid at early times, i.e., when $k\eta \gg 1$ (when the Hubble parameter is negligible compared to the mode frequency). This gives

$$B(k) = 0, \quad (7.203)$$

$$A(k) = -e^{i\varphi} \frac{\sqrt{\pi}}{2}, \quad (7.204)$$

so that the solution is finally

$$u_X = e^{i\varphi} \sqrt{\frac{1}{2k}} \left(1 - \frac{i}{k\eta} \right) e^{-ik\eta} \quad (7.205)$$

For the scalar modes, we want to relate this exact solution to the asymptotic form

$$u_S = C_1 z_S = C_1 a \frac{\dot{\phi}}{H} \quad (7.206)$$

when $k\eta \rightarrow 0$. This gives

$$|C_1|^2 = \frac{1}{2k^3} \frac{(8\pi G V_k)^3}{3V_k'^2} = \frac{1}{k^3} \frac{16\pi^2 G^2 V_k}{3\epsilon_k}. \quad (7.207)$$

Although we are in the quasi de Sitter case, all quantities such that H, V slowly vary with time. Therefore, for a given mode, we evaluate them at the epoch when the mode under consideration exits the Hubble radius. This is what the subscript k means in the above equation. Therefore, we obtain the power spectrum of the Bardeen potential in the matter era

$$k^3 \Phi_{\text{mat}}^2 = \frac{48\pi^2 G^2 V_k}{25\epsilon_k}. \quad (7.208)$$

Equivalently, we have for the tensor modes

$$k^3 \bar{E}_{ij} \bar{E}^{ij} = \frac{256\pi^2 G^2 V_k}{3}. \quad (7.209)$$

Note that the amplitude of the tensor modes directly gives the energy scale of inflation, contrarily to the amplitude of the scalar modes. Comparing the scalar and tensor modes, we recover the well known formula

$$\frac{k^3 \bar{\bar{E}}_{ij} \bar{\bar{E}}^{ij}}{k^3 \Phi_{\text{mat}}^2} = \frac{400 \epsilon_k}{9}, \quad (7.210)$$

which, after some manipulations can also be expressed into a ratio between the scalar and the tensor part of the C_ℓ (however the formula is far from accurate, especially when there is a cosmological constant). At this point, we should also prove that although the fluctuations are of quantum origin, they behave classically today, so that the amplitude we just computed corresponds to the variance of the perturbations whom our universe is a realization. This subtle point, known as the quantum-to-classical transition, would deserve a careful analysis, so we will simply refer the reader to some technical papers (Guth & Pi 1985 ; Polarski & Starobinsky 1985; Lesgourges et al. 1997; Starobinsky 1998; Y 1998).

The k -dependence of ϵ and V translate into the fact that both spectra are not exactly proportional to k^{-3} . In general one can approximate the spectra (at least for a limited range of wavenumbers) as power laws close to k^{-3} , so that it is convenient to define the spectral index of the scalar and tensor perturbations as

$$k^3 |\Phi_{\text{mat}}|^2 = A_S k^{n_S - 1}, \quad (7.211)$$

$$k^3 \bar{\bar{E}}_{ij} \bar{\bar{E}}^{ij} = A_T k^{n_T}, \quad (7.212)$$

with

$$n_S \sim 1 - 4\epsilon_k + 2\delta_k, \quad (7.213)$$

$$n_T \sim -2\epsilon_k, \quad (7.214)$$

so that $n_S \sim 1$ and $n_T \sim 0$. At higher order in the slow-roll parameters, one also expects deviations from the pure power-law spectra. This is usually referred to as the “running” of the spectral index ($dn_S/d \ln k \neq 0$).

However for a specific inflationary model, the four observable quantities A_S , A_T , n_S and n_T can be expressed at lowest order in term of the physical quantities V_k and the two slow-roll parameters ϵ_k and δ_k , so that there exist some consistency relations (between the tensor spectral index and the scalar-to-tensor ratio), which in principle allow to test inflation, and to reconstruct the potential on a small region (as we can have access to V as well as V' and V'' with the slow roll parameters). Note however that this consistency relation crucially relies on the detection of the tensor modes (and hence, on the B -polarization of CMB as it is probably the most efficient way to detect the tensor modes), which may very well be an extraordinarily difficult task if ϵ_k happens to be very small.

8. Conclusion

The study of the inflationary perturbations that we have presented here is of course far from complete. We did not derive the calculation of the spectral indices as a function of the slow-roll parameters, nor of the running of the spectral indices. However we hope to have made it clear that this can be done analytically within the slow-roll approximation (as well as numerically, of course). Let us emphasize that many alternative models such as the ekpyrotic universe or the pre Big-Bang scenario can also be studied within this framework, as the core ingredient (quantum fluctuations which are expelled from the Hubble radius) are present. What changes is the matter content of the Universe during this phase as well as the dynamics of the scale factor.

Also many epicycles are possible: one could consider inflation with several scalar fields, which usually produce isocurvature fluctuations, inflation with a non smooth potential, in which case the slow roll conditions may not be satisfied at some epochs, which then translates into features in the perturbation spectrum (step-like, or with a break in the spectra index), etc.

References

- A.H. Guth 1981, Phys. Rev. D 23, 347
- A. Vilenkin & L. Ford, Phys. Rev. D 26, 1231
- A.D. Linde 1982, Phys. Lett. 116B, 335
- A.A. Starobinsky 1982, Phys. Lett. 117B, 175
- A.D. Linde, "Particle Physics and Inflationary Cosmology" 1990, Harwood, Chur, Switzerland
- A.R. Liddle & D.H. Lyth, "Cosmological inflation and large-scale structure" 2000, Cambridge University Press, Cambridge, England
- J. Lesgourgues, "Modèles inflationnaires et anisotropies du fond diffus" 1998, Thèse de Doctorat de l'Université de Tours
- D. Langlois, "Sur les inhomogénéités en cosmologie et les fluides relativistes" 1994, Thèse de Doctorat de l'Université Paris VI
- D.J. Schwarz, C.A. Terro-Escalante & A.A. Garcia 2001, Phys. Lett. 517B, 243
- A.D. Linde 1985, Phys. Lett. 129B, 177
- A. Guth & S.Y. Pi 1985, Phys. Rev. D **32**, 1899
- D. Polarski & A.A. Starobinsky 1996, Class. Quant. Grav. 13, 377
- J. Lesgourgues, D. Polarski & A.A. Starobinsky 1997, Nucl. Phys. B497, 479
C. Kiefer, D. Polarski &
- A.A. Starobinsky 1998, Int. J. Mod. Phys. D7, 455
- C. Kiefer, J. Lesgourgues, D. Polarski & A.A. Starobinsky, Class. Quant. Grav. 15, L67

Chapter 9

CMB OBSERVATIONAL TECHNIQUES AND RECENT RESULTS

Edward L. Wright

*UCLA Astronomy
PO Box 951562
Los Angeles, CA 90095-1562
USA
wright@astro.ucla.edu*

Abstract The Cosmic Microwave Background (CMB) consists of photons that were last created about 2 months after the Big Bang, and last scattered about 380,000 years after the Big Bang. The spectrum of the CMB is very close to a blackbody at 2.725 K, and upper limits on any deviations from the CMB from a blackbody place strong constraints on energy transfer between the CMB and matter at all redshifts less than 2 million. The CMB is very nearly isotropic, but a dipole anisotropy of $\pm 3.346(17)$ mK shows that the Solar System barycenter is moving at 368 ± 2 km/sec relative to the observable Universe. The dipole corresponds to a spherical harmonic index $\ell = 1$. The higher indices $\ell \geq 2$ indicate intrinsic inhomogeneities in the Universe that existed at the time of last scattering. While the photons have traveled freely only since the time of last scattering, the inhomogeneities traced by the CMB photons have been in place since the inflationary epoch only 10^{-35} sec after the Big Bang. These intrinsic anisotropies are much smaller in amplitude than the dipole anisotropy, with $\Delta T \leq 100 \mu\text{K}$. Electron scattering of the anisotropic radiation field produces an anisotropic linear polarization in the CMB with amplitudes $\leq 5 \mu\text{K}$. Detailed studies of the angular power spectrum of the temperature and linear polarization anisotropies have yielded precise values for many cosmological parameters. This paper will discuss the techniques necessary to measure signals that are 100 million times smaller than the emission from the instrument and briefly describe results from experiments up to *WMAP*.

Keywords: Cosmic microwave background, instrumentation.

1. Introduction

The Cosmic Microwave Background (CMB) was first seen via its effect on the interstellar CN radical (Adams, 1941) but the significance of this datum was not realized until after 1965 (Thaddeus, 1972; Kaiser and Wright, 1990). In fact, Herzberg, 1950 calculated a 2.3 K excitation temperature for the CN transition and said it had “of course only a very restricted meaning.” Later work by Roth et al., 1993 obtained a value for $T_o = 2.729^{+0.023}_{-0.031}$ K at the CN 1-0 wavelength of 2.64 mm which is still remarkably accurate.

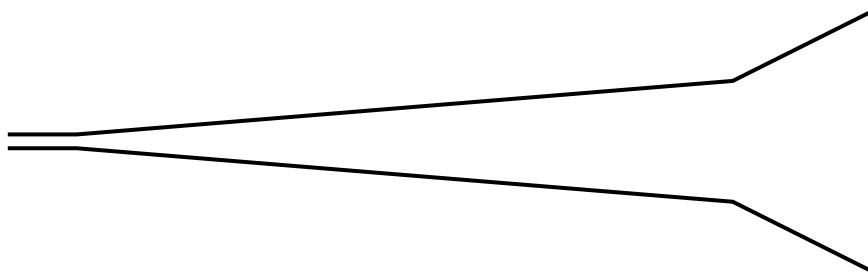


Figure 9.1. The flared horn used by Dicke during World War II.

A second notable missed opportunity to discover the CMB occurred during World War II radar research. In a paper reporting measurements of the atmospheric opacity at 1.25 cm wavelength using zenith angle scans, an upper limit of $T_o < 20$ K is given (Dicke et al., 1946). The Dicke switch and the differential radiometer were invented for this work. Since the reference load was at room temperature, the large difference signal of 250 K at the zenith did not allow for a precise determination of the cosmic temperature. The antenna was a flared horn (Figure 9.1) which was specifically designed for low sidelobes. This missed opportunity is especially ironic since Dicke was actually building a radiometer to look for the CMB when he heard about the Penzias and Wilson, 1965 result (Dicke et al., 1965).

The most accurate measurements of the CMB spectrum to date have come from the Far InfraRed Absolute Spectrophotometer (FIRAS) on the COsmic Background Explorer (COBE) (Boggess et al., 1992). In contradiction to its name, FIRAS was a fully differential spectrograph that only measured the difference between the sky and an internal reference source that was very nearly a blackbody. Figure 9.2 shows the interferograms observed by FIRAS for the sky and for the external calibrator (XC) at three different temperatures, all taken with the internal calibrator (IC) at 2.759 K. Data from the entire FIRAS dataset show that the rms deviation from a blackbody is only 50 parts per million of the peak I_ν of the blackbody (Fixsen et al., 1996) and a recalibration of the thermometers on the external calibrator yield a blackbody temperature of

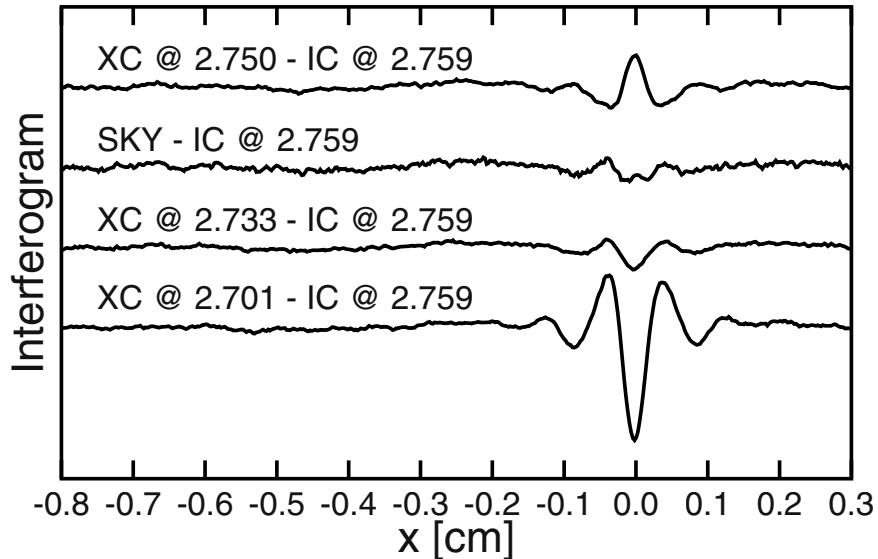


Figure 9.2. Four interferograms taken by FIRAS. Three show observations of the external calibrator at different temperatures straddling the sky temperature, while the fourth shows observations of the sky. All were taken with the internal calibrator at 2.759 K.

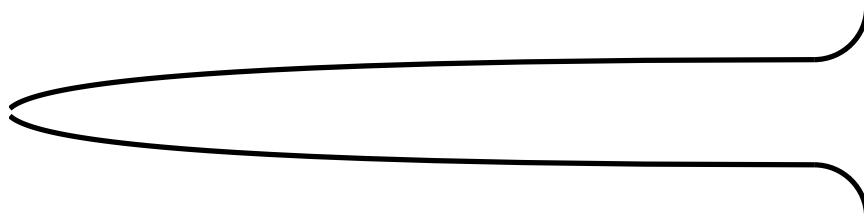


Figure 9.3. The FIRAS horn: a compound parabolic concentrator with a trumpet bell flare to reduce sidelobes.

2.725 ± 0.001 K (Mather et al., 1999). FIRAS also had a flared horn to reduce sidelobes as seen in Figure 9.3.

Shortly after the Cosmic Microwave Background (CMB) was discovered, the first anisotropy in the CMB was seen: the dipole pattern due to the motion of the observer relative to the rest of the Universe (Conklin, 1969). After confirmation by Henry, 1971 and by Corey and Wilkinson, 1976 the fourth “discovery” of the dipole (Smoot et al., 1977) showed a very definite cosine pattern as expected for a Doppler effect, and placed an upper limit on any further variations in T_{CMB} . Further improvements in the measurement of the dipole anisotropy were made by the Differential Microwave Radiometers (DMR) experiment on COBE (Bennett et al., 1996 and by the Wilkinson Microwave

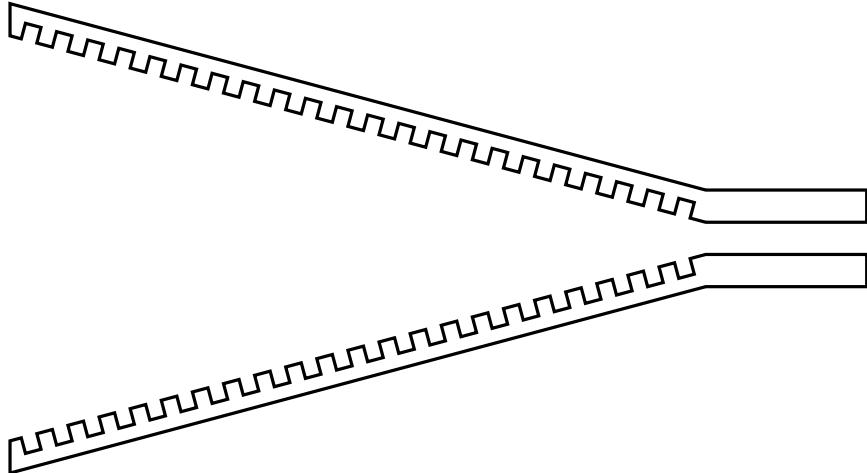


Figure 9.4. Cross-section of corrugated horn. The grooves in the walls of the horn act like shorted $\lambda/4$ stubs, and as such appear like open circuits to the waves propagating in the horn. As a result the modes in the horn are very similar to free-space TEM modes which minimizes the discontinuity at the edge of the horn, and hence minimizes the sidelobes.

Anisotropy Probe (Bennett et al., 2003b). Both the DMR and *WMAP* use corrugated horns to reduce sidelobes, as shown in figure 9.4. Everyone of these experiments used a differential radiometer which measured the difference between two widely separated spots on the sky.

Experiments to measure smaller angular scales use radio telescopes with dishes to make a beam with a smaller angular spread than a horn. Horns would be used to feed the dishes. A large edge taper should be used to avoid having the beam from the feed spill over the edge of the dish, as seen in Figure 9.5. Usually there is stuff behind the dish that one would rather not look at, such as the ground or the thermal radiator system in *WMAP*. Figure 9.5 illustrates a Gaussian illumination of the primary with the edge of the dish at the 2σ point, which corresponds to an edge taper of e^{-2} or about 9 db. Figure 9.6 shows that sidelobes of a circular aperture with a Gaussian illumination pattern get much smaller for increasing edge taper, and that the angular resolution only declines slightly. *WMAP* used edge tapers of 13 to 21 db, although the illumination patterns were not symmetric or Gaussian (Page et al., 2003b).

The first theoretical predictions of $\Delta T/T = 10^{-2}$ (Sachs and Wolfe, 1967) and $\Delta T/T = 10^{-3.5}$ (Silk, 1968) were superseded by predictions based on cold dark matter (Peebles, 1982, Bond and Efstathiou, 1987). These CDM predictions were consistent with the small anisotropy seen by COBE and furthermore predicted a large peak at a particular angular scale due to acoustic oscillations in the baryon/photon fluid prior to recombination. The position

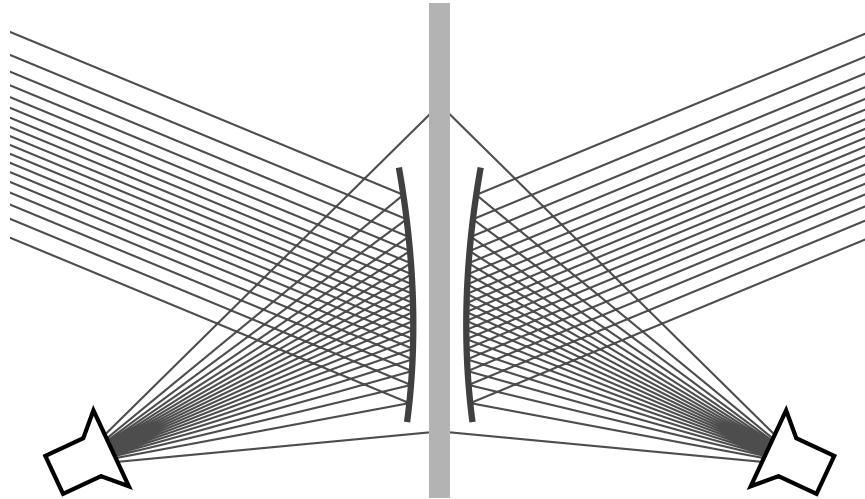


Figure 9.5. A differential radiometer using back-to-back off-axis paraboloidal dishes. The feeds are designed to mainly illuminate the center of the dishes with minimal spillover past the edges.

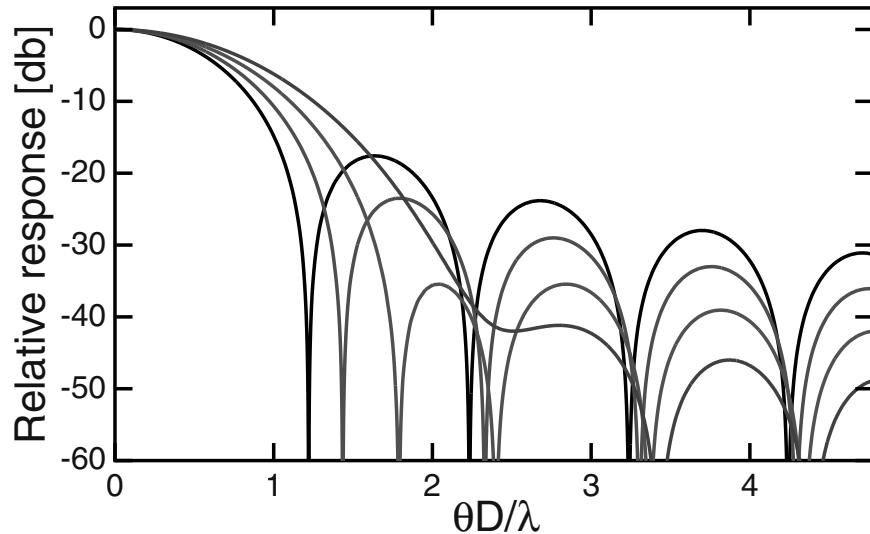


Figure 9.6. The point spread function of a circular aperture for 4 different values of the edge taper with Gaussian illumination. The four curves are for uniform illumination or 0 db taper, 9, 18 and 27 db taper. The sidelobe level decreases with increasing taper, while the width of the main beam increases slightly.

of this big peak and other peaks in the angular power spectrum of the CMB anisotropy depends on a combination of the density parameter Ω_m and the

vacuum energy density Ω_V , so this peak provides a means to determine the density of the Universe (Jungman et al., 1996). A tentative detection of the big peak at the position predicted for a flat Universe had been made by 1994 (Scott et al., 1995). The peak was localized to $\ell_{pk} = 229 \pm 8.5$ (Knox and Page, 2000) by the beginning of 2000. Later the BOOMERanG group claimed to have made a dramatic improvement in this datum to $\ell_{pk} = 197 \pm 6$ (de Bernardis et al., 2000). This smaller value for ℓ_{pk} favored a moderately closed model for the Universe. But improved data on the peak position from *WMAP* (Page et al., 2003c) gives $\ell_{pk} = 220.1 \pm 0.8$ which is consistent with a flat Λ CDM model.

Polarization of the CMB was shown to be $< 300 \mu\text{K}$ (Lubin and Smoot, 1981). This observation used a differential polarimeter that was only sensitive to linear polarization. COBE put a limit of $< 15 \mu\text{K}$ on the polarization anisotropy. The linear polarization of the CMB was first detected by DASI POL (Kovac et al., 2002), and the cross-correlation of the temperature and polarization anisotropies was confirmed by *WMAP* (Kogut et al., 2003).

The detected polarization level is an order of magnitude lower than the anisotropy. The observed polarization is caused by electron scattering during the late stages of recombination on small angular scales and after reionization on large angular scales. The magnitude of the polarization on small angular scales depends on the anisotropy being in place at recombination, as is the case for primordial adiabatic perturbations but not for topological defects; the electron scattering cross-section; and the recombination coefficient of hydrogen. The detection of this polarization is a very strong confirmation of the standard model for CMB anisotropy.

Because polarization is a vector field, two distinct modes or patterns can arise (Kamionkowski et al., 1997, Seljak and Zaldarriaga, 1997): the gradient of a scalar field (the ‘‘E’’ mode) or the curl of a vector field (the ‘‘B’’ mode). Electron scattering only produces the E mode. Electron scattering gives a polarization pattern that is correlated with the temperature anisotropy, so the E modes can be detected by cross-correlating the polarization with the temperature. The B modes cannot be detected this way, and the predicted level of the B modes is at least another order of magnitude below the E modes, or two orders of magnitude below the temperature anisotropy.

2. Observational Techniques

The most important part of any CMB experiment is the modulation scheme that allows one to measure μK signals in the presence of $\sim 100 \text{ K}$ instrumental foregrounds. A good modulation scheme is much more important than high sensitivity, since detector noise can always be beaten down as $1/\sqrt{t}$ by integrating longer, while a systematic error is wrong forever.

Chopping

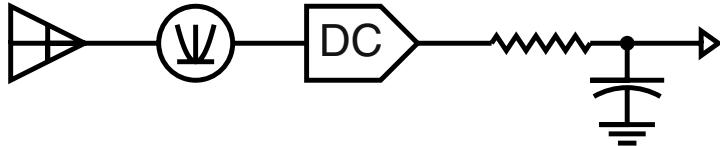


Figure 9.7. A total power radiometer with a bolometer [a square-law device indicated by the parabolic curve] feeding a DC amplifier. This design will have too much $1/f$ noise to be useful.

The first step in any modulation scheme is the chopping scheme. The instrument sketched out in Figure 9.7 will not succeed because the first stage of amplification is at zero frequency (DC), and all electronic circuits suffer from either $1/f$ noise or drifting baselines corresponding to $1/f^2$ noise. Figure 9.7 shows a bolometer detector where the radiation goes directly into a square-law device. In terms of radio engineering, this is similar to the crystal sets that were used in the 1910's. Modern bolometers running at temperatures below 0.3 K actually have enough sensitivity to make this design superior to radio frequency amplifier designs, but some form of chopping is absolutely required.

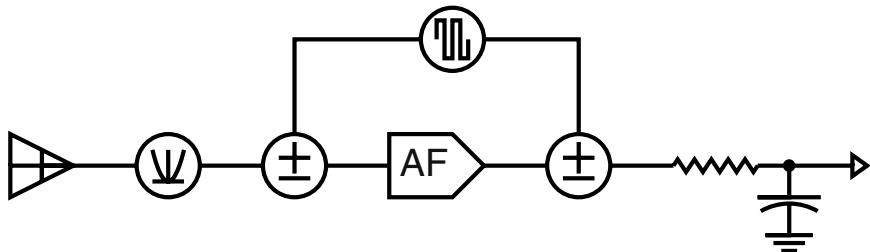


Figure 9.8. A total power radiometer with an AC-biased bolometer feeding an audio amplifier followed by a lockin amplifier. This design is used in BOOMERanG and will be used in the Planck HFI.

The least obtrusive chopping scheme involves biasing the bolometer with alternating current (AC). This is illustrated in Figure 9.8. The bias supply is connected to an audio frequency (AF) source, shown here as a square wave oscillator, and this causes the responsivity of the bolometer to change sign at an audio frequency rate. The output of the bolometer then goes through an AF amplifier and into a phase sensitive demodulator and low pass filter, or a lockin amplifier. The $1/f$ knee of an AC-biased bolometer can be lower than 0.01 Hz (Wilbanks et al., 1990). While AC bias removes the problem of $1/f$ noise due to the amplifier, there can still be $1/f$ or $1/f^2$ noise from the atmosphere or

drifting temperatures in the instrument. Thus a good scanning strategy is still needed with AC-biased bolometers.

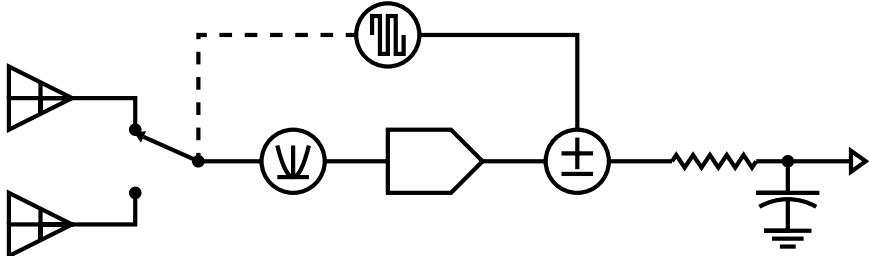


Figure 9.9. A differential radiometer with an optical chopper so the bolometer looks alternately at two sky different sky positions. The bolometer feeds an audio amplifier followed by a lockin amplifier.

A differential radiometer like the COBE DMR looks alternately at two different sky positions and measures the difference between the brightnesses at these two positions. Figure 9.9 shows a differential radiometer with a bolometric detector. This system using a chopping secondary is fairly common in infrared astronomy.

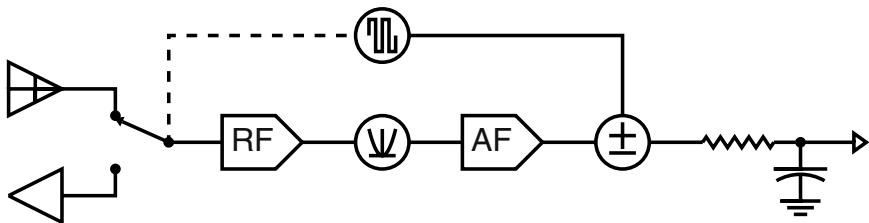


Figure 9.10. A differential radiometer chopping against a load, using an RF amplifier prior to the square-law detector. This kind of instrument is required when measuring the absolute temperature of the CMB, and was used by Penzias and Wilson, 1965.

Figure 9.10 shows a radiometer using a radio frequency (RF) amplifier that is chopping against a load. One might think that with the first stage of amplification occurring at a high frequency, chopping would not be necessary, but in practice RF amplifiers have gain fluctuations that contribute multiplicative $1/f$ noise. Chopping against a load is necessary when measuring the absolute temperature of the CMB, T_o .

In terms of antique radio technology, this radiometer with an RF amplifier leading to a square-law device is a *tuned RF* receiver which was the state-of-the art in 1929. The modern superheterodyne circuit for radio receivers with amplification and filtering at an intermediate frequency (IF) was used by the COBE DMR, but the primary advantage of a superheterodyne receiver over a

tuned RF receiver is its improved selectivity. Since the CMB is a very broad band signal, selectivity beyond that provided by a RF filter is seldom desired.

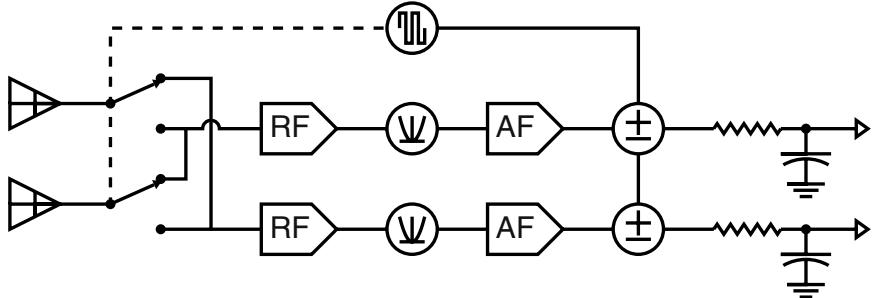


Figure 9.11. A dual channel differential radiometer where the chopper acts like a reversing switch. Practical systems with arrays like SCUBA act like this when running with chopper throws smaller than the array.

Finally one can set up a chopping system with two antennae and two amplifier chains, so that the chopper reverses the connections between the horns and the amplifiers. Figure 9.11 shows such a scheme. In reality this setup would not be very practical, but the same effect is obtained when using an array of detectors like SCUBA behind a chopper with a throw that is less than the size of the array.

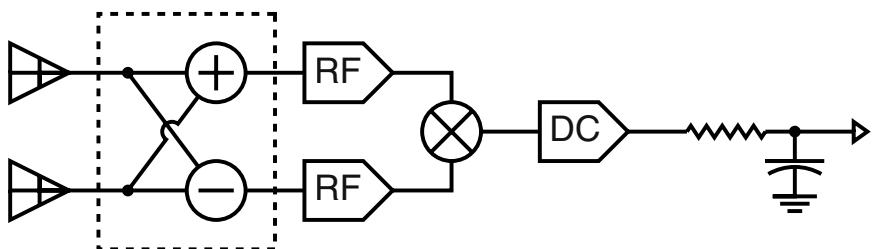


Figure 9.12. A correlation radiometer where the chopper is replaced by a hybrid circuit. Separate amplifier chains amplify the sum and difference signals which are then multiplied.

A practical microwave radiometer that has the same sensitivity as the system shown in Figure 9.11 is the correlation radiometer shown in Figure 9.12. A hybrid circuit at the input forms the sum and difference voltages $(V_A + V_B)/\sqrt{2}$ and $(V_A - V_B)/\sqrt{2}$. These are separately amplified and then multiplied, giving an output proportional to $V_A^2 - V_B^2$ which is the desired difference in the powers arriving at the two horns.

There are two practical difficulties with the correlation radiometer. The first is that one will get too much $1/f$ noise from the multiplier and the DC amplifier following it. This can be solved by introducing phase switches into both

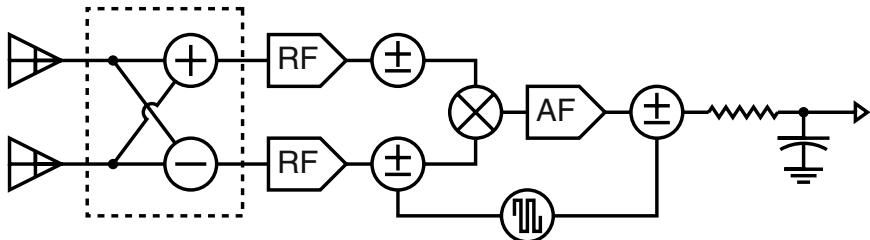


Figure 9.13. A correlation radiometer with a phase switch inserted in both arms. One of the phase switches is driven to modulate the output of the multiplier so a lockin amplifier can be used.

amplifier chains, and then toggling just one of them. This causes the sign of the output product to toggle, so the output of the multiplier can be amplified at audio frequencies and then fed to a lockin amplifier. This is shown in Figure 9.13.

The second practical difficulty is the implementation of the multiplier needed for the correlation radiometer. This multiplier has to form a product in a picosecond in order to handle the 94 GHz output of the highest frequency *WMAP* band. This is handled by using another hybrid followed by square law detectors, which corresponds to using a *quarter-square multiplier*. The ancient Egyptians used tables of squares to multiply using the formula $A \times B = [(A+B)^2 - (A-B)^2]/4$. Both *WMAP* and the Planck LFI use this technique, giving what is termed a differential pseudo-correlation radiometer.

These chopping schemes have an effect on the signal-to-noise ratio of the experiment as follows:

- Let the mapping speed of the total power, one-horned system like Figure 9.8 be 1.00. This is the system planned for the Planck HFI.
- Then the system chopping against a load like Figure 9.10 spends only 50% of its time looking at the sky, so the sky signal is $\sqrt{2}$ noisier. The noise on the load measurement is also $\sqrt{2}$ noisier because the load is observed only 50% of the time. The difference output is then 2 times noisier, which corresponds to a mapping speed of only 0.25 relative to the total power system.
- The two-horned differential radiometer like Figure 9.9 is looking at two parts of the sky at once, so it has a mapping speed of 0.5 relative to the total power system.
- The two-horned differential radiometers with two amplifier chains like Figure 9.11 or Figure 9.13 achieve a mapping speed equal to the total power system, at the expense of doubling the number of horns and amplifiers. This is the system used by *WMAP*.

- The Planck LFI is like *WMAP* but one of the horns is replaced by a load, so its mapping speed is 0.5 relative to the total power system.

Scanning

Any experiment to map N_p pixels will need to collect $N_d \geq N_p$ data points. One would like to see that a typical time history that might be produced by some systematic effect will correspond to an element in the N_d -dimensional data space that is orthogonal or nearly orthogonal to the N_p -dimensional subspace that corresponds to the time histories that can be generated by scanning a map. This can be achieved by imposing more than two distinct modulations in the experiment, since the sky is a two dimensional object. For example, the COBE DMR chopped between two beams 100 times per second, spun to interchange those beams every 73 seconds, precessed that spin axis around the circle 94° away from the Sun every 104 minutes, and then moved that circle around the sky once per year as the Earth went around the Sun. This is a four way modulation. *WMAP* chops between two beams 2500 times per second, spins to interchange those beams every 132 seconds, precesses its spin axis around a circle 157.5° from the Sun once per hour, and follows the annual motion of the Sun again giving a four way modulation.

On the other hand ARCHEOPS only scanned around a circle of constant elevation and then let the center of the circle move in right ascension as the Earth turned. This provides only a two way modulation. Since the sky itself is a two dimensional function, just about any time history of drifting baselines is consistent with some pattern on the sky. Thus ARCHEOPS is very vulnerable to striping. This can be seen in the last panel of Figure 2 of astro-ph/0310788 (Hamilton et al., 2003) which clearly shows correlated residuals aligned with the scan path. These stripes have a low enough amplitude to not interfere with measurements of the temperature-temperature angular power spectrum C_ℓ^{TT} , but they would ruin a measurement of the polarization power spectrum C_ℓ^{EE} .

Stripes are caused by small, asymmetric reference sets for pixels in the map. The reference set for the i^{th} pixel consists of the other pixels in the map that are used to establish the baseline for the i^{th} pixel. In a differential experiment like COBE or *WMAP* the reference set is the circle of radius equal to the chopper throw centered on the i^{th} pixel, or a subset of this circle. This gives a large reference set so differential experiments have nearly uncorrelated noise per pixel and thus no stripes.

For a one-horned experiment like ARCHEOPS or Planck the reference set is a line of pixels passing through the i^{th} pixel along the scan direction. The length of the reference set along the scan circle is determined by the $1/f$ knee of the output, and is of order $\pm\omega/f_{\text{knee}}$ where ω is the angular scan rate of the instrument. Observations both before and after the i^{th} pixel can be used to set

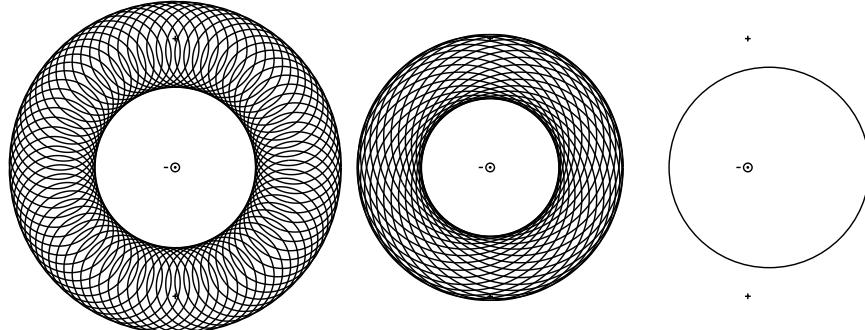


Figure 9.14. The “hourly” scan pattern of COBE, WMAP and Planck from left to right. In each panel the anti-solar direction is in the center, and plus signs denote the North and South ecliptic poles. An equiangular azimuthal projection is used. COBE scanned this area in one orbit of about 103 minutes, WMAP scans its pattern every hour, while Planck spends several hours integrating on one scan circle of radius 70° radius shown offset from anti-Sun by a 15° precession angle.

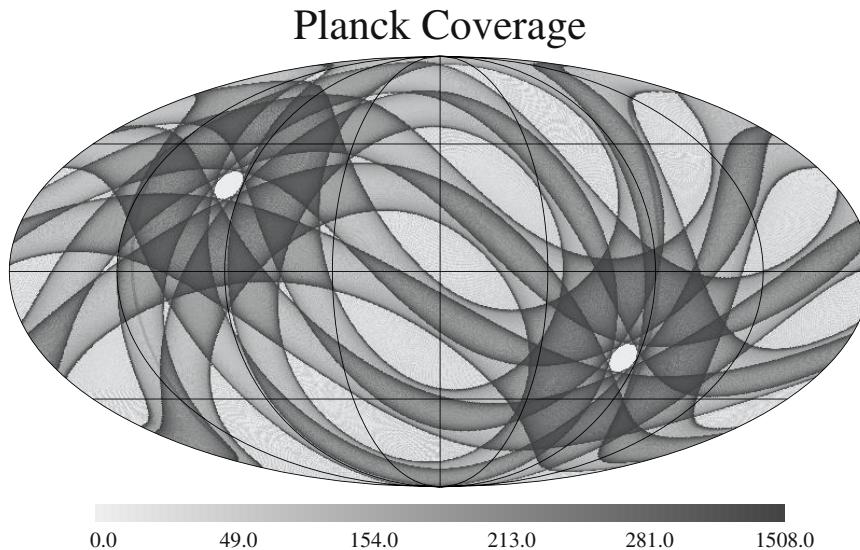


Figure 9.15. Simulated coverage by the Planck mission assuming a 70° scan radius and a 15° precession radius with 10 precession cycles per year. This plot is in galactic coordinates. Note the asymmetry between the coverage patterns in the North and south ecliptic hemispheres.

the baseline so the reference set always has inversion symmetry. A description of the minimum variance method for processing data from one-horned radiometers using a “pre-whitening” filter and time-ordered processing techniques is given by Wright, 1996. The width of the pre-whitening filter determines the length of the reference set. When several scan circles pass through

the i^{th} pixel in different directions then the reference set becomes larger and more symmetric. If scans pass through the i^{th} pixel in all directions (modulo 180° because of the inversion symmetry) then the reference set is symmetric and there are no stripes. If ω/f_{knee} is large and there is a large range of scan angles then the reference set is large and the noise per pixel is nearly uncorrelated.

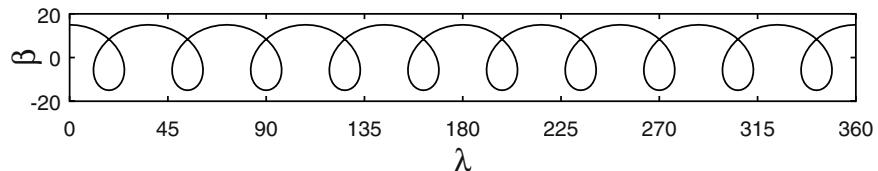


Figure 9.16. The path of the Planck scan circle center when the precession angle is 15° and the precession rate is 10 cycles per year. The difference between the Northern and Southern loops of the cycloid causes the asymmetry in the coverage pattern seen in Figure 9.15.

Figure 9.14 shows the amount of sky scanned by COBE, *WMAP* and Planck in about an hour. These scan patterns move around the sky once per year as the Earth orbits the Sun. The center of the Planck scan circle precesses around the anti-Sun slowly, with perhaps 10 turns per year. The range of scan angles through a pixel is always 180° for the COBE scan pattern. For the *WMAP* pattern the region near the ecliptic only sees a range of 45° in scan angles, while the median range of scan angles is 95° . The median range of scan angles is only 51° for Planck.

The range of scan angles through a pixel is also crucial in determining the ability of a given experiment to make reliable measurements of polarization. Observing the same pixel with different orientations of the instrumental axes provides the data needed to separate true celestial polarizations from instrumental effects.

The simulated Planck coverage map created while determining the range of scan angles, shown in Figure 9.15, illustrates an interesting asymmetry that is inherent in this mission's planned slow precession. If Planck precesses only 10 times per year on a 15° radius circle, then the scan circle motion due to precession is only 2.5 times higher than the rate due to the annual motion. Thus in one ecliptic hemisphere the net motion is 3.5 times the annual rate while in the opposite hemisphere the net motion is only 1.5 times the annual rate. This results in an asymmetric path for the scan circle center shown in Figure 9.16. This asymmetry will certainly make testing any North-South asymmetry (Eriksen et al., 2003) much more difficult. Planck would have a much better scanning strategy if the precession rate were close to the geometric mean of the spin period and one year. This would be about 10 hours, or several hundred precession periods per year.

Frequency Range

One time history that will always be consistent with a pattern on the sky is obtained by scanning over the Milky Way. The only way to make this nearly orthogonal to a true CMB pattern on the sky is to observe a large range of frequencies. The spectrum of the Milky Way on large angular scales as measured by FIRAS is given in Wright et al., 1991. The ratio between the CMB anisotropy signal and this galactic spectrum peaks at 72 GHz. For higher frequencies the rising thermal dust emission spectrum starts to dominate over the CMB signal. At frequencies lower than 72 GHz the galactic foreground is dominated by free-free and synchrotron emission. An experiment to measure the primary CMB anisotropy would like to observe a range of frequencies covering \pm a factor of three from 72 GHz, or from 24 GHz to 216 GHz. But the thermal Sunyaev-Zeldovich effect goes through zero at about 220 GHz so extending the high frequency limit to 400 GHz is clearly a good idea. The *WMAP* mission only covers the peak and the low frequency side of the peak in the CMB:galaxy ratio, while the *Planck* mission will extend the high frequency coverage to more than 800 GHz.

Sensitivity

Once a good chopping and scanning strategy is planned, a detector system with enough sensitivity to map the CMB anisotropy is needed. The primary anisotropy of the CMB extends up to $\ell \approx 2000$ so there are about 4 million spots on the sky that need to be measured. The anisotropy is about $38 \mu\text{K}$ in each spot so the integrated “monopole” sensitivity, $\sigma_{pix}/\sqrt{N_{pix}}$, needs to be about 19 nK in order to reach a signal-to-noise ratio of 1 per spot on the primary anisotropy. A SNR of 1 marks the “point of diminishing returns” when measuring the variance of a Gaussian signal. When the SNR per pixel is < 1 then the error on C_ℓ improves like one over the integration time, while when the SNR per pixel is > 1 then the error on C_ℓ is limited by cosmic variance and does not improve at all with increased integration time. But to measure E-mode polarization one would like 10 times more sensitivity, and to measure the B-mode polarization one would like at least 100 times more sensitivity.

WMAP will achieve a monopole sensitivity of 23 nK in 4 years which is well into the region of diminishing returns on C_ℓ^{TT} for the $\ell < 900$ range compatible with the *WMAP* angular resolution. But the *WMAP* sensitivity is far from the point of diminishing returns for polarization measurements.

How can one reach these sensitivity goals? The goal of a monopole sensitivity of 19 nK can be achieved in one year with a sensitivity of $107 \mu\text{K}$ in one second. With a bandwidth of 18 GHz (25% of the 72 GHz optimal frequency) the system temperature requirement is $T_{sys} = 14 \text{ K}$ for a single radiometer channel, using the Dicke radiometer equation $\Delta T = T_{sys}/\sqrt{Bt}$. The best cur-

rent performance of High Electron Mobility Transistor (HEMT) amplifiers is about 0.3 K/GHz for cryogenic HEMTs, which is a bit too high. Hence an experiment designed to map the whole sky to the point of diminishing returns for C_ℓ^{TT} would need to have at least two channels. *WMAP* has 20 channels with two polarizations on each of the 10 differencing assemblies, but only achieves 1.5 K/GHz with passively cooled HEMTs at running at ≈ 90 K. However, the absence of expendable cryogens allows *WMAP* to operate for several years and easily surpass the sensitivity goal.

The $1/f$ gain fluctuations in HEMTs require a high chopping frequency. Prior to the lockin amplifier, the variance of the output of a HEMT radiometer in a 1 Hz bandwidth ($t_{int} = 0.5$ sec) centered at f is given by

$$\text{var}(\Delta T) = \frac{2T_{sys}^2}{B} + \left(\frac{\Delta G}{G}(f) \right)^2 T_{sys}^2 \quad (9.1)$$

The $1/f$ gain fluctuations are given by

$$\left(\frac{\Delta G}{G}(f) \right)^2 = b^2 \left(\frac{1 \text{ Hz}}{f} \right)^\alpha \quad (9.2)$$

Typically $\alpha \approx 1$ and $b = 10^{-5}/\sqrt{\text{Hz}}$ for warm HEMTs and $10^{-4}/\sqrt{\text{Hz}}$ for cryogenic HEMTS, and the bandwidth B is 10's of GHz, so the $1/f$ knee frequency is

$$f_{knee} = (b^2 B / 2)^{1/\alpha} \quad (9.3)$$

which ranges from 20 to 1000 Hz for the *WMAP* radiometers (Jarosik et al., 2003b). The chopping frequency f_c must be higher than this to avoid excess noise due to gain fluctuations.

The post-lockin noise variance in 1 Hz centered at f is

$$\text{var}(\Delta T) = 4 T_{sys}^2 \left[\frac{2}{B} + \left(\frac{\Delta G}{G}(f_c) \right)^2 \right] + T_{off}^2 \left(\frac{\Delta G}{G}(f) \right)^2 + \Delta T_{off}(f)^2 \quad (9.4)$$

which still shows $1/f$ noise due to gain fluctuations but they are only driven by the imbalance in the radiometer, T_{off} . The factor of “4” in front of T_{sys}^2 is the increased noise due to chopping. If $T_{off} \ll T_{sys}$ then the knee frequency is much lower:

$$f'_{knee} = \left(\frac{b^2 B T_{off}^2}{8 T_{sys}^2 [1 + (f_{knee}/f_c)^\alpha]} \right)^{1/\alpha}. \quad (9.5)$$

This is 0.04 Hz for the worst case *WMAP* radiometer, W4, which has both the highest bandwidth and the highest offset. Ideally the post-lockin $1/f$ knee

frequency f'_{knee} should be lower than all the scan frequencies but the *WMAP* spin frequency is only 0.008 Hz so this ideal was not achieved for the W4. $\Delta T_{off}(f)^2$ is the power spectrum of the offset drifts which typically show $1/f^2$ behavior and dominate the noise at very low frequencies.

Note that the quantum limit on coherent receivers is $0.5h\nu/k$ or 0.024 K/GHz (Wright, 1999). This corresponds to 0.5 photons per mode. The 0.3 K/GHz for cryogenic HEMTs is about 7 photons/mode. At 72 GHz the CMB has only $\bar{n} = 0.4$ photons per mode, where $\bar{n} = (\exp[h\nu/kT] - 1)^{-1}$ is the mean number of photons per mode. Thus a background-limited incoherent detector could be much more sensitive than a coherent radiometer using HEMTs.

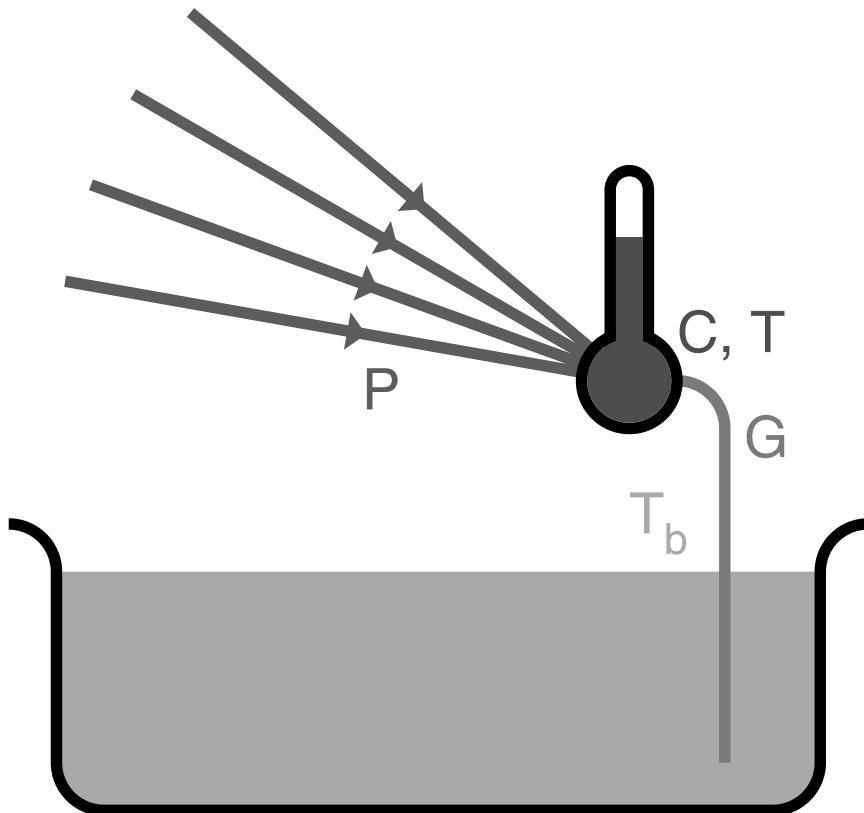


Figure 9.17. A cartoon representation of a bolometer: a small thermometer at the focus of a beam of radiation, thermally linked to a bath.

Consider a bolometric radiometer with an 18 GHz at 72 GHz with a diffraction-limited throughput, $A\Omega = \lambda^2$. A background limited (BLIP) system would have a temperature sensitivity of $19 \mu\text{K}$ in 1 second. This is already 6 times better than the $107 \mu\text{K}$ in one second needed to reach the point of diminishing

returns for C_ℓ^{TT} with a single channel. However, this bolometer would have to have a noise equivalent power (NEP) less than $7 \times 10^{-18} \text{ W}/\sqrt{\text{Hz}}$, which is still difficult to achieve. The required NEPs are higher and thus easier to achieve at higher frequencies so bolometers are definitely the technology of choice for frequencies higher than 94 GHz.

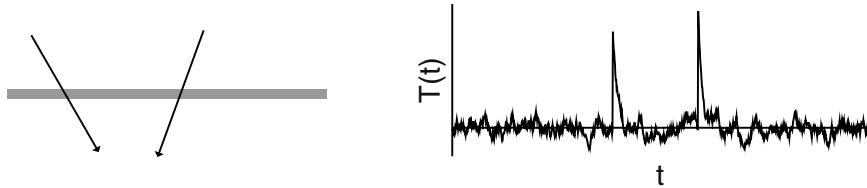


Figure 9.18. The bolometer shown at left responds to the energy deposited by ionizing particles passing through its absorber, leading to the impulsive signals known as glitches shown at right.

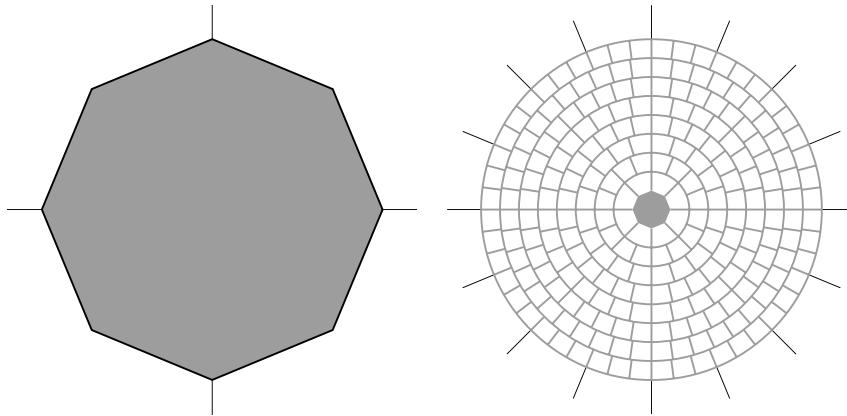


Figure 9.19. The FIRAS bolometer on the left compared to a spiderweb bolometer on the right.

Bolometers are just thermometers weakly coupled to a thermal bath by a conductance G . Radiation is focussed on the bolometer causing its temperature to rise by $\Delta T = P/G$ (see Figure 9.17). The thermal time constant of a bolometer is given by $\tau = C/G$ where $C = dQ/dT$ is the heat capacity of the bolometer in ergs/K. From the definition of entropy $S = k \ln \Omega$ with Ω being the state density, and $dS = dQ/T$, we find that

$$\ln \Omega_{bolo} = \ln \Omega_0 + (Q - Q_0)/kT_b - 0.5(Q - Q_0)^2/(kT_b^2 C) + \dots \quad (9.6)$$

where Q_\circ is the energy of the bolometer at the bath temperature T_b , which gives

$$\frac{d \ln \Omega}{dQ} = \frac{1}{kT} = \frac{1}{kT_b} \left(1 - \frac{Q - Q_\circ}{CT_b} \right) + \dots \quad (9.7)$$

and thus $T = T_b + (Q - Q_\circ)/C + \dots$ as required. The thermal bath has a much larger heat capacity so the overall density of states is

$$\Omega_{bolo}\Omega_{bath} \propto \exp[-0.5(Q - Q_\circ)^2/(kT_b^2 C)] \quad (9.8)$$

which is a Gaussian with a standard deviation of the energy in the bolometer of $\sigma(Q) = T\sqrt{kC}$. This corresponds to a standard deviation of the power $\sigma(P) = \sigma(Q)/\tau$ and since the noise bandwidth of a simple lowpass filter with time constant τ is $1/4\tau$, a noise equivalent power of $\text{NEP} = T\sqrt{4kC/\tau}$.

Clearly one obtains the best performance for a given time constant with a detector that has the lowest possible heat capacity. The heat capacity of a crystal varies like $C \propto (T/\Theta_D)^3$, where Θ_D is the Debye temperature. Diamond has the highest Debye temperature of any crystal, so FIRAS used an 8 mm diameter, 25 μm thick disk of diamond as a bolometer (Mather et al., 1993). Diamond is transparent, so a very thin layer of gold was applied to give a surface resistance close to the 377 ohms/square impedance of free space. On the back side of the diamond layer an impedance of 267 ohms/square gives a broadband absorption. Chromium was alloyed with the gold to stabilize the layer. The temperature of the bolometer was measured with a small silicon resistance thermometer. Running at $T = 1.6$ K, the FIRAS bolometers achieved an optical NEP of about $10^{-14} \text{ W}/\sqrt{\text{Hz}}$.

Since $C \propto T^3$ and $\text{NEP} = T\sqrt{4kC/\tau}$ the NEP scales like $T^{2.5}$. This means that a FIRAS-like bolometer running at 0.1 K could achieve an NEP of $10^{-17} \text{ W}/\sqrt{\text{Hz}}$.

A bolometer is sensitive to any source of heat, not just microwave photons, so charged particles passing through the absorbing layer lead to impulsive signals called glitches as seen in Figure 9.18. These events occur most frequently at the stratospheric altitudes where balloon-borne experiments operate. An important improvement in bolometer design was the use of mesh absorbers, since there is no need to fill an area $\sim \lambda^2$ in order to absorb radiation with wavelength λ . Figure 9.19 shows how the area that is sensitive to charged particles can be cut using a spiderweb bolometer (Bock et al., 1995). This cuts the mass and hence the heat capacity of the absorber. BOOMERanG used spiderweb bolometers.

Antenna-coupled bolometers (Schwarz and Ulrich, 1977) offer another way to achieve a small heat capacity and a small area sensitive to charged particles. Radiation is absorbed by an antenna and then coupled into a transmission line which brings it to a very small absorbing thermometer.

3. Recent Observations

In this paper I will discuss the new observations that have been released in the year prior to this meeting: September 2002 to September 2003. I will discuss these results in time order.

DASIOPOL

The Degree Angular Scale Interferometer (DASI) is a very small interferometric array that operates at 26-36 GHz and the South Pole. After measuring the angular power spectrum of the anisotropy (Halverson et al., 2002) the instrument was converted into a polarization sensitive interferometer which detected the E mode polarization at 5.5σ by looking at a small patch of sky for most of a year of integration time (Kovac et al., 2002). The level agreed well with the solid predictions for adiabatic primordial perturbations. Since the measured quantity was the EE autocorrelation, the 5.5σ corresponds to a 9% accuracy in the polarization amplitude.

The TE cross-correlation was also seen, but with only 50% accuracy. As expected, the B modes were not seen.

ARCHEOPS

ARCHEOPS is a balloon-borne experiment built to test the detectors and the cryogenic system planned for the ESA Planck Explorer mission High Frequency Instrument. ARCHEOPS has bolometers cooled below 0.1 K, and thus achieves a very high instantaneous sensitivity and was able to map a substantial fraction of the sky with good SNR in only 1 night of observing. This large sky coverage provided a lower level of cosmic or sample variance, so the ARCHEOPS data gave much better results on the low- ℓ side of ℓ_{pk} . This gave a peak location of $\ell_{pk} = 220 \pm 6$ (Benoît et al., 2003).

ACBAR

ACBAR is a bolometric camera array mounted on the VIPER 2 meter diameter telescope at the South Pole. It operates at several frequencies on both sides of the peak of the spectrum of the CMB, and thus can be used to make a sensitive test for the Sunyaev-Zeldovich effect. However, the data released to date are only at one frequency. ACBAR was able to measure the CMB angular power spectrum up to $\ell > 2000$ (Kuo et al., 2002), but the size of the surveyed regions was so small that ACBAR's ℓ resolution was too limited to provide much information about ℓ_{pk} . ACBAR was able with its high observing frequency to show that the excess at $\ell > 2000$ seen by CBI (Mason et al., 2003) is not a primary CMB anisotropy. It could be due to point sources, or it could

be due to the S-Z effect, both of which would be much weaker in the ACBAR band than in the CBI 26-36 GHz band.

WMAP

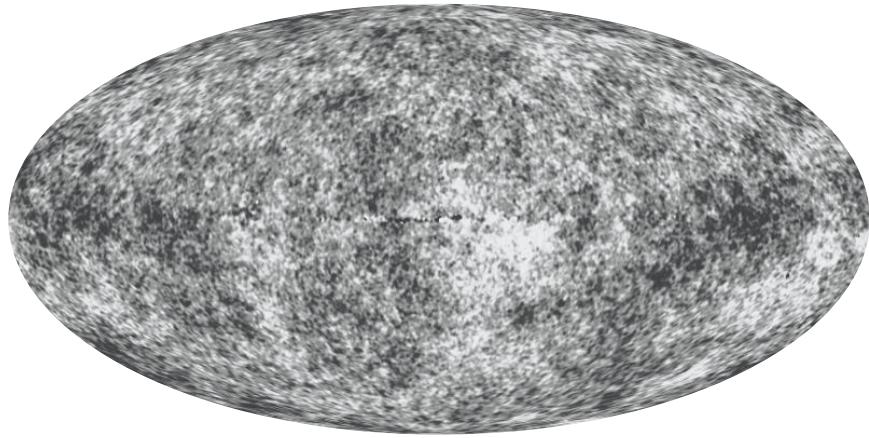


Figure 9.20. A “no galaxy” map made from an internal linear combination of the 5 *WMAP* bands, smoothed to 1° resolution.

The *WMAP* satellite, launched on 30 June 2001, released its first year results on 11 Feb 2003. Simultaneously the mission was renamed the Wilkinson Microwave Anisotropy Probe to honor the late David T. Wilkinson who was a key member of both the COBE and the *WMAP* teams until his death in September 2002.

WMAP observed at 5 frequencies: 23, 33, 41, 61 and 94 GHz. From maps in these 5 bands, an internal linear combination map has been constructed that cancels almost all of the Milky Way foreground while preserving the CMB anisotropy. Figure 9.20 shows this map on a gray scale. All bands were smoothed to $\approx 1^\circ$ resolution so the linear combination could be made without worrying about the different beamsizes in the different bands. After this smoothing 53% of the sky was within $\pm 53 \mu\text{K}$ of the median of the map, implying an RMS ΔT of $73 \mu\text{K}$ in this smoothing beam. This is considerably higher than the $30 \mu\text{K}$ RMS seen by the COBE DMR at a 10° smoothing because a 1° beam picks up a large part of the big first acoustic peak.

The *WMAP* results and their cosmological significance were described in 13 papers and will not be repeated here. Bennett et al., 2003a gave a description of the *WMAP* mission. Bennett et al., 2003b summarized the results from first year of *WMAP* observations. Bennett et al., 2003c described the observations of galactic and extragalactic foreground sources. Hinshaw et al., 2003b gave the angular power spectrum derived from the *WMAP* maps. Hinshaw et al.,

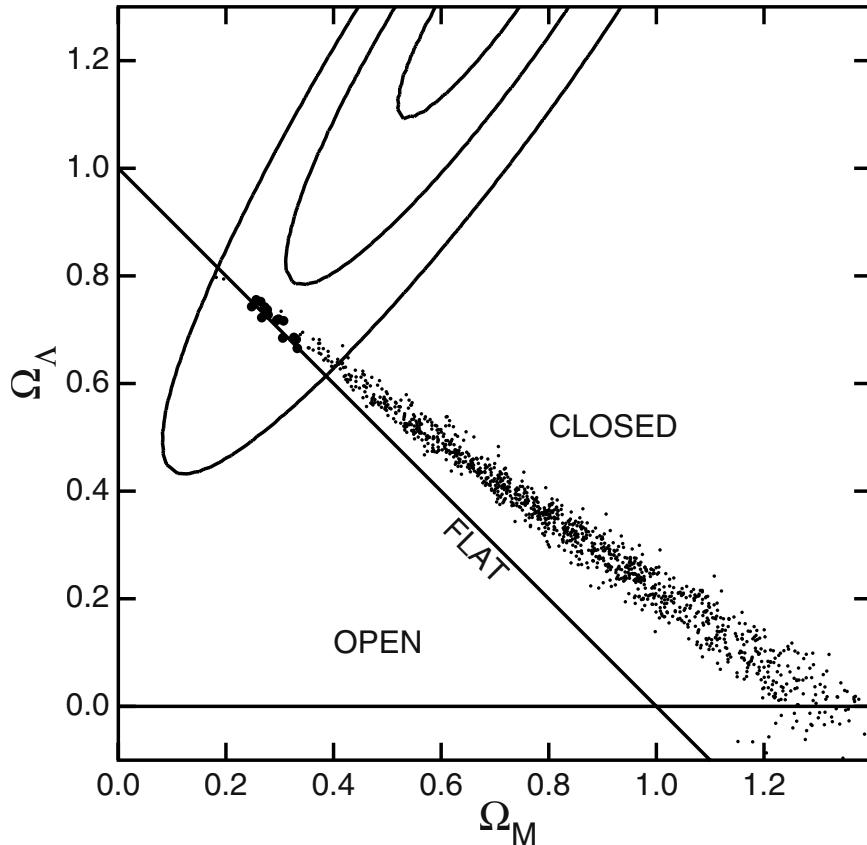


Figure 9.21. Cloud of points from a Monte Carlo Markov chain sampling of the likelihood of models fit to the *WMAP* plus other CMB datasets. The size of the points indicates how consistent the model is with the HST Key Project on the Distance Scale value for the Hubble constant. The contours show the likelihood computed for 230 Type Ia supernovae (Tonry et al., 2003).

2003a described the *WMAP* data processing and systematic error limits. Page et al., 2003a discussed the beam sizes and window functions for the *WMAP* experiment. Page et al., 2003c discussed results that can be derived simply from the positions and heights of the peaks and valleys in the angular power spectrum. Spergel et al., 2003 described the cosmological parameters derived by fitting the *WMAP* data and other datasets. Verde et al., 2003 described the fitting methods used. Peiris et al., 2003 described the consequences of the *WMAP* results for inflationary models. Jarosik et al., 2003a described the on-orbit performance of the *WMAP* radiometers. Kogut et al., 2003 described the *WMAP* observations of polarization in the CMB. Barnes et al., 2003 described

the large angle sidelobes of the *WMAP* telescopes. Komatsu et al., 2003 addressed the limits on non-Gaussianity that can be derived from the *WMAP* data.

Using the *WMAP* data plus CBI and ACBAR, the position of the big peak in the angular power spectrum was found to be $\ell_{pk} = 220.1 \pm 0.8$. The position of the big peak defines a track in the Ω_m - Ω_V plane shown in Figure 9.21.

The ratios of the anisotropy powers below the peak at $\ell \approx 50$, at the big peak at $\ell \approx 220$, in the trough at $\ell \approx 412$, and at the second peak at $\ell \approx 546$ were precisely determined using the *WMAP* data which has a single consistent calibration for all ℓ 's. Previously, these ℓ ranges had been measured by different experiments having different calibrations so the ratios were poorly determined. Knowing these ratios determined the photon:baryon:CDM density ratios, and since the photon density was precisely determined by FIRAS on COBE, accurate values for the baryon density and the dark matter density were obtained. These values are $\Omega_b h^2 = 0.0224 \pm 4\%$, and $\Omega_m h^2 = 0.135 \pm 7\%$. The ratio of CDM to baryon densities from the *WMAP* data is 5.0:1.

Because the matter density $\Omega_m h^2$ was fairly well constrained by the amplitudes, the positions of a point in Figure 9.21 served to define a value of the Hubble constant. The size of the points in Figure 9.21 indicates how well this derived Hubble constant agrees with the $H_0 = 72 \pm 8$ from the HST Key Project (Freedman et al., 2001). Shown as contours are the $\Delta\chi^2 = 1, 4, \& 9$ contours from my fits to 230 SNe Ia (Tonry et al., 2003). Clearly the CMB data, the HST data, and the SNe data are all consistent at a three-way crossing that is very close to the flat Universe line. Assuming the Universe actually is flat, the age of the Universe is very well determined: 13.7 ± 0.2 Gyr.

WMAP also found a TE (temperature-polarization) cross-correlation. At small angles the TE amplitude was perfectly consistent with the standard picture of the recombination era. But there was also a large angle TE signal that gave an estimate for the electron scattering optical depth since reionization: $\tau = 0.17 \pm 0.04$. Based on this the epoch of reionization was 200 million years after the Big Bang.

4. Summary

Measurements of the CMB anisotropy made in the last 5 years have moved cosmology into a new era of precise parameter determination and the ability to probe the conditions during the inflationary epoch. These results depend on the study of perturbations that are still in the linear, small amplitude regime, and thus are not confounded by non-linearities and the difficulties associated with hydrodynamics.

Acknowledgments

The *WMAP* mission is made possible by the support of the Office of Space Sciences at NASA Headquarters and by the hard and capable work of scores of scientists, engineers, technicians, machinists, data analysts, budget analysts, managers, administrative staff, and reviewers.

References

- Adams, W. S. 1941, *ApJ*, 93, 11
- Barnes, C., Hill, R. S., Hinshaw, G., Page, L., Bennett, C. L., Halpern, M., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Tucker, G. S., Wollack, E. & Wright, E. L. 2003, *ApJS*, 148, 51
- Bennett, C. L., Banday, A. J., Górski, K. M., Hinshaw, G., Jackson, P., Keegstra, P., Kogut, A., Smoot, G. F., Wilkinson, D. T. & Wright, E. L. 1996, *ApJL*, 464, L1
- Bennett, C. L., Bay, M., Halpern, M., Hinshaw, G., Jackson, C., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Wilkinson, D. T., Wollack, E. & Wright, E. L. 2003a, *ApJ*, 583, 1
- Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Wollack, E., Wright, E. L., Barnes, C., Greason, M. R., Hill, R. S., Komatsu, E., Nolta, M. R., Odegard, N., Peiris, H. V., Verde, L. & Weiland, J. L. 2003b, *ApJS*, 148, 1
- Bennett, C. L., Hill, R. S., Hinshaw, G., Nolta, M. R., Odegard, N., Page, L., Spergel, D. N., Weiland, J. L., Wright, E. L., Halpern, M., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Tucker, G. S., and Wollack, E. 2003c, *ApJS*, 148, 97
- Benoît, A., Ade, P., Amblard, A., Ansari, R., Aubourg, É., Bargot, S., Bartlett, J. G., Bernard, J.-P., Bhatia, R. S., Blanchard, A., Bock, J. J., Boscaleri, A., Bouchet, F. R., Bourrachot, A., Camus, P., Couchot, F., de Bernardis, P., Delabrouille, J., Désert, F.-X., Doré, O., Douspis, M., Dumoulin, L., Dupac, X., Filliatre, P., Fosalba, P., Ganga, K., Gannaway, F., Gautier, B., Giard, M., Giraud-Héraud, Y., Gispert, R., Guglielmi, L., Hamilton, J.-C., Hanany, S., Henrot-Versillé, S., Kaplan, J., Lagache, G., Lamarre, J.-M., Lange, A. E., Macías-Pérez, J. F., Madet, K., Maffei, B., Magneville, C., Marrone, D. P., Masi, S., Mayet, F., Murphy, A., Naraghi, F., Nati, F., Patanchon, G., Perrin, G., Piat, M., Ponthieu, N., Prunet, S., Puget, J.-L., Renault, C., Rosset, C., Santos, D., Starobinsky, A., Strukov, I., Sudiwala, R. V., Teyssier, R., Tristram, M., Tucker, C., Vanel, J.-C., Vibert, D., Wakui, E. & Yvon, D. 2003, *A&A*, 399, L19
- Bock, J. J., Chen, D., Mauskopf, P. D. & Lange, A. E. 1995, *Space Science Reviews*, 74, 229–235.

- Boggess, N. W., Mather, J. C., Weiss, R., Bennett, C. L., Cheng, E. S., Dwek, E., Gulkis, S., Hauser, M. G., Janssen, M. A., Kelsall, T., Meyer, S. S., Moseley, S. H., Murdock, T. L., Shafer, R. A., Silverberg, R. F., Smoot, G. F., Wilkinson, D. T. & Wright, E. L. 1992, *ApJ*, 397, 420
- Bond, J. R. & Efstathiou, G. 1987, *MNRAS*, 226, 655
- Conklin, E. K. 1969, *Nature*, 222, 971–972.
- Corey, B. E. and Wilkinson, D. T. 1976, *Bull. Am. Astr. Soc.*, 8, 351
- de Bernardis, P., Ade, P. A. R., Bock, J. J., Bond, J. R., Borrill, J., Boscaleri, A., Coble, K., Crill, B. P., De Gasperis, G., Farese, P. C., Ferreira, P. G., Ganga, K., Giacometti, M., Hivon, E., Hristov, V. V., Iacoangeli, A., Jaffe, A. H., Lange, A. E., Martinis, L., Masi, S., Mason, P. V., Mauskopf, P. D., Melchiorri, A., Miglio, L., Montroy, T., Netterfield, C. B., Pascale, E., Piacentini, F., Pogosyan, D., Prunet, S., Rao, S., Romeo, G., Ruhl, J. E., Scaramuzzi, F., Sforna, D. & Vittorio, N. 2000, *Nature*, 404, 955
- Dicke, R. H., Beringer, R., Kyhl, R. L. & Vane, A. B. 1946, *Physical Review*, 70, 340
- Dicke, R. H., Peebles, P. J. E., Roll, P. G. & Wilkinson, D. T. 1965, *ApJ*, 142, 414
- Eriksen, H. K., Hansen, F. K., Banday, A. J., Gorski, K. M. & Lilje, P. B. 2003, ArXiv Astrophysics e-prints. astro-ph/0307507.
- Fixsen, D. J., Cheng, E. S., Gales, J. M., Mather, J. C., Shafer, R. A. & Wright, E. L. 1996, *ApJ*, 473, 576
- Freedman, W. L., Madore, B. F., Gibson, B. K., Ferrarese, L., Kelson, D. D., Sakai, S., Mould, J. R., Kennicutt, R. C., Ford, H. C., Graham, J. A., Huchra, J. P., Hughes, S. M. G., Illingworth, G. D., Macri, L. M. & Stetson, P. B. 2001, *ApJ*, 553, 47
- Halverson, N. W., Leitch, E. M., Pryke, C., Kovac, J., Carlstrom, J. E., Holzapfel, W. L., Dragovan, M., Cartwright, J. K., Mason, B. S., Padin, S., Pearson, T. J., Readhead, A. C. S. & Shepherd, M. C. 2002, *ApJ*, 568, 38
- Hamilton, J. ., Benoît, A. & Collaboration, 2003, ArXiv Astrophysics e-prints.
- Henry, P. S. 1971, *Nature*, 231, 516
- Herzberg, G. 1950, New York: Van Nostrand Reinhold, 1950, 2nd ed.
- Hinshaw, G., Barnes, C., Bennett, C. L., Greason, M. R., Halpern, M., Hill, R. S., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Odegard, N., Page, L., Spergel, D. N., Tucker, G. S., Weiland, J. L., Wollack, E. & Wright, E. L. 2003a, *ApJS*, 148, 63
- Hinshaw, G., Spergel, D. N., Verde, L., Hill, R. S., Meyer, S. S., Barnes, C., Bennett, C. L., Halpern, M., Jarosik, N., Kogut, A., Komatsu, E., Limon, M., Page, L., Tucker, G. S., Weiland, J. L., Wollack, E. & Wright, E. L. 2003b, *ApJS*, 148, 135

- Jarosik, N., Barnes, C., Bennett, C. L., Halpern, M., Hinshaw, G., Kogut, A., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Weiland, J. L., Wollack, E. & Wright, E. L. 2003a, ApJS, 148, 29
- Jarosik, N., Bennett, C. L., Halpern, M., Hinshaw, G., Kogut, A., Limon, M., Meyer, S. S., Page, L., Pospieszalski, M., Spergel, D. N., Tucker, G. S., Wilkinson, D. T., Wollack, E., Wright, E. L. & Zhang, Z. 2003b, ApJS, 145, 413
- Jungman, G., Kamionkowski, M., Kosowsky, A. & Spergel, D. N. 1996, Phys. Rev. Lett., 76, 1007
- Kaiser, M. E. & Wright, E. L. 1990, ApJL, 356, L1
- Kamionkowski, M., Kosowsky, A. & Stebbins, A. 1997, Phys. Rev. D, 55, 7368
- Knox, L. & Page, L. 2000, Phys. Rev. Lett., 85, 1366
- Kogut, A., Spergel, D. N., Barnes, C., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Limon, M., Meyer, S. S., Page, L., Tucker, G. S., Wollack, E. & Wright, E. L. 2003, ApJS, 148, 161
- Komatsu, E., Kogut, A., Nolta, M. R., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Verde, L., Wollack, E. & Wright, E. L. 2003, ApJS, 148, 119
- Kovac, J. M., Leitch, E. M., Pryke, C., Carlstrom, J. E., Halverson, N. W. & Holzapfel, W. L. 2002, Nature, 420, 772
- Kuo, C. L. et al. 2002, ApJ. astro-ph/0212289.
- Lubin, P. M. & Smoot, G. F. 1981, ApJ, 245, 1
- Mason, B. S., Pearson, T. J., Readhead, A. C. S., Shepherd, M. C., Sievers, J., Udomprasert, P. S., Cartwright, J. K., Farmer, A. J., Padin, S., Myers, S. T., Bond, J. R., Contaldi, C. R., Pen, U., Prunet, S., Pogosyan, D., Carlstrom, J. E., Kovac, J., Leitch, E. M., Pryke, C., Halverson, N. W., Holzapfel, W. L., Altamirano, P., Bronfman, L., Casassus, S., May, J. & Joy, M. 2003, ApJ, 591, 540
- Mather, J. C., Fixsen, D. J. & Shafer, R. A. 1993, In *Proc. SPIE Vol. 2019, p. 168–179, Infrared Spaceborne Remote Sensing, Marija S. Scholl; Ed.*, pages 168–179.
- Mather, J. C., Fixsen, D. J., Shafer, R. A., Mosier, C. & Wilkinson, D. T. 1999, ApJ, 512, 511
- Page, L., Barnes, C., Hinshaw, G., Spergel, D. N., Weiland, J. L., Wollack, E., Bennett, C. L., Halpern, M., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Tucker, G. S. & Wright, E. L. (2003a). ApJS, 148, 39–50.
- Page, L., Jackson, C., Barnes, C., Bennett, C., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Spergel, D. N., Tucker, G. S., Wilkinson, D. T., Wollack, E. & Wright, E. L. 2003b, ApJ, 585, 566

- Page, L., Nolta, M. R., Barnes, C., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Peiris, H. V., Spergel, D. N., Tucker, G. S., Wollack, E. & Wright, E. L. 2003c, *ApJS*, 148, 233
- Peebles, P. J. E. 1982, *ApJL*, 263, L1
- Peiris, H. V., Komatsu, E., Verde, L., Spergel, D. N., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Tucker, G. S., Wollack, E. & Wright, E. L. 2003, *ApJS*, 148, 213
- Penzias, A. A. and Wilson, R. W. 1965, *ApJ*, 142, 419
- Roth, K. C., Meyer, D. M. & Hawkins, I. 1993, *ApJL*, 413, L67
- Sachs, R. K. and Wolfe, A. M. 1967, *ApJ*, 147, 73
- Schwarz, S. E. and Ulrich, B. T. 1977, *Journal of Applied Physics*, 48, 1870
- Scott, D., Silk, J. & White, M. 1995, *Science*, 268, 829
- Seljak, U. & Zaldarriaga, M. 1997, *Physical Review Letters*, 78, 2054
- Silk, J. 1968, *ApJ*, 151, 459.
- Smoot, G. F., Goernstein, M. V. & Muller, R. A. 1977, *prl*, 39, 898
- Spergel, D. N., Verde, L., Peiris, H. V., Komatsu, E., Nolta, M. R., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Tucker, G. S., Weiland, J. L., Wollack, E. & Wright, E. L. 2003, *ApJS*, 148, 175.
- Thaddeus, P. (1972). *ARA&A*, 10, 305.
- Tonry, J., Schmidt, B. P., Barris, B., Candia, P., Challis, P., Cloociatti, A., L., Coil A., Filipenko, A. V., Garnavich, P., et al. 2003, *ApJ*, in press, astro-ph/0305008.
- Verde, L., Peiris, H. V., Spergel, D. N., Nolta, M. R., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Tucker, G. S., Wollack, E. & Wright, E. L. 2003, *ApJS*, 148, 195–211.
- Wilbanks, T., Devlin, M., Lange, A. E., Beeman, J. W. & Sato, S. 1990, *IEEE Transactions on Nuclear Science*, 37, 566
- Wright, E. L. 1999, *New Astronomy Review*, 43, 201
- Wright, E. L., Mather, J. C., Bennett, C. L., Cheng, E. S., Shafer, R. A., Fixsen, D. J., Eplee, R. E., Isaacman, R. B., Read, S. M., Boggess, N. W., Gulkis, S., Hauser, M. G., Janssen, M., Kelsall, T., Lubin, P. M., Meyer, S. S., Moseley, S. H., Murdock, T. L., Silverberg, R. F., Smoot, G. F., Weiss, R. & Wilkinson, D. T. 1991, *ApJ*, 381, 200
- Wright, E.L. (1996). ArXiv Astrophysics e-prints. astro-ph/9612006.

Chapter 10

FLUCTUATIONS IN THE COSMIC MICROWAVE BACKGROUND

Andrew H. Jaffe

Imperial College Astrophysics

a.jaffe@imperial.ac.uk

Abstract

The Cosmic Microwave Background (CMB) gives us a snapshot of the universe when it was hotter, denser and simpler than today. It is relic radiation from the Big Bang, left over from the transition from an opaque and ionized primordial plasma to a transparent, neutral gas that eventually cooled enough to form the stars, galaxies and clusters that we observe. In these lectures, I will discuss the cosmological evolution of the CMB and of fluctuations in the photons, dark matter and baryons, producing the patterns that we observe today. I will also discuss the statistical tools we have developed to characterize these fluctuations and use them to derive the impact of various cosmological parameters upon the CMB.

Keywords: Cosmic microwave background, theory.

1. Introduction

After the discovery (Hubble, 1929; Hubble, 1958) of the expansion of the Universe, Tolman, 1934 applied Einstein's theory of General Relativity to the case of a Universe containing black-body radiation – photons governed by the Planck distribution, described by a specific temperature. He discovered that the Universe could be seen as a giant black-body cavity, and that the photons maintained the Planck distribution, with the wavelength redshifting with the expansion, and the temperature decreasing. In now-familiar language, if the scale factor is $a(z)$, the wavelength of radiation scales as $\lambda(t) \propto a(t)$ and the temperature as $T(t) \propto 1/a(t)$.

As what would eventually be called Big Bang cosmology was developed, the evolution of this radiation field was worked out in more detail. By the 1960s, Dicke and his colleagues at Princeton University in New Jersey had

begun to build the hardware required to search for it. Simultaneously, Penzias and Wilson of Bell Labs, also in New Jersey, were calibrating a new microwave antenna with which they hoped to make astronomical observations in the as-yet unexplored centimeter waveband. Rather than the sky full of individual sources that radio astronomers saw, Penzias and Wilson, after eliminating myriad sources of possible systematic error, found the sky abuzz with a uniform radiation field with an effective temperature of 3.5K. Unfamiliar with cosmology, they consulted, among others, their colleague B. Burke, who, as it happened, was aware of the efforts of Dicke's Princeton team – Penzias and Wilson had scooped them. The results appeared in a pair of papers, the experimental results from Penzias and Wilson, 1965 and the theoretical underpinnings by Dicke et al., 1965. This was the discovery of the Cosmic Microwave Background.

Soon thereafter, more details of the evolution of the CMB were worked out. Rather than being completely uniform, the CMB on the largest scales would reflect our motion through the universal bath of photons; it would exhibit a so-called dipole pattern, hotter in our direction of motion and colder in the opposite. This dipole was discovered by Smoot et al., 1977. But it was also realized that temperature anisotropies in the present-day CMB would necessarily arise. Broadly speaking, the same physics that would eventually be responsible for the large-scale structure of the Universe would also produce deviations from a uniform CMB temperature.

In these lectures, I will discuss some of the details of the generation of these anisotropies. (Other lectures will cover other aspects of the CMB. E. Wright discusses observations of the CMB and J. Bartlett discusses details of CMB polarization.) In such a short space I cannot do justice to such a broad and important subject; I commend various articles and reviews (Scott et al. 1995; Hu et al., 1997; Kosowsky, 1999; Kamionkowski and Kosowsky, 1999; Church et al., 2001; Kosowsky, 2002; Hu and Dodelson, 2002) for a more detailed treatment.

In §10.4, I will discuss the evolution of density perturbations in an expanding Universe and in §10.5 the plasma oscillations thereby induced. In §10.6 I will introduce that statistical tools to describe the distribution of CMB temperatures on the sky, and in §10.7 how the cosmological parameters influence the distribution of temperatures. Finally, in §10.8 I will briefly review how we actually analyze CMB data and conclude in §10.8.

2. Cosmological Preliminaries

The Cosmic Microwave Background (as discussed, for example, in Ned Wright's contribution to this volume), is, along with the expansion of the Universe and the abundance of the light elements, one of the three so-called “pi-

lars of the Big Bang". These three pieces of evidence imply that the Universe has been expanding and cooling for roughly fifteen billion years.

As we have seen in other lectures at this school, and in textbooks (Kolb and Turner, 1990; Peebles, 1993; Peacock, 1999), the early Universe was denser and hotter. If we look back far enough, we will find a time when the Universe was so hot that the electrons were stripped from their nuclei; matter in the Universe was in the form of a hot, ionized plasma. In this section, we will trace the evolution of the Universe at this time and its repercussions in the visible Universe today.

As background, we will need to recall the evolution of a Universe governed by the Freedmann-Robertson-Walker (FRW) metric, and of the different components thereof.

The metric in an FRW universe is given by (in units with $c = 1$)

$$ds^2 = dt^2 - a^2(t) \left[\frac{dr^2}{1 - r^2/R^2} + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \right] \quad (10.1)$$

where t is the time coordinate, (r, θ, ϕ) are spherical polar coordinates for the spatial part of the metric, and $a(t)$ is the "scale factor," defined so that physical lengths scale proportional to a . The constant parameter R gives the radius of curvature; $1/R$ is positive, negative or 0 for spaces of positive, negative or zero (flat) curvature. Note that conventions differ: here, $a(t)$ is dimensionless, the coordinate distance, r , and the radius of curvature R have units of length. The metric can be written in a variety of other forms useful in various circumstances. From the scale factor, we can define the redshift, z : $1 + z = a_0/a$, where a zero subscript refers to the present day.

The dynamics of the FRW universe depend upon the matter content. The total density of an FRW Universe is ρ . To do this, let us consider a slightly simplified version of the Universe, consisting of components labelled $i = 1, \dots, n$, where each component has a density, ρ_i , and a pressure, p_i .

For completeness, we present here the equations governing the evolution of an FRW Universe. First, we define the expansion rate as the value of \dot{a}/a , where the overdot is a time derivative; the value of the expansion rate today is the Hubble Constant,

$$H_0 \equiv \frac{\dot{a}_0}{a_0} = 100h \text{ kms}^{-1} \text{Mpc}^{-1}; \quad (10.2)$$

the Hubble Space Telescope Key Project (Freedman et al., 2001) gives $h = 0.72 \pm 0.03$. The *critical density* is given by

$$\rho_c = \frac{3H_0^2}{8\pi G}; \quad (10.3)$$

if the overall density $\rho_{\text{tot}} = \rho_c$, the Universe is geometrically flat; if $\rho > \rho_c$, it is closed, and if $\rho < \rho_c$, open. We therefore define $\Omega_i = \rho_i/\rho_c$, the present-day contribution of component i to the critical density.

With these definitions and the Einstein Equations of General Relativity, we can show that

$$\begin{aligned} \frac{\dot{a}}{a} &= H_0 E(z) \\ &= H_0 \sqrt{\Omega_m(1+z)^3 + \Omega_\Lambda + \Omega_r(1+z)^4 + \Omega_Q(1+z)^{3(1+w)} + \Omega_k(1+z)^2} \end{aligned} \quad (10.4)$$

where m refers to pressureless matter, Λ to a “cosmological constant” ($w_\Lambda = -1$), r to radiation, and Q to a component with arbitrary equation of state parameter, $w = p_Q/\rho_Q$. Finally, in a slight misuse of nomenclature, $\Omega_k = 1 - \Omega_m - \Omega_r - \Omega_\Lambda - \Omega_Q$ gives the contribution from curvature. The powers of $(1+z)$ are just the factor by which the density of each component redshifts over time, and indeed this curvature equation defining Ω_k holds at all times.

The remaining evolution equation is

$$\ddot{\frac{a}{a}} = \frac{1}{2} H_0^2 \left[\Omega_m(1+z)^3 + 2\Omega_r(1+z)^4 - 2\Omega_\Lambda + \Omega_Q(1+w)(1+z)^{3(1+w)} \right]. \quad (10.5)$$

From these, we see that the overall evolution depends on the component which dominates the density at any time. In particular, we see that in the past the density approaches $\Omega_{\text{tot}} = 1$ as $t \rightarrow 0$, and that, no matter what component dominates today, radiation dominated at early times, followed, in most models, by matter domination at $z_{\text{eq}} = \Omega_m/\Omega_\gamma \sim 40,000$.

3. The Last Scattering Surface

From the evolution of the temperature, $T \propto (1+z)$, and the number density of non-relativistic particles, $n \propto (1+z)^3$, we can infer that the universe was once sufficiently hot and dense to ionize the hydrogen that today makes up most of the baryon density of the Universe. Naively, we might expect this to occur when $T \sim 1 \text{ Ry} = 13.6 \text{ eV}$, the binding energy of Hydrogen.

To do the calculation in more detail, for the baryons (nuclei) and electrons, we define the ionization fraction $X(t)$, the ratio of the density of ions to neutrals; if we assume overall charge-neutrality, this is equal to the ratio of the free electron number density to that of neutrals. We also assume that the number density of any massive species (i) is large enough that it can be described by a Boltzmann distribution,

$$n_i = g_i \frac{m_i T}{2\pi} \exp \frac{\mu_i - m_i}{T} \quad (10.6)$$

for a species with mass, m_i , chemical potential, μ_i , and “statistical weight,” g_i , at a temperature, T . Charge neutrality and baryon- and lepton-number conservation enforce $n_p = n_e$ and $n_B = n_p + n_H$, and the reaction $e + p \rightarrow H$ requires $\mu_p + \mu_e = \mu_H$. Putting all of these together (left as an “exercise for the reader”!), we can derive the Saha equation:

$$\frac{1 - X(z)}{X(z)^2} = \frac{n_B}{n_\gamma} \frac{4\sqrt{2}\zeta(3)}{\sqrt{\pi}} \left[\frac{T(z)}{m_e} \right]^{3/2} e^{1 \text{ Ry}/T(z)} \quad (10.7)$$

We show the Hydrogen ionization fraction in figure 10.1. We find a very rapid transition from $X \simeq 1$ to $X \simeq 0$ at $z \simeq 1,100$, corresponding to $T(z) = 0.3\text{eV}$. This is considerably lower than our naive expectation of 13.6eV , due to the small prefactors on the right hand side of Eq. 10.7, themselves due to the very small value of $n_B/n_\gamma = 2.7 \times 10^{-8}(\Omega_B h^2)$: there are many more photons than baryons, and so even the small fraction in the high-energy tail of the Boltzmann distribution are sufficient to keep the Universe ionized.

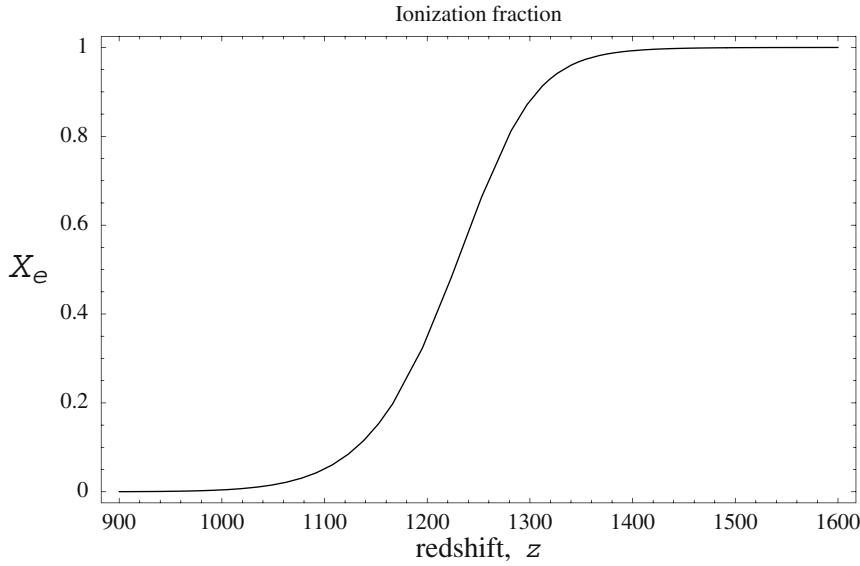


Figure 10.1. The ionization fraction, x_e , as a function of redshift

Reionization

If we follow the evolution of the ionization fraction yet further in time, we would find that the universe never completely recombines; rather, we are left with a residual ionization fraction $X(0) \sim 10^{-4}$. As the universe expands and

cools, any reaction with a rate, $\Gamma(z)$, which falls below the expansion rate, \dot{a}/a “freezes out.”

But we find that the universe today is, in fact, ionized, with $X(0) \simeq 1 \gg 10^{-4}$; this is largely surmised from features in the spectra of QSOs, which, outside of dense clumps (galaxies and proto-galaxies), does not show suppression due to the Lyman-alpha transition that would be present if the intergalactic medium was neutral. More recently, some evidence of this suppression, known as the Gunn-Peterson effect (Gunn and Peterson, 1965) has been seen at $z \simeq 6$ (Becker et al., 2001; Djorgovski et al., 2001; Fan et al., 2002), implying the presence of some neutral gas. However, observations of CMB polarization by the WMAP satellite (see the lectures at this school from J. Bartlett) seem to imply that the epoch of reionization occurs much earlier, at $z \simeq 17$.

4. Perturbations on Large and Small Scales

In the previous section, we analyzed the evolution of the overall ionization fraction; now we will start to consider inhomogeneities in the primordial plasma. These inhomogeneities will eventually become the large-scale structures we see in the Universe today, but they grew from tiny seeds at the last scattering surface. To handle this situation, we need to use relativistic perturbation theory in an expanding Universe. Well inside the horizon, we can use ordinary Newtonian gravity, but on larger scales the details of General Relativity come into play. In GR we are free to define the space and time coordinate systems as we please. For example, in one coordinate system two nearby points may have different densities at the same time; by redefining the time coordinate at one of these points, we can change the density contrast between the two points – in fact we can even “gauge it away” for all time. This freedom indicates the extra care that must be taken especially on superhorizon scales where there is no natural way to define coordinate systems, although a gauge invariant formalism was developed (Bardeen, 1980). Details of relativistic perturbation theory lie beyond the scope of these lectures, but is discussed at length in the textbooks cited in the introduction, as well as in reviews (Efstathiou, 1990; Mukhanov et al., 1992).

Nonetheless, we can get a feel for the evolution of perturbations by considering the smallest scales, where Newtonian gravity is sufficient. We start in physical (non-expanding) coordinates and simply write down the equations of motion:

$$\begin{aligned} \frac{\partial \rho_i}{\partial t} + \nabla \cdot (\rho_i \mathbf{v}_i) &= 0 && \text{continuity} \\ \frac{\partial \mathbf{v}_i}{\partial t} + (\mathbf{v}_i \cdot \nabla) \mathbf{v}_i &= -\nabla \Phi - \frac{1}{\rho_i} \nabla p && \text{Euler} \\ \nabla^2 \Phi &= 4\pi G \rho && \text{Poisson .} \end{aligned} \quad (10.8)$$

Although these are the equations that we would right down from “first principles” in a Newtonian analysis, they are also the small-scale and small-velocity limit of General Relativity. To account for the expansion of the Universe, we change to comoving coordinates, \mathbf{r} and peculiar velocity, \mathbf{u} , defined from physical coordinates, \mathbf{x} , and velocities, \mathbf{v} as

$$\mathbf{x} = \mathbf{r}/a(t); \quad \mathbf{u} = \mathbf{v}/a - (\dot{a}/a)\mathbf{x}. \quad (10.9)$$

The zeroth-order solution is the homogeneous FRW universe of Section 10.2, Eqns. 10.3-10.5, and we define the density perturbation, δ_i , by $\rho(\mathbf{x}, t) = \bar{\rho}(t)[1 + \delta(\mathbf{x}, t)]$. With these definitions, our perturbation variables are the density perturbation δ and the peculiar velocity, \mathbf{u} , taken to be of the same order in a perturbation series. Finally, we Fourier transform all of the spatially-varying quantities into comoving wavenumber \mathbf{k} , combine the resulting equations, and take all of the second-order combinations to be vanishingly small. This eliminates all spatial derivatives and removes mode-coupling terms containing more than one value of the wavenumber. The resulting equation is a single second-order differential equation at each value of \mathbf{k} ,

$$\frac{d^2\delta_{\mathbf{k}}}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\delta_{\mathbf{k}}}{dt} = \left[4\pi G\bar{\rho} - \left(\frac{c_s k}{a}\right)^2\right]\delta_{\mathbf{k}} = 4\pi G\bar{\rho}\left[1 - \left(\frac{k}{k_J}\right)^2\right]\delta_{\mathbf{k}} \quad (10.10)$$

where we have defined the *Jeans wavenumber*, k_J , and adiabatic sound speed, c_s , from

$$\lambda_J = \frac{2\pi a}{k_J} = c_s \left(\frac{\pi}{G\bar{\rho}}\right)^{1/2}; \quad c_s^2 = \left.\frac{\partial p}{\partial \rho}\right|_S. \quad (10.11)$$

Although we can't write down a general solution to this equation, we can examine it in various special cases and limits.

In a flat Universe, dominated by pressureless matter like Cold Dark Matter, we have $a \propto t^{2/3}$ and, by definition, $p = c_s = 0$. The solution to the second-order equation is $\delta_{\mathbf{k}} = A_{\mathbf{k}}t^{2/3} + B_{\mathbf{k}}t^{-1}$. The first term, proportional to $t^{2/3} \propto a(t)$, is the *growing mode*, and the t^{-1} corresponds to the *decaying mode*. In a non-flat matter-dominated universe we cannot always write down the functions of time in closed form, but very generally we do find growing and decaying terms. Moreover, plugging these back into the original first order equations, we can solve for the peculiar velocity, \mathbf{u} , and find that the irrotational flow, satisfying $\nabla \times \mathbf{u}_{\parallel} = 0$, evolves as $u_{\parallel} \propto a^{-1/2}$, and the vorticity, satisfying $\nabla \cdot \mathbf{u}_{\perp} = 0$, evolves as $v_{\perp} \propto a^{-2}$. That is, all peculiar velocities decay, but vorticity — angular momentum — decays much more rapidly.

In a Universe dominated instead by curvature or an accelerating component such as a cosmological constant, the expansion of the universe is more rapid than in a matter dominated universe. In both of these cases, there is no longer a growing mode: neither solution to Equation 10.10 grows with time.

Finally we must deal with perturbations not to pressureless matter but to baryons and photons. At early times, the universe is radiation dominated, and $p_r = \rho/3$ for radiation. Before the epoch of recombination, the same Thomson scattering processes that keep the baryons ionized also keep the radiation tightly-coupled to the ions. The nuclei have pressure $p_B = (5kT/3m_p)\rho \ll p_r$ for hydrogen gas at temperature T. We can thus consider a single “radiation/baryon fluid” with sound speed

$$c_s = \frac{c}{\sqrt{3}} \left(\frac{3}{4} \frac{\rho_B}{\rho_\gamma} + 1 \right)^{-1/2} \equiv \frac{c}{\sqrt{3}} (1 + R)^{-1/2}. \quad (10.12)$$

where we have defined $R = (3/4)(\rho_B/\rho_\gamma)$, so any *baryon loading* ($R > 0$) decreases the sound speed and the pressure relative to the case of pure radiation.

On small scales, these sources of pressure determine the evolution of perturbations. Consider once again Eq. 10.10. When the wavenumber $k > k_J$, or conversely when the wavelength is smaller than the Jeans length, pressure dominates over gravity and the fluid oscillates with angular frequency $\omega \sim c_s k$. The detailed solution actually involves Bessel functions when expansion is correctly taken into account, and there are additional complications due to gravitational interactions with any pressureless component such as CDM which can continue to collapse.

Before recombination, the radiation pressure is so great that the Jeans length is greater than the horizon size, and so no perturbations within the horizon can grow; they can only oscillate as sound waves. Conversely, after recombination, the pressure drops precipitously and all of a sudden, perturbations within the horizon can grow.

In order to understand the evolution of perturbation on larger scales (compared to the scale of the Horizon), we need to consider the most general perturbation to the various constituents of the universe (dark matter particles, photons, baryons, ...), which would let their densities and velocities each vary individually. Since the number of degrees of freedom contributing to the entropy come almost entirely from radiation, the specific entropy, or entropy per particle, is defined by

$$s = \frac{n_\gamma}{n_b} \propto \frac{T^3}{n_b} \propto \frac{\rho_\gamma^{3/4}}{\rho_b} \quad (10.13)$$

for baryons, labeled by b . We can then write a change in pressure as

$$dp(\rho, s) = \frac{1}{3} \frac{1}{1+R} d\rho + \frac{1}{3} \frac{\rho_B}{1+R} ds \quad (10.14)$$

However, we can generically decompose a perturbation into *isocurvature* modes, in which the local density of the constituents vary to preserve a constant total

density, and *adiabatic* perturbations in which the specific entropy, s , is constant, but in which the total density, or, equivalently, the local curvature, varies. Thus, an adiabatic perturbation has $ds = 0$, while an isocurvature perturbation has $d\rho = 0$. In fact, the later evolution of perturbations is very different for isocurvature and adiabatic perturbations, which evolve independently on large scales, and result in qualitatively different power spectra of CMB fluctuations. A priori, we might expect that whatever is responsible for perturbations would treat all species equally giving pure curvature perturbations, leaving the ratio of the various number densities constant. Indeed, the *simplest* inflationary scenarios do not produce isocurvature perturbations (see the lectures by A. Riazuelo at this school) but more generally there is no reason not to expect isocurvature modes. Except, as we shall see, the data are fit very well indeed by the pure adiabatic curvature perturbations!

On scales larger than the horizon we need to use the full equations of General Relativity to determine the evolution of perturbations, and it is in this case that the lack of a fixed coordinate system is the most problematic. The evolution of density perturbations depends on the chosen coordinate system or gauge. For example, in the longitudinal or conformal-newtonian gauge, we can write the perturbed metric as

$$ds^2 = (1 + 2\Phi)dt^2 - a^2(t)[(1 - 2\Psi)\gamma_{ij} + h_{ij}]dx_idx_j \quad (10.15)$$

where Φ and Ψ are the potential functions corresponding to the Newtonian potential on small scales, γ_{ij} is the spatial part of the unperturbed FRW metric, Eq. 10.1, and h_{ij} is the tensor part of the metric perturbation, corresponding to gravitational waves. (Note that different authors have different conventions with $\pm\Psi \leftrightarrow \pm\Phi$.)

5. Oscillations in the Primordial Plasma

Armed with an understanding of the evolution of the ionization fraction and of perturbations, we can actually understand most of the qualitative features of the CMB.

After the epoch of last scattering, photons freely stream through the universe, their temperature carrying the imprint of the conditions of the plasma from which they were freed. As we saw above in Figure 10.1, the transition from a completely-ionized, tightly-couple plasma of photons and baryons to the free-streaming epoch of neutral gas happens on a very short timescale ($\delta z \simeq 200$, corresponding to 100,000 years at $t \simeq 400,000$ years); on scales larger than about 30,000Mpc \sim 100,000 light-years we can treat the surface of last scattering as infinitesimally thin. In that case, present-day photons in

some direction \hat{n} will have a temperature perturbation

$$\frac{\Delta T}{T} \Big|_0 = \frac{1}{4} \frac{\delta \rho_\gamma}{\rho_\gamma} \Big|_{\text{LSS}} + (\mathbf{v}_0 - \mathbf{v}_{\text{LSS}}) \cdot \hat{\mathbf{n}} + \int_{\text{LSS}}^0 d\eta h_{ij} \hat{n}_i \hat{n}_j \quad (10.16)$$

where “0” refers to today, “LSS” refers to the point on the last scattering surface “pointed at” by a geodesic from the observer at “0”, ρ_γ is the photon density (and, since $T \propto \rho^4$ for a black body, the first term gives the initial temperature perturbation), \mathbf{v} is the peculiar velocity (so the second term is a doppler shift), and h_{ij} is the metric perturbation (so the third term is due to the photons’ falling into and climbing out of potential wells along the line of sight, and the integral is along the geodesic (dependent on the overall geometry of the Universe).

But any complete description of the evolution of perturbations in the universe will link all of these terms: initial velocity and density perturbations to the various components (baryons, dark matter, photons) evolve prior to last scattering as discussed above, and so photon overdensities occur in potential wells, and velocity perturbations occur in response to gravitational and pressure forces. Indeed, to solve this problem in its most general form, we must resort to the Boltzmann equation. The Boltzmann equation gives the evolution of the distribution function, $f_i(x_\mu, p_\mu)$ for a particle of species i with position x_μ and momentum p_μ . In its most general form, the Boltzmann equation is formally

$$\frac{d}{dt} f_i = C[f_j, \dots] \quad (10.17)$$

where the left-hand side is the Liouville operator (or total derivative), d/dt , acting on the distribution function, and the right-hand side is the collision operator, which depends explicitly on the details of scattering amongst the particle types labelled by i . The Boltzmann equation is simply a statement of the Liouville theorem: in the absence of collisions, phase space is conserved ($df/dt = 0$). An example of this is the Newtonian continuity equation, $\partial\rho/\partial t + \nabla \cdot (\mathbf{v}\rho) = 0$. A full solution in the GR case is quite complicated, and we only give a flavor of the full derivation here. First, we calculate the Liouville operator by using the chain rule and the geodesic equation:

$$\frac{df}{dt} = \frac{\partial f}{\partial x^\mu} \frac{dx^\mu}{dt} + \frac{\partial f}{\partial p^\mu} \frac{dp^\mu}{dt} = p^\mu \frac{\partial f}{\partial x^\mu} - \frac{\partial f}{\partial p^\mu} \Gamma_{\alpha\beta}^\mu p^\alpha p^\beta, \quad (10.18)$$

where $\Gamma_{\alpha\beta}^\mu$ are the connection coefficients, which we calculate in, say, the longitudinal gauge in an expanding FRW background. We then expand around the solution to the spatially homogeneous zeroth-order equation with no collision term,

$$\frac{\partial f^{(0)}}{\partial t} - \frac{\dot{a}}{a} p \frac{\partial f^{(0)}}{\partial p} = 0. \quad (10.19)$$

where p is the magnitude of the photon wavevector (i.e., the photon frequency). If there are no collisions, $\partial f / \partial t = 0$, the solution is $f^{(0)}(t, p_\mu) = f^{(0)}(pa)$. For relativistic particles like photons, the distribution function is only a function of pa , another way of saying that the photon frequency redshifts with the expansion of the Universe.

If we subtract this zeroth order solution, fourier transform the x coordinates, convert the time coordinate to conformal time, η , defined by $d\eta = dt/a$, and ignore vector and tensor perturbations (discussed in the lectures by J. Bartlett on polarization at this school), the Liouville operator becomes a first-order partial differential operator for $f^{(1)}(\mathbf{k}, p, \eta)$, depending also on the general-relativistic potentials, Φ and Ψ . We further define the temperature fluctuation at a point, $\Theta(k, \mu) = f^{(1)}(T_0 \partial f^{(0)} / \partial T_0)^{-1}$ where T_0 is the average temperature and $\mu = \cos \theta$ in the polar coordinates for wavevector k .

The collision term, $C[f]$ is the rate for interactions to change $f(\mathbf{k}, p, \eta)$. For photons, this is from Thomson scattering off of electrons, with differential cross section

$$\frac{d\sigma}{d\Omega} = \frac{3}{8\pi} \sigma_T \frac{1}{2} (1 + \cos^2 \theta) = \frac{\sigma_T}{4\pi} \left[P_0(\cos \theta) + \frac{1}{2} P_2(\cos \theta) \right] \quad (10.20)$$

where σ_T is the total Thomson cross section and in the second equality we have rewritten the cross section in terms of the Legendre polynomials $P_\ell(\mu)$. The angular wavenumber, ℓ , corresponds to perturbations with an angular scale $\theta \sim 180^\circ/\ell$. The scattering at a point is due entirely to the monopole (P_0 term) and quadrupole (P_2) moments of the radiation field at that point. With this in mind, we define the moments $\Theta_\ell(k) = \int_{-1}^1 d\mu P_\ell(\mu) \Theta(k, \mu)$ (be aware that there are different conventions for the definition of Θ_ℓ in the literature). Putting all of this together, the Boltzmann equation becomes

$$\frac{\partial \Theta}{\partial \eta} + ik\mu\Theta + \frac{\partial \Psi}{\partial \eta} - ik\mu\Phi = -\frac{\sigma_T n_e}{a} \left[\Theta - \Theta_0 - \frac{1}{2} P_2(\mu) \Theta_2 + v\mu \right] \quad (10.21)$$

where v is the velocity of the baryons. In practice, this is turned into a hierarchy of equations for Θ_ℓ and solved numerically. Detailed solutions were investigated (Wilson and Silk, 1981; Silk and Wilson, 1981; Bond and Efstathiou, 1984), as was a more heuristic treatment (Hu and Sugiyama, 1995b; Hu and Sugiyama, 1995a; Hu and Sugiyama, 1996); Seljak and Zaldarriaga 1996; Zaldarriaga and Seljak, 2000; Zaldarriaga and Seljak, published a very fast algorithm for the solution and have made it available as the CMBFAST program.

Large Scales: The Sachs-Wolfe effect

Here, we can only investigate some of the simplest qualitative features of the solution. First consider the effect of free-streaming since recombination,

so the collision term $C \equiv 0$. This is just $df/dt = 0$, but now in a perturbed universe: $\Theta' + ik\mu\Theta = ik\mu\Phi$. The solution to this equation is

$$\Theta(\eta) - \Phi(\eta) = (\Theta_{\text{LSS}} - \Phi_{\text{LSS}})e^{-ik\mu(\eta - \eta_{\text{LSS}})} \quad (10.22)$$

where ‘‘LSS’’ refers to quantities on the surface of last scattering, and we have assumed that $\Phi = \text{const}$ as in a flat, matter-dominated FRW universe without a cosmological constant. Since we don’t have separate access to the potential today, the Φ term on the left-hand side just results in an overall change in the average temperature, so the two terms *in toto* give the observed temperature fluctuation. We will also simplify the calculation by considering only the largest scales, $k\eta \rightarrow 0$ (superhorizon scales at last scattering), so the observable temperature fluctuation is $\Theta_{\text{obs}} = (\Theta_{\text{LSS}} - \Phi_{\text{LSS}})$.

To complete the solution, we need to know the relationship between the potential, Φ , and the temperature, Θ , on the surface of last scattering. The easiest way to see this (Peacock, 1999; White and Hu, 1997) is to consider two different coordinate systems perturbed around an FRW background. The first is the longitudinal or Newtonian gauge in which our potentials Φ and Ψ are defined. In the other ‘‘comoving’’ frame, the time coordinate is defined as the proper time in the rest frame of the matter at a given point – in this system, density fluctuations vanish at a given time. The difference between the two coordinate systems is just the relativistic time-dilation effect: clocks run slower in a denser environment. From the definition of our perturbed metric, Eq. 10.15, clocks in the Newtonian frame must differ by an interval $ds^2 = (1 + 2\Phi)dt^2$ or $\delta s \simeq (1 + \Phi)dt$, so the potential gives the amount of time dilation, equivalent to the difference in times between the same event in the different coordinate systems: $\delta t/t = \Phi$. Since $T \propto 1/a$, changing from one time to another is equivalent to changing from one temperature to another, $\Theta = -\delta a/a$. But in a flat, matter-dominated universe, $t \propto a^{2/3}$, so $\Theta = -(2/3)\Phi$, so finally

$$\Theta_{\text{obs}}(\eta) = -\frac{1}{3}\Phi_{\text{LSS}} \quad (10.23)$$

This is the result first derived by Sachs and Wolfe (1967), which obtains (a) on large scales; (b) for adiabatic initial conditions, so $\Theta = -(2/3)\Phi$; and (c) in a Universe with $\Phi = \text{const}$. In particular, if we allow the potential to vary with time, we find in general $\Theta = -2 \int d\eta (\partial\Phi/\partial\eta)$ due to blue- and red-shifting as photons fall into and out of potential wells along the line of sight – the *Integrated Sachs-Wolfe effect* (ISW). In particular, if there had been *no* fluctuations on the surface of last scattering, $\Delta T/T = -2\Phi$ (Jaffe et al., 1994).

Small scales: Acoustic oscillations

On smaller scales, we must track the details of the interactions amongst the various components: photons, baryons and dark matter. Hu & Sugiy-

mama (1996) combined the Boltzmann equation with the Euler equation for the baryons to get a second-order differential equation linking the temperature fluctuation, Θ , to the potentials Φ and Ψ . Oscillations to the photon density are driven by gravitational potential fluctuations, damped by the expansion of the Universe, with a restoring force due to pressure — sound waves. As we saw above, the pressure force is decreased by baryon loading. This enables us to trace the evolution of fluctuations on small scales. We defer a detailed account of the evolution of these sound waves to sec 10.7 below.

6. The Power Spectrum of CMB Fluctuations

So far, we have discussed the evolution of perturbations in the dark matter, photons and baryons at any given point, for some particular set of initial conditions. However, our theories are not so specific. Rather, they give us only statistical information about the initial perturbations.

It turns out that everything will be easier to understand in Fourier space, so we first define our conventions:

$$\delta_{\mathbf{k}} = \int d^3r \delta(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}}, \quad \text{and} \quad \delta(\mathbf{r}) = \int \frac{d^3k}{(2\pi)^3} \delta_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{r}}. \quad (10.24)$$

In particular, we will rely on the power spectrum, defined as

$$\langle \delta_{\mathbf{k}} \delta_{\mathbf{k}'} \rangle = (2\pi)^3 \delta_D(\mathbf{k} + \mathbf{k}') P(k) \quad (10.25)$$

where δ_D is the Dirac delta function, whose arguments enforce the statistical isotropy of the density field (i.e., that the correlation function, $\langle \delta(\mathbf{r}) \delta(\mathbf{r}') \rangle = \xi(|\mathbf{r} - \mathbf{r}'|)$ is only a function of the distance between the points), and angle brackets refer to an “ensemble average,” which, depending on the context, may also be considered an ergodic or spatial average.

Similarly, we can transform the temperature fluctuations on the sky using the spherical harmonics, $Y_{\ell m}(\hat{\mathbf{x}})$

$$\frac{\Delta T}{T}(\hat{\mathbf{x}}) = \sum_{\ell=0, \dots} \sum_{m=-\ell, \ell} a_{\ell m} Y_{\ell m}(\hat{\mathbf{x}}), \quad \text{and} \quad a_{\ell m} = \int d^2\hat{\mathbf{x}} Y_{\ell m}^*(\hat{\mathbf{x}}) \frac{\Delta T}{T}(\hat{\mathbf{x}}). \quad (10.26)$$

We can then define the power spectrum of temperature fluctuations, C_{ℓ} , (Bond and Efstathiou, 1987)

$$\langle a_{\ell m} a_{\ell' m'} \rangle = C_{\ell} \delta_{\ell \ell'} \delta_{m, -m'} \quad (10.27)$$

where the Kronecker deltas enforce two-dimensional statistical isotropy on the sphere (the correlation function $\langle \Delta T(\hat{\mathbf{x}}) \Delta T(\hat{\mathbf{x}}') \rangle = C(\hat{\mathbf{x}} \cdot \hat{\mathbf{x}}')$ is only a function of the angular distance between the points). The power spectrum is related to

the three-dimensional temperature fluctuations defined in the previous sections as $C_\ell = (2/\pi) \int dk k^2 |\Theta_\ell(k, \eta_0)|^2$

In the previous section we saw that temperature fluctuations on the sky are the solution to a linear differential equation relating the temperature to the underlying potential fluctuations or more specifically to the primordial potential fluctuations (as created by, say, inflation). Similarly, the mass fluctuations, $\delta(\mathbf{r})$, are also linearly related to the primordial potential fluctuations.

As we said above, our theories provide only a statistical description of the temperature fluctuations on the sky. That is, the theories specify the probability density, $\Pr(a_{00}, a_{10}, a_{1-1}, a_{11}, a_{20}, \dots, a_{\ell m}, \dots)$, a function simultaneously of all the $a_{\ell m}$. For an isotropic theory with Gaussian fluctuations, this simplifies vastly to

$$\Pr(a_{\ell m}) = \frac{1}{\sqrt{2\pi C_\ell}} \exp \left[-\frac{1}{2} \frac{|a_{\ell m}|^2}{C_\ell} \right] \quad (10.28)$$

for all $m = -\ell, \dots, \ell$ at every ℓ . That is, the $a_{\ell m}$ have a Gaussian (normal) distribution with zero mean and variance given by C_ℓ . Similarly, for density fluctuations, the $\delta(\mathbf{k})$ have a Gaussian distribution with zero mean and variance $P(k)$.

But these distributions also hint at some fundamental limits to our measurements. At every ℓ , there are at most $2\ell + 1$ quantities that we can measure, the individual $a_{\ell m}$. Even with a perfect measurement, we need to infer from these quantities the “underlying” variance from which these quantities are drawn: the observed variance of the $a_{\ell m}$ at best provides an estimate of the power spectrum, C_ℓ . With a gaussian distribution, we can compute the variance of a single coefficient, $\text{var}(|a_{\ell m}|^2) \equiv \langle |a_{\ell m}|^4 \rangle - \langle |a_{\ell m}|^2 \rangle^2 = 3C_\ell^2 - C_\ell^2 = 2C_\ell^2$. If there are $2\ell + 1$ measurements at a given ℓ , the variance of the estimate is then $\sigma_\ell^2 = 2C_\ell^2/(2\ell + 1)$; for low ℓ , the error bar is comparable to the measurement. This is a lower limit, as well-instrument noise and systematic problems in realistic experiments will only increase the error.

It is typical in the study of cosmological perturbations to use not $P(k)$ and C_ℓ themselves, but rather the quantities $\Delta^2(k) = k^3 P(k)/(2\pi^2)$ and $\delta T_\ell^2 = \ell(\ell + 1)C_\ell/(2\pi)$. These “dimensionless” quantities give the contribution to the total variance in density or temperature from a given 3D or spherical wavenumber, or even more heuristically, the mean-square fluctuation at wavelength $\lambda \sim 2\pi/k$ or angular scale $\theta \sim 180^\circ/\ell$. In addition, the Sachs-Wolfe effect has $\ell(\ell + 1)C_\ell/(2\pi) = \text{const}$ at low ℓ .

7. The CMB and Cosmological Parameters

Finally, we can understand the behavior of the CMB power spectrum just defined using the physical processes and mathematics of the previous sec-

tions. On scales larger than the sound horizon at last scattering, only the initial potential perturbations have any effect on the power spectrum, and we see $\ell(\ell + 1)C_\ell \simeq \text{const}$ from the Sachs-Wolfe effect. Deviations from a constant arise from two sources. First, there is a contribution from the Integrated Sachs-Wolfe effect, if the potential is time-varying (at a fixed location), as in a Universe with $\Omega_m \neq 1$ and/or $\Omega_{\text{tot}} \neq 1$. Second, there may be contributions from sources of anisotropy other than scalar perturbations to the metric. In particular, the presence of a background of gravitational radiation would generically produce large-scale anisotropy; in practice the only way to distinguish this from scalar perturbations is by using polarization (see J. Bartlett's lecture).

On smaller scales, we consider the sound waves discussed in sec 10.5.0. First, consider waves entering the (sound) horizon around the time of last scattering: these are the largest waves that could have formed a coherent structure at this time. Indeed, by determining the characteristic *angular* scales of the CMB fluctuation pattern, and matching this to the physical scale of the sound horizon at last scattering we can determine the *angular diameter distance* to the last scattering surface, which is mostly dependent on the geometry of the Universe: in a flat universe, angular and physical scales obey the usual Euclidean formulae; in a closed (positively curved) universe, geodesics converge and a given physical scale corresponds to a larger angular scale (and hence smaller multipole ℓ); conversely, in a negatively curved Universe the same physical scale corresponds to a smaller angular scale and larger ℓ .

Consider now a wave that enters the horizon some time considerably before Last Scattering, when the density of the Universe is still dominated by radiation, and the Baryons are tightly-coupled to the photons. Although the dark matter is pressureless, the dominant radiation has pressure $p \simeq \rho/[3(1 + R)]$, where $R = (3/4)(\rho_B/\rho_\gamma)$ defines the “baryon loading” as before. Although the dark matter can continue to collapse, the radiation rebounds when the pressure and density become sufficiently high. Eventually, gravity may take over again and cause the perturbation to collapse yet again, one or more times. Larger and larger scales, entering the horizon at later and later times, will thus experience fewer and fewer collapse and rebound cycles. Moreover, because of the effect of the baryons on the pressure, the strength of the rebound is decreased as we increase the baryon density. It is this cycle of collapse and rebound that we see as peaks in the CMB power spectrum, often called *acoustic peaks* after the acoustic waves responsible for them. We thus use their heights to measure the relative contributions of baryon and photons to the pressure, and their angular scale to determine the geometry, as well as the history of the sound speed in the baryon-photon plasma. (And of course other cosmological parameters also affect the spectrum in yet other ways)

There are yet other physical effects that affect the pattern of CMB fluctuations. Although the photons and baryons are tightly bound to one another via scattering, the coupling is not perfect. Hence, there is a scale (known as the *Silk damping scale*) below which the photons can stream freely and wash out perturbations. This free-streaming damps perturbations on small scales.

All of these effects are included in codes like CMBFAST (Seljak and Zaldarriaga 1996; Zaldarriaga and Seljak, 2000, <http://www.cmbfast.org/>) and CAMB (Lewis et al., 2000, <http://camb.info/>) which solve the combined Boltzmann and linearized Einstein equations in an expanding Universe. These codes allow one to calculate the CMB temperature power spectrum for a given model. A sample of spectra for various input cosmological parameters is shown in Figure 10.2.

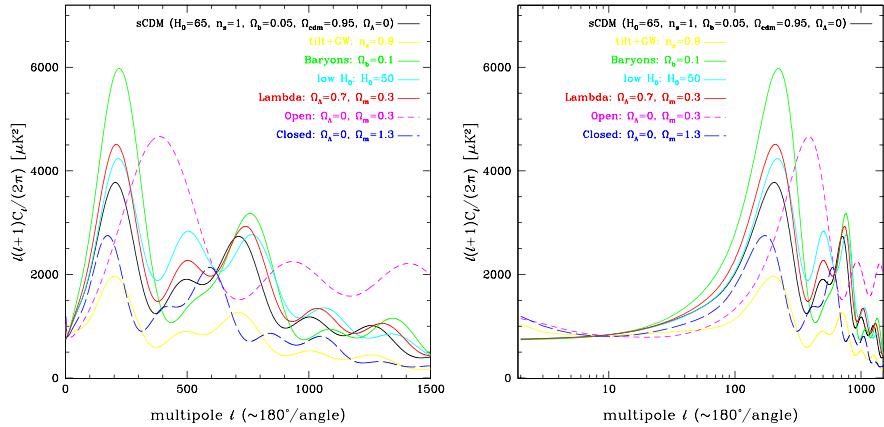


Figure 10.2. A sample of theoretical power spectra for various cosmological parameters, as marked.

The evolution of a single wavelength of perturbation is determined by a second-order differential equation and thus has two independent solutions, as we saw in our discussion of the evolution of perturbations in sec 10.4. When the Hubble length is smaller than the perturbation wavelength (which is the case at very early times, from soon after the perturbations creation in the inflationary epoch) only one of these independent solutions does not decay with time. As long as recombination occurs sufficiently long after the epoch of inflation (typically not a problem, since inflation must occur well before electroweak unification at $T_{ew} \sim 1$ TeV, so $T_{infl}/T_{lss} \gg 10^9$) all perturbations are “squeezed” into the same state. A result is that all acoustic oscillations of a given wavelength all have the same temporal phase. This coherence is important to achieving the multiple peak structures seen in Figure 10.2.

The matter transport caused by the pressure and gravitational potential gradients means there are *velocity* perturbations as well. Hence, the photons can scatter off of moving electrons, which generates a net linear polarization of the photons (Hu and White, 1997). For the sound waves we are considering, the velocities are greatest when the density contrast is smallest, and vice versa: the velocity is out of phase with the density—and hence the polarization signal is out of phase with the temperature. Unfortunately, due to the relative inefficiency of scattering off of the moving electrons, the polarization fraction is only about 10%, and the polarization spectra are correspondingly suppressed.

8. Conclusions

Of course, things get *really* exciting only when all of these theoretical constructs confront actual data. This task is taken up at some length in the lectures by Ned Wright at this school. Here, we must be satisfied with presenting the currently available data (Fig. 10.3), dominated by the WMAP experiment (Bennett et al., 2003), and a “best-fit” model calculated by the WMAP team. The most important features were, in fact, strongly indicated by the data available prior to WMAP [e.g., MAXIMA and BOOMERanG (Jaffe et al., 2001; Knox and Page, 2000)], and are:

- The general structure of a Sachs-Wolfe plateau at low ℓ , acoustic peaks at sub-degree scales, and damping at still smaller scales, seems to obtain. This is an extremely strong argument that the general paradigm of adiabatic initial conditions evolving under gravity and baryon-loaded radiation pressure holds.
- Within this paradigm, the results are consistent with a flat Universe ($\Omega_m = 1.02 \pm 0.02$) and a nearly scale-invariant initial spectrum of perturbations $n_s = 0.93 \pm 0.03$. Both of these are as expected if *inflation* acted in the early Universe.
- Combined with other cosmological measurements, the data seem to indicate a matter density $\Omega_m = 0.27 \pm 0.04$ and a “dark energy” density $\Omega_\Lambda = 0.73 \pm 0.04$.
- The data are consistent with other cosmological observations, including the Hubble constant, $H_0 = 71^{+0.04}_{-0.03}$ km/s/Mpc, and the baryon density, $\Omega_B h^2 = 0.0224 \pm 0.0009$.

There are, however, some puzzles and anomalies. The polarization data indicates that the Universe reionized at redshift $z_r \simeq 20^{+10}_{-9}$ rather than the expected $z_r \simeq 6$; there are also some outliers from the smooth best-fit power spectra at low and moderate ℓ which may just be statistical flukes but are suf-

ficiently unlikely that a physical explanation would be tantalizing (Niarchou et al., 2003).

We expect that all of these mysteries will be cleared up — and new ones generated — by the release of further data from WMAP expected in 2004 and beyond, many ongoing ground- and balloon-based observing campaigns, and eventually the Planck Surveyor satellite, due to be launched in 2007.

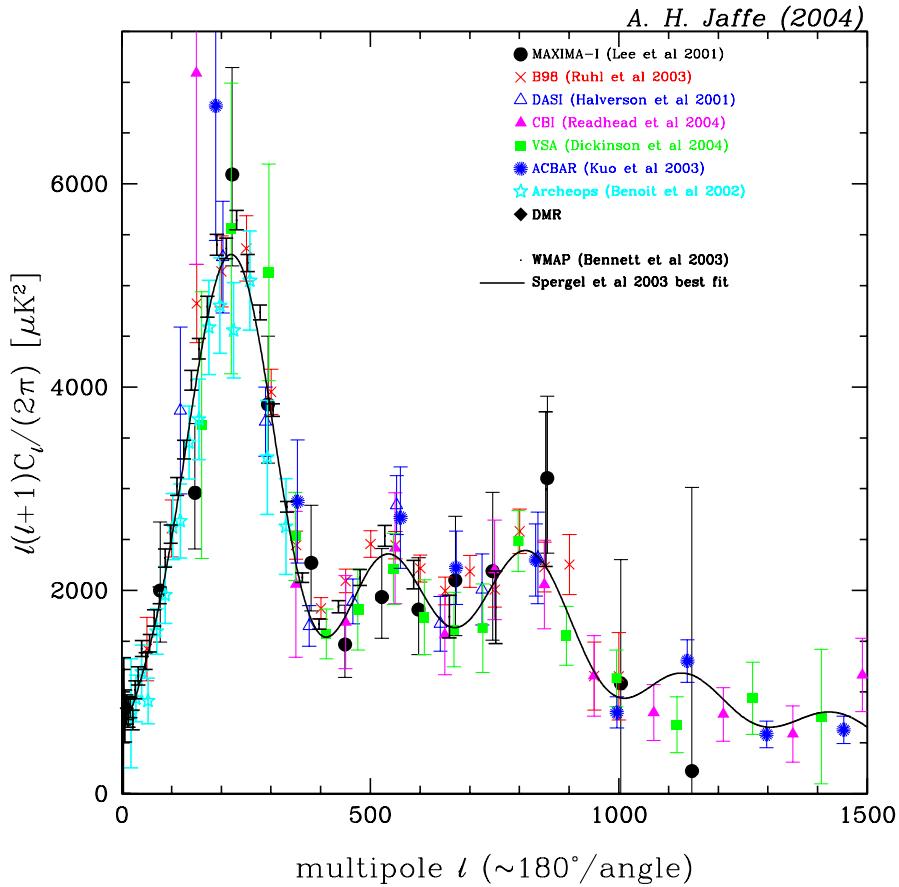


Figure 10.3. CMB anisotropy data circa March 2004.

Acknowledgments

I would like to thank the organizers of the Cargese school, and in particular Alain Blanchard, for a wonderful time. I would also like to thank my many CMB collaborators in the MAXIMA, BOOMERANG, COMBAT and Planck

collaborations for helping me to understand the physics of the CMB. Finally, I would like to thank Sarah Church and Lloyd Knox, in collaboration with whom some of sec 10.7 was originally written (Church et al., 2001). These lectures were prepared under funding from PPARC in the UK

References

- Bardeen, J. M. 1980, Phys. Rev. D, 22, 1882
 Becker, R. H., Fan, X., White, R. L., Strauss, M. A., Narayanan, V. K., Lupton, R. H., Gunn, J. E., Annis, J., Bahcall, N. A., Brinkmann, J., Connolly, A. J., Csabai, I., Czarapata, P. C., Doi, M., Heckman, T. M., Hennessy, G. S., Ivezić, Ž., Knapp, G. R., Lamb, D. Q., McKay, T. A., Munn, J. A., Nash, T., Nichol, R., Pier, J. R., Richards, G. T., Schneider, D. P., Stoughton, C., Szalay, A. S., Thakar, A. R. & York, D. G. 2001, AJ, 122, 2850
 Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Wollack, E., Wright, E. L., Barnes, C., Greason, M. R., Hill, R. S., Komatsu, E., Nolta, M. R., Odegard, N., Peiris, H. V., Verde, L. & Weiland, J. L. 2003, ApJS, 148, 1
 Bond, J. R. & Efstathiou, G. 1984, ApJL, 285 L45
 Bond, J. R. & Efstathiou, G. 1987, MNRAS, 226, 655
 Church, S., Jaffe, A. & Knox, L. 2001, CMB and Inflation: the report from Snowmass 2001 pages 11203–+.
 Dicke, R. H., Peebles, P. J. E., Roll, P. G. & Wilkinson, D. T. 1965, Cosmic black-body radiation. ApJ, 142, 414
 Djorgovski, S. G., Castro, S., Stern, D., and Mahabal, A. A. 2001, ApJL, 560, L5
 Efstathiou, G. 1990, In Peacock, J.A., Heavens, A.F., and Davies, A.T., editors, *Physics of the Early Universe, Proc. of the Scottish Universities Summer School in Physics*, volume 36 of *SUSSP Proceedings*, page 361, Edinburgh. SUSSP.
 Fan, X., Narayanan, V. K., Strauss, M. A., White, R. L., Becker, R. H., Pentericci, L. & Rix, H. 2002, AJ, 123, 1247
 Freedman, W. L., Madore, B. F., Gibson, B. K., Ferrarese, L., Kelson, D. D., Sakai, S., Mould, J. R., Kennicutt, R. C., Ford, H. C., Graham, J. A., Huchra, J. P., Hughes, S. M. G., Illingworth, G. D., Macri, L. M. & Stetson, P. B. 2001, ApJ, 553, 47
 Gunn, J. E. & Peterson, B. A. 1965, ApJ, 142, 1633
 Hu, W. & Dodelson, S. 2002, ARA&A, 40, 171
 Hu, W. & Sugiyama, N. 1995, ApJ, 444, 489
 Hu, W. & Sugiyama, N. 1995b, Phys. Rev. D, 51, 2599
 Hu, W. & Sugiyama, N. 1996, ApJ, 471, 542
 Hu, W., Sugiyama, N. & Silk, J. 1997 Nature, 386, 37

- Hu, W. & White, M. 1997, New Astronomy, 2, 323
- Hubble, E. 1929, *Proceedings of the National Academy of Science*, 15, 168
- Hubble, E. 1958, *The realm of the nebulae*. New York: Dover, 1958.
- Jaffe, A. H., Ade, P. A., Balbi, A., Bock, J. J., Bond, J. R., Borrill, J., Boscaleri, A., Coble, K., Crill, B. P., de Bernardis, P., Farese, P., Ferreira, P. G., Ganga, K., Giacometti, M., Hanany, S., Hivon, E., Hristov, V. V., Iacoangeli, A., Lange, A. E., Lee, A. T., Martinis, L., Masi, S., Mauskopf, P. D., Melchiorri, A., Montroy, T., Netterfield, C. B., Oh, S., Pascale, E., Piacentini, F., Pogosyan, D., Prunet, S., Rabii, B., Rao, S., Richards, P. L., Romeo, G., Ruhl, J. E., Scaramuzzi, F., Sforna, D., Smoot, G. F., Stompor, R., Winant, C. D. & Wu, J. H. 2001, Phys. Rev. Lett., 86, 3475
- Jaffe, A. H., Stebbins, A. & Frieman, J. A. 1994, ApJ, 420, 9
- Kamionkowski, M. & Kosowsky, A. 1999, Ann. Rev. Nucl. Part. Sci., 49, 77
- Knox, L. & Page, L. 2000, Phys. Rev. Lett., 85, 1366
- Kolb, E. W. & Turner, M. S. 1990, *The Early Universe*. Frontiers in Physics, Addison-Wesley, Reading, MA.
- Kosowsky, A. 1999, New Astronomy Review, 43, 157
- Kosowsky, A. 2002, *The cosmic microwave background*. In Modern Cosmology, 219.
- Lewis, A., Challinor, A. & Lasenby, A. 2000, ApJ, 538, 473
- Mukhanov, V.F., Feldman, H.A. & Brandenberger, R.H. 1992, Physics Reports, 215(5–6), 203
- Niarchou, A., Jaffe, A. H., and Pogosian, L. 2003, ArXiv Astrophysics e-prints.
- Peacock, J. A. 1999, *Cosmological physics*. Cosmological physics. Publisher: Cambridge, UK: Cambridge University Press, 1999. ISBN: 0521422701.
- Peebles, P. J. E. 1993, *Principles of physical cosmology*. Princeton Series in Physics, Princeton University Press, Princeton, NJ.
- Penzias, A. A. & Wilson, R. W 1965, ApJ, 142, 419
- Scott, D., Silk, J. & White, M. 1995, Science, 268, 829
- Seljak, U. & Zaldarriaga, M. 1996, ApJ, 469, 437
- Silk, J. & Wilson, M. L. 1981, ApJL, 244, L37
- Smoot, G. F., Gorenstein, M. V. & Muller, R. A. 1977, Physical Review Letters, 39, 898
- Tolman, R. C. 1934, *Relativity, Thermodynamics, and Cosmology* Oxford University Press, Oxford.
- White, M. & Hu, W. 1997, A&A, 321, 8
- Wilson, M. L. & Silk, J. 1981, ApJ, 243, 14
- Zaldarriaga, M. & Seljak, U. <http://www.cmbfast.org/>.
- Zaldarriaga, M. & Seljak, U. 2000, ApJS, 129, 431

Chapter 11

SUPERNOVAE AS ASTROPHYSICAL OBJECTS

Bruno Leibundgut¹

*European Southern Observatory
Karl-Schwarzschild-Strasse 2, D-85748 Garching
Germany
bleibundgut@eso.org*

Abstract Supernovae are a prominent component of modern astrophysics. They are responsible for a major part of the chemical enrichment in the universe and the main recycling mechanism in galaxies. The physics of these explosions is fairly well, although not completely, understood.

This article reviews the current understanding of the supernova explosions. After a brief historical introduction the two main classes of supernovae are described starting from the classification scheme currently employed. The different energy inputs for supernovae are presented. Despite their rather different energy sources supernovae from different types reach very similar luminosities. A notable exception to this are the Gamma-Ray Bursts, which are several orders of magnitude more energetic. The characteristics of each supernova type are presented.

Keywords: Supernovae; explosion physics; chemical enrichment

1. Some History

Stellar explosions are among the most violent events in the universe. The appearance of new stars, "stella novae" from their Latin designation, has always intrigued astronomers as documented in the ancient Chinese and Korean records (see Clark & Stephenson 1977 Clark & Stephenson 1977 and Murdin & Murdin 1978 Murdin & Murdin 1978 for reviews of the historic supernovae). The 'stella nova' in Cassiopeia, which appeared in November 1572, was observed by the young Tycho Brahe. It strongly impressed him and led him to conclude that the object, due to lack of an observable parallax, had to be outside the solar system and could not be explained in the Ptolemaic system of crystal spheres Brahe 1573. Some 150 years later Lundmark Lundmark 1925

suggested that there are two classes of novae, an 'upper class' which would be about 10 magnitudes brighter than a 'lower class' of novae. The latter would correspond to the well known Galactic novae. He based his proposal mostly on the observation of the nova 'S Andromeda' observed in 1885 (designated SN 1885A in modern nomenclature), which appeared that much brighter than a sample of about two dozen regular novae in the Andromeda galaxy. Lundmark later seemingly was the first to suggest the name 'super-nova' Lundmark 1932. Walter Baade and Fritz Zwicky Baade & Zwicky 1934 as well realised that stellar explosions come in different flavours.

It was Walter Baade who made the connection between the historical supernovae and the observed emission nebulae at their positions, thus identifying the remnants of the explosions. The most prominent object is of course the Crab Nebular (Messier 1), the leftover from the supernova in 1054 Baade 1942; Mayall & Oort 1942. With extensive observations of bright supernovae Minkowski Minkowski 1941 introduced two subclasses. Zwicky Zwicky 1965 refined the classification scheme for supernovae further. However, for several decades only two main classes were maintained until in the early eighties it became clear that at least one further subclass needs to be added.

The understanding of supernova physics is moving rapidly and at any given time only a snapshot is possible. The most recent reviews can be found in Hillebrandt & Leibundgut Hillebrandt. & Leibundgut 2003 and Weiler Weiler 2003. More specific reviews on Type Ia supernovae are Leibundgut Leibundgut 2000 and Meikle Meikle 2000. The Type II supernovae have been discussed in Arnett et al. Arnett et al. 1989 and McCray McCray 1993 in connection with SN 1987A.

2. Supernova classification

The modern classification of supernovae is based on the spectroscopy around maximum light (e.g. Filippenko 1997; Turatto 2003 - see also Fig. 11.1). The distinction is done through the presence (or absence) of hydrogen lines leading to the classes of Type II supernovae (or Type I supernovae). The hydrogen-deficient supernovae are further subdivided into groups which display prominent absorption near 6150Å attributed to Si II for the Type Ia supernovae and others which show Na and oxygen lines, called Type Ib/c supernovae. The presence/absence of helium lines is often argued for a separation into the Type Ib/Type Ic supernovae, respectively. One should note that these classifications are based on optical spectroscopy. The IR He I (λ 10830Å) line has been observed in many Type Ic supernovae by now. Moreover, there is one 'cross-over' class of Type IIf supernovae. These typically start out as hydrogen bearing supernovae before the hydrogen lines disappear and the objects start to resemble Type Ib/c supernovae. They are the major link showing that the SNe Ib/c are

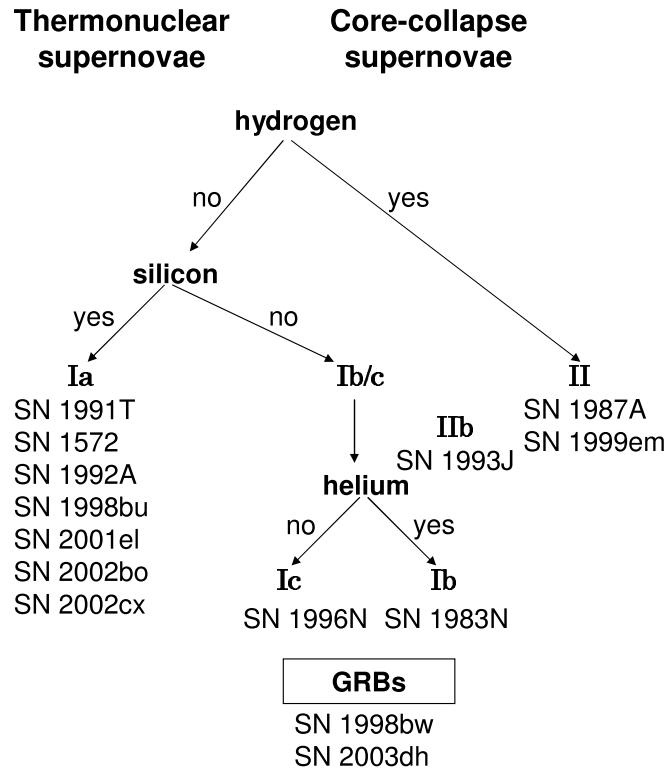


Figure 11.1. Classification scheme for supernovae

core-collapse objects. Figure 11.1 also lists some typical example for each SN class.

Most importantly, a physical picture for this classification scheme has emerged. The Type Ia supernovae (SNe Ia hereafter) are coming from thermonuclear explosions of objects, which have shed hydrogen and helium during their progenitor evolution. Hence no traces from these elements are observed in these explosions. All other supernovae most likely come from core-collapse explosions. The signature for these objects are their oxygen and calcium rich spectra at late phases. The general characteristics of the supernova types are shown in Table 11.1.

The entries in Table 11.1 can be traced back to observational results. The mass and size of the progenitor stars can be evaluated best from the bolometric light curves, i.e. the integrated energy radiated in the electromagnetic spec-

Table 11.1. Characteristics of supernova types

Thermonuclear Supernovae	Core-collapse Supernovae
from low-mass stars ($M < 8M_{\odot}$)	high-mass stars ($M > 8M_{\odot}$)
highly evolved stars (white dwarfs)	large envelopes (core still burning)
explosive carbon and oxygen burning	burning due to compression of the infalling material
binary system required	single stars (binaries for SNe Ib/c)
complete disruption	neutron star (or black hole) remnant

trum (e.g. Leibundgut. & Pinto 1992; Arnett et al. 1989. The presence of only higher elements in SNe Ia is an indication that they come from stars that have shed their envelopes a long time ago. The core-collapse in massive stars is triggered by the fact that only endo-thermal nuclear fusion is possible beyond the iron peak, the most tightly bound nuclei, and the pressure in the centre of the star is removed triggering the collapse. The burning mechanisms in the two types is entirely different. While SNe Ia are incinerated from the inside out, with a flame moving through the star, burning oxygen and carbon explosively up to the iron-peak elements, the core-collapse produces high-density and high-temperature zones near the surface of the forming neutron star where material is compressed and nuclear burning is initiated. Since the explosion in SNe Ia are suspected to come from white dwarfs, which are normally simply cooling by radiating away the energy they gain from the gravitational shrinking of the star, they need to be triggered into an explosion. The natural point is the situation where the electron pressure can no longer support the white dwarf, which happens when it reaches the Chandrasekhar mass Chandrasekhar 1931. Hence most models predict a companion star losing mass that is accreted by the white dwarf to grow to the Chandrasekhar mass. It should be noted that currently no consensus has been reached on what the companion star is (e.g. Branch et al. 1995; Livio 2000. For SNe II there is no need to invoke a companion star for the explosion. However, for the subtype of the SNe Ib/c where the hydrogen layer has been lost, a companion is often suggested as the mechanism to remove the hydrogen envelope. Contrary to the SNe Ia here the companion is not a donor, but used to expel the envelope. And finally, core-collapse supernovae do leave a compact remnant, either a neutron star or a black hole, while the thermonuclear explosions burn and disrupt the star entirely.

3. Input Energy

It is interesting to evaluate the energy sources of the two explosion mechanisms. Gravity is the drive behind the core-collapse supernovae. The collapse of about $1.5 M_{\odot}$ to nuclear densities or beyond release about 10^{53} erg. Most of this energy is radiated in anti-neutrinos, which escape in the formation of neutrons out of protons and electrons. About 10^{51} erg go into kinetic energy pushing the envelope away and only 10^{49} erg go into electromagnetic radiation signalling the death of the star across the universe. The thermonuclear explosions draw their energy from the energy difference of the binding energy of oxygen and carbon compared to the iron-peak elements. About one solar mass of O and C are burned and an energy of 10^{49} erg is released in electro-magnetic radiation.

There are several effects that can influence the electro-magnetic display of supernovae. Shocks further convert kinetic energy into radiation. Some of the radiation can not escape the dense explosion and only when the debris expand and adiabatically cool is some of it released. Energy that went into ionising the envelope is released when the material cools down enough so that the atoms recombine again. For supernovae with extended envelope this recombination can create an extended plateau phase in the light curve, where the expansion of the atmosphere is balanced by the inward moving wave of recombination. One of the best observed examples is SN 1999em Hamuy et al. 2001; Elmhamdi et al. 2003 where the plateau lasted for about 100 days.

However, the largest energy reservoir is stored in radioactive isotopes that release γ -rays after typical decay times. The most important channel is the γ -decay of ^{56}Ni into ^{56}Co and then stable ^{56}Fe (e.g. Diehl & Timmes 1998). For the core-collapse supernovae this channel provides the energy input for the late light curves (after the plateau phase), while it is the only energy input for SNe Ia Leibundgut & Suntzeff 2003. Bolometric light curves can be used to track the change in escape fraction of the γ -rays from the supernova ejecta Leibundgut. & Pinto 1992; Contardo et al. 2000.

For massive supernovae the absolute luminosity after about 120 days, together with the age of the supernova, gives a relatively accurate measure of the amount of ^{56}Co synthesised in the explosion Hamuy et al. 2003B; Elmhamdi et al. 2003. This measurement is now available for many core-collapse supernovae and is typically a factor 10 less than assumed in thermonuclear supernovae but spans almost a factor of 100 Pastorello et al. 2004.

The long and rich light curve observed for SN 1987A is a clear demonstration of how the various physical effects form the light curve Leibundgut & Suntzeff 2003. It shows many of the described features and some more.

Hypernovae have been added to the list of supernovae and they represent the high energy end (at least in their kinematics) with the high velocities observed

in these objects. The connection of gamma-ray burst with supernovae has now been generally accepted with the observations of SN 2003dh/GRB030329 (e.g. Stanek et al. 2003; Matheson et al. 2003; Hjorth. et al. 2003. It should be noted that already SN 1998bw/GRB980425 showed all the signatures of a supernova Galama et al. 1998; Patat et al. 2001. These hypernovae are characterised by the absence of hydrogen and helium and very high expansion velocities observed in the spectra Mazzali et al. 2002; Mazzali et al. 2003. In some cases no gamma-ray burst is observed, like for SN 2002ap. The amount of nickel synthesised in these explosions is substantial, up to about $0.5 M_{\odot}$ Sollerman et al. 2002. The kinetic energies inferred from the line widths are also substantially larger than the ones of regular SNe II. In many aspects they appear to be similar to the SNe Ib/c with high kinetic energy.

4. Core-collapse supernovae

The richness in appearance of the core-collapse supernovae is due to their varied progenitor histories. An example of the appearance of SNe II is given in Figure 11.2. The spectra have all been observed near maximum light and show $H\alpha$ in these objects. The typical light curves of SNe II displays a long plateau of about 100 days after the maximum. The most prominent and best observed case after SN 1987A is SN 1999em Hamuy et al. 2001; Elmhamdi et al. 2003. While SN 1987A displays a strong P Cygni line, it is almost not visible in SN 1993J. The latter lost this line in its evolution completely and only after about one year did $H\alpha$ reemerge in the nebular spectrum Filippenko et al 1994. The case of SN 1988Z is different again. In this case, the hydrogen is excited in circumstellar material shocked by the supernova shock. The time of explosion for this object is not known very well Stathakis & Sadler 1991 and it is likely that it exploded several months before it was discovered. The emission is dominated by the shock energy and not recombination or radioactive decay as in all other supernovae.

This special case of supernovae interacting early on with their dense circumstellar environment is discussed in Chevalier & Fransson Chevalier & Fransson and Leibundgut Leibundgut 1994. They typically have very slow light curves and spectra that show emission lines but very little absorption. The best studied cases so far are SN 1986J Leibundgut et al. 1991, SN 1988Z Turatto 1993, SN 1995N Fransson 2002 and SN 1998S Fassia et al. 2000. These objects often can be observed for many years and show very little evolution.

Most Gamma-Ray Bursts are now also classified as supernovae. Their connection became clear with the observations of SN 1998bw/GRB980425 Galama et al. 1998; Patat et al. 2001 and more recently with SN 2003dh/GRB030329 Hjorth. et al. 2003; Matheson et al. 2003. Other examples include SN 2002ap Mazzali et al. 2002 and SN 1999cy Germany et al. 2000; tur00. These objects

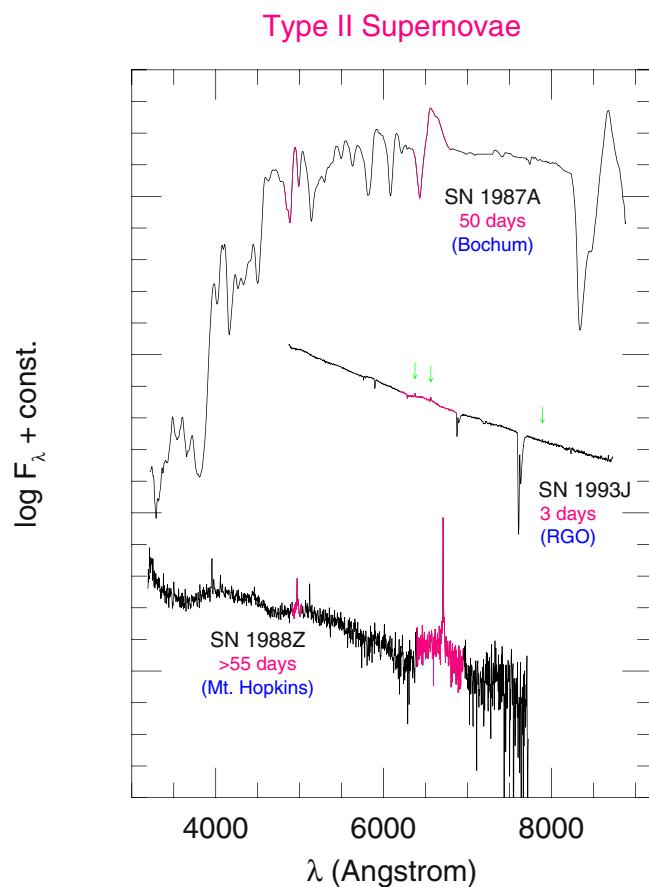


Figure 11.2. Examples of Type II supernova spectra. The references for the spectra are SN 1987A: Hanuschik & Dachs 1988; SN 1993J: Lewis et al. 1994; SN 1988Z: Filippenko 1991.

probably have a core that collapses to a black hole and produce relativistic jets. The interaction of these jets with the stellar atmosphere and surrounding material produce the γ -ray flashes and the optical afterglows McFadyen et al. 2003.

5. Type Ia supernovae

Although thermonuclear supernovae have simpler underlying physics than the core-collapse supernovae, there still remain formidable hurdles to fully understand them Hillebrandt & Niemeyer 2000. The observational material that has been assembled in the last decade is considerable and many nearby supernovae are now observed with exquisite detail. The last four years have seen dramatic progress in recognising peculiar events and also determining specific characteristics. The situation a few years ago is described in Leibundgut Leibundgut 2000. Since then the peculiar SN 2000cx Li et al. 2001; Candia et al. 2003 and SN 2002cx Li et al. 2003 have been observed. A truly particular case has been discovered in SN 2001ic Hamuy et al. 2003a, which displayed a strong, broad-lined H α emission after about 90 days past the maximum. This supernovae early on had all the signatures of a bright SN Ia with what looked like residual H α emission from the host galaxy. The spectral sequence later showed that the hydrogen emission is intrinsic to the supernova and indicates that this explosion occurred inside a dense hydrogen cocoon. These objects have thrown a dark shadow over the light curve – luminosity relations that have been used in the past to normalise the peak luminosity Phillips et al 1999; Riess et al. 1998; Perlmutter et al. 1997 and derive cosmological distances. They highlight the fact that not all SNe Ia are identical. The differences provide us with a tool to further investigate the true nature of these explosions. Of most import are currently the distribution of the synthesised nickel mass in the explosion, the mass of the progenitor star - and hence the ejecta - and the explosion energy. These affect the light curves in very specific ways Pinto & Eastman 2000.

The ability to determine a normalised peak luminosity for these objects has allowed us to measure accurate cosmological distances. The implications of this capability and the inferred cosmological results will be described in the following paper.

6. Conclusions

The term supernovae describes very different explosive processes in astronomy. There are essentially two object classes that provide an observable display that is rather similar. One is the core-collapse in a massive star where the freed gravitational energy is turned into radiation in many different ways. The rich variety of core-collapse supernovae is due to the many evolutionary chan-

nels that lead to such explosions. In addition, core-collapse in massive stars occurs in rather dense astronomical environments – near or in star forming regions – and in stars at different evolutionary phases at the outside.

The thermonuclear explosions are much closer to what has been termed a 'standard bomb' as they presumably are coming from a highly evolved stellar state, white dwarfs, that represent an old population.

The physics of the explosions are not completely understood in both cases. For the core-collapse supernovae there are also many effects which influence the appearance of the supernovae Leibundgut & Suntzeff 2003. The variations are a possible way to improve our understanding of these objects.

References

- Arnett, W.D., Bahcall, J.N., Kirshner, R.P. & Woosley, S. E. 1989, ARA&A, 27, 629
 Baade, W. 1942, ApJ, 46, 188
 Baade, W. Zwicky, F., 1934, Publ. Nat. Acad. Sci., 20, 254
 Brahe, T. 1573, De Stella Nova, e.g. in J.L.E. Dreyer ed., Tychonis Brahe Dani Opera Omnia (Copenhagen: Libraria Gyldendaliana (1913-1929))
 Branch, D., Livio, M., Yungelson, L.R., Boffi, F.R. & Baron, E. 1995, PASP, 107, 1019
 Candia, P., et al. 2003, PASP, 115, 277
 Chandrasekhar, S. 1931, MNRAS, 91, 456
 Chevalier, R.A. & Fransson, C. 2003, in Supernovae and Gamma-Ray Bursts, ed. K. Weiler, Heidelberg: Springer, 171
 Clark, D. J. & Stephenson, F. 1977, *The Historical Supernovae*, New York: Pergamon Press
 Contardo, G., Leibundgut, B. & Vacca, W.D. 2000, A&A, 359, 876
 Diehl, R. & Timmes, F. X. 1998, PASP, 110, 637
 Elmhamdi, A. et al. 2003, MNRAS, 338, 939
 Fassia, A. et al. 2000, MNRAS, 318, 1093
 Filippenko, A. V. 1991, SN 1987A and Other Supernovae, eds. I.J. Danziger & K. Kjär, (Garching: ESO), 343
 Filippenko, A.V., Matheson, T. & Barth, A.J. 1994, AJ, 108, 2220
 Filippenko, A.V. 1997, ARA&A, 35, 309
 Fransson, C. et al. 2002, ApJ, 572, 350
 Galama, T.J. et al. 1998, Nature, 395, 670
 Germany, L.M., Reiss, D.J., Sadler, E.M., Schmidt, B.P. & Stubbs, C.W. 2000, ApJ, 533, 320
 Hamuy, M. et al., 2001, ApJ, 558, 615
 Hamuy, M. et al. 2003a, Nature, 424, 651
 Hamuy, M. 2003b, ApJ, 582, 905

- Hanuschik, R.W. & Dachs, J. 1988, A&A, 205, 135
- Hillebrandt, W. & Leibundgut, B. 2003, From Twilight to Highlight: The Physics of Supernovae, Heidelberg: Springer
- Hillebrandt, W. & Niemeyer, J. 2000, ARA&A, 38, 191
- Hjorth, J. et al. 2003, Nature, 423, 847
- Leibundgut, B., in Circumstellar Media in Late Stages of Stellar Evolution, eds. R. Clegg, I. Stevens & P. Meikle, (Cambridge: Cambridge University Press), 100
- Leibundgut, B. 2000, A&AR, 10, 179
- Leibundgut, B., et al. 1991, ApJ, 372, 531
- Leibundgut, B. & Pinto, P.A. 1992, ApJ, 401, 49
- Leibundgut, B. & Suntzeff, N.B. 2003, in Supernovae and Gamma-Ray Bursts, ed. K. Weiler, Heidelberg: Springer, 77
- Lewis, J. R. et al. 1994, MNRAS, 266, L27
- Li, W., et al. 2001, PASP, 113, 1178
- Li, W., et al. 2003, PASP, 115, 453
- Livio, M. 2000, in *Type Ia Supernovae: Theory and Cosmology*, eds. J.C. Niemeyer and J.W. Truran, Cambridge, Cambridge University Press, 33
- Lundmark, K.E. 1925, MNRAS, 85, 865
- Lundmark, K.E. 1932, Lund Observatory Circ. 8
- Matheson, T., et al. 2003, ApJ, 599, 394
- Mayall, N.U. & Oort, J.H. 1942, PASP, 54, 95
- Mazzali, P., et al. 2002, ApJ, 572, L61
- Mazzali, P., et al. 2003, ApJ, 599, L95
- McCray, R. 1993, ARA&A, 31, 175
- McFadyen, A.I., Woosley, S.E. & Heger, A. 2001, ApJ, 550, 410
- Meikle, W.P.S. 2000, MNRAS, 314, 782
- Minkowski, R. 1941, PASP, 53, 224
- Murdin, P. & Murdin, L. 1978, *Supernovae*, Cambridge: Cambridge University Press
- Pastorello, A., et al. 2004, MNRAS, 347, 74
- Patat, F., et al. 2001, ApJ, 555, 900
- Perlmutter, S., et al. 1997, ApJ, 483, 565
- Phillips, M.M., Lira, P., Suntzeff, N.B., Schommer, R.A., Hamuy, M. & Maza, J., 1999 AJ, 118, 1766
- Pinto, P. A. & Eastman, R. G. 2000, ApJ, 530, 757
- Riess, A. G., et al. 1998, AJ, 116, 1009
- Sollerman, J., et al. 2002, A&A, 386, 944
- Stanek, K.Z., et al. 2003, ApJ, 591, L13
- Stathakis, R.A. & Sadler, E.M. 1991, MNRAS, 250, 786
- Turatto, M., et al. 1993, MNRAS, 262, 128

- Turatto, M. 2003, in Supernovae and Gamma-Ray Bursts, ed. K. Weiler, Heidelberg: Springer, 21
- Turatto, M., et al. 2000, ApJ, 534, L57
- Weiler, K. 2003, Supernovae and Gamma-Ray Bursts, Heidelberg: Springer
- Zwicky, F. 1965, Stellar Structure, eds. L.H. Aller and D.B. McLaughlin, (Chicago: University of Chicago Press), 367

Chapter 12

COSMOLOGY WITH SUPERNOVAE

Bruno Leibundgut¹

*European Southern Observatory
Karl-Schwarzschild-Strasse 2, D-85748 Garching
Germany
bleibundgut@eso.org*

Abstract Modern cosmology is using many different methods to determine the structure of the universe. Supernovae are among the most important ones due to their extreme luminosity, the time-variability that allows to separate the different supernova explosions and relative ease with which they can be observed.

Since the recognition of supernovae as a separate class of astrophysical objects they have been proposed and used to measure the distance scale and the geometry of the universe.

There are several independent applications with supernovae to measure the current expansion rate, Hubble's constant, and the expansion history of the universe. The latter has led to the surprising discovery that the expansion is actually accelerating and a new component for the universe is needed. Supernovae are also poised to be a major player in the characterisation of the nature of the dark energy.

Keywords: Cosmology - supernovae - distance scale

1. Introduction

Cosmology with supernovae has developed over the second half of the last century. Their extreme luminosities always made them attractive candidates to measure large distances. Various methods were devised to use supernovae to measure cosmological parameters ranging from simple standard candle paradigms to physical explanations of the supernova explosions and subsequent derivation of distances. Essentially, supernovae have been used to determine luminosity distances, i.e. the comparison of the observed flux to the total emitted radiation. The trick is to find a reliable way to measure the absolute luminosity of the objects.

There are two major cosmological parameters that can be determined through supernova observations. They are the classical parameters, which govern the expansion of the universe in Friedmann-Robertson-Walker models: the Hubble constant, H_0 , and the deceleration parameter, q_0 . The former sets the scale of the universe and the magnitude of the current expansion of space and the latter describes the change of the expansion with time (e.g. Sandage 1961; Sandage 1988; Weinberg 1972; Peebles 1993; Peacock 1999). There is a rich literature on the Hubble constant and Type Ia supernovae (SNe Ia) (see Branch & Tammann 1992; Branch 1998; Leibundgut 2001 for reviews). The deceleration parameter has been replaced by more modern formulations specifically including the cosmological constant or some variants thereof Carroll et al. 1993.

2. The Hubble constant

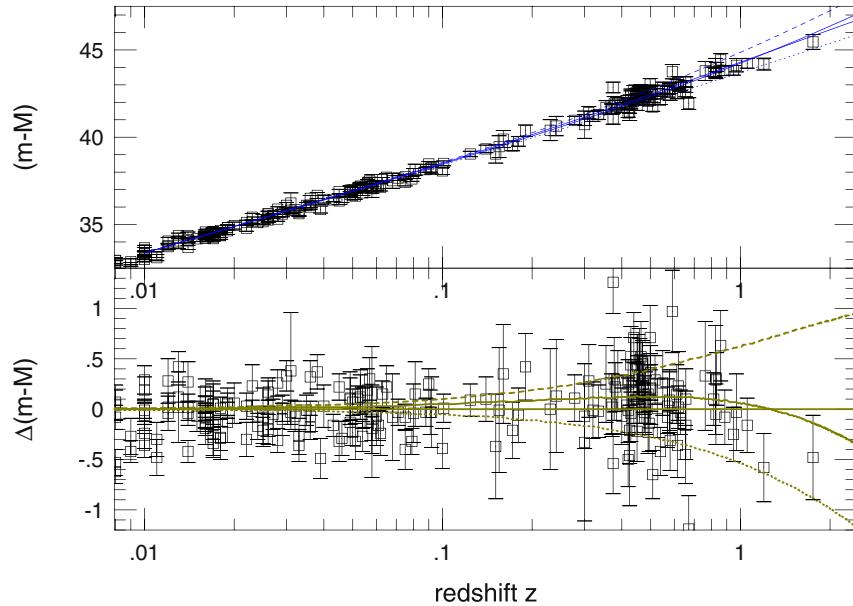
Type Ia supernovae

The best way to show that objects provide good luminosity distances is to plot them in a Hubble diagram. Originally, this diagram was using recession velocity *vs.* apparent magnitude Hubble 1936; Sandage 1961. The underlying assumptions are that the Hubble law holds and that the objects are all of the same luminosity, i.e. standard candles, so that the apparent brightness directly reflects distance.

Since no astronomical standard candle is known – all proposed objects have been shown to be essentially non-uniform in one way or another – we nowadays have to calculate and plot the distance modulus for the objects. The scatter around the linear expansion line is less than 0.2 magnitudes or 20% Tonry et al. 2003. Independent of our ignorance of the exact explosion mechanism or the radiation transport in the explosions this proves that SNe Ia can reliably be used as a distance indicator in the local universe. This situation is very much comparable to the Cepheid stars, where the period-luminosity relation is based on empirical data of objects in the Magellanic Clouds.

The simplest method is to assume that all supernovae are identical. This is, of course, not true (see previous paper) but it turns out that the subclass of the Type Ia supernovae is indeed rather homogeneous. The first to plot a Hubble diagram of Type Ia Supernovae was Kowal Kowal 1968. There are essentially three quantities that can be derived from such a Hubble diagram in the nearby universe: the slope of the expansion line, the scatter around the expansion line and the value of the local Hubble constant from the intercept at zero redshift (e.g. Tammann & Leibundgut 1990; Leibundgut. & Pinto 1992). The slope gives an indication of the local expansion and for a linear expansion in an isotropic universe it has a fixed value. The scatter around the expansion line provides a measure of the accuracy of the standard candle and the measurement errors. The intercept of the line, finally, together with an estimate of the

Figure 12.1. Hubble diagram of nearby Type Ia supernovae. The distances are derived from light curve shape corrected luminosities (data from Tonry et al. 2003). The solid line is for an empty universe ($\Omega_\Lambda = \Omega_M = 0$), the dotted line for an Einstein-de Sitter model ($\Omega_\Lambda = 0, \Omega_M = 1$) and the dashed line for a model with no matter and all cosmological constant ($\Omega_\Lambda = 1, \Omega_M = 0$). The concordance model ($\Omega_\Lambda = 0.7, \Omega_M = 0.3$) is shown as the line fitting the data best. The bottom panel shows all distances relative to the empty universe model.



absolute luminosity gives the Hubble constant. It has become clear that SNe Ia in the optical are not standard candles and have their luminosity normalised to be used as distance indicators Phillips 1993; Hamuy et al. 1996; Riess et al. 1996; Riess et al. 1998; Perlmutter et al. 1997; Phillips et al. 1999; Goldhaber et al. 2001. A systematic comparison of these different methods has been done Drell et al. 2000; Leibundgut 2000 and it has been shown that they are not internally consistent. The reason for this technical problem is still not clear and needs to be investigated. More accurate light curve data are becoming available (e.g. Krisciunas et al. 2001; Li 2001; Li 2003; Krisciunas et al. 2003; Candia et al. 2003; Benetti 2004) and it should be possible to further investigate the correlations between light curve shape, colour and luminosity of SNe Ia. For these reasons modern Hubble diagrams show the distance modulus ($m-M$) rather than the directly observed apparent magnitude m . The latest version with over 200 SNe Ia has been assembled by Tonry (Tonry et al. 2003 see also Barris 2004; Riess et al. 2004b) and is shown in Figure 12.1. It should be noted that SNe Ia may be standard candles in the near-infrared Krisciunas et al. 2004. The first significant IR sample shows very small scatter without prior

correction for light curves shape. For the derivation of the Hubble constant the (normalised) luminosity of the SNe Ia has to be known. The most direct way to achieve this is through the distance ladder and in particular the calibration of nearby SNe Ia by Cepheids (for the most recent results see Saha et al. 1999; Freedman et al. 2001). The main discrepancy for the published values of the Hubble constant from SNe Ia is coming from the different interpretations of the Cepheids and application of the light curve shape correction.

Core-collapse supernovae

The brilliance of core-collapse supernovae has enticed people to investigate their capabilities as distance indicators as well. Following early work by Baade 1926, originally done for Cepheid stars, the expanding photosphere method (EPM, Eastman et al. 1996) has been applied to several supernovae. The most comprehensive data sample has been assembled by Hamuy 2001. A critical test has become the distance to SN 1999em, which was determined through EPM Leonard et al. 2001; Hamuy et al. 2002; Elmhamdi et al. 2003 and which also has a Cepheid distance available Leonard et al. 2003. The discrepancy is most likely attributable to the fact that the correction factor for the dilution of the black body flux in EPM are strongly model dependent and need to be calculated for each supernova individually.

Recently, Mario Hamuy has realised that the expansion velocity and the luminosity during the plateau phase correlate and that Type II SNe may be quite good distance indicators Hamuy. & Pinto. 2002. The distance accuracy achieved this way can be better than 20%. These determinations are based on the physical understanding of the plateau phase of SNe II and are linked to physics of the supernova atmosphere. This means that they are independent of the *distance ladder*, which is needed, e.g., for the SNe Ia. Typical values for the Hubble constant from SNe II are in the range of 65 to 75 km s⁻¹ Mpc⁻¹ Hamuy 2003.

3. The expansion history of the universe

The recent paradigm changes in the cosmological model are based on several new insights. The flatness of space-time as measured by the cosmic microwave background (CMB) fluctuations and the recognition, that the global matter density is near 30% of the critical density, require an additional component in the energy content of the universe. At the same time the observation that the luminosity distances derived from Type Ia supernovae are larger than the expectation in any non-accelerated universe model have conspired to change our view of the history of the cosmic expansion. The three measurements are complementary to each other and the combination of any two of them provides independent evidence for an additional component in the Friedman equation.

However, only the supernova measurement gives a direct indication that we need a repulsive component in the universe. It will also be the supernovae that will provide a first indication to the nature of the dark energy.

There are several fundamental tests that will need to be performed until we can be sure that the current paradigm will persist. It is very appealing to think we know all constituents of the universe by now, but further surprises may still be in store for us. The testing has to concentrate on the reliability of the individual measurements. The Type Ia supernovae have been criticised for the fact that they are based on a rather simple assumption, namely that the distances derived from them are accurate. Many publications oversimplify this picture by calling Type Ia supernovae standard candles. This is not only incorrect, also it is misleading and belittles the result. The tests done on supernovae are solid and the theoretical work is progressing steadily.

Since SNe Ia are not standard candles, it is not admissible to simply assume a constant luminosity. Instead, one has to adopt that the luminosity normalisation of the distant objects follows what has been found in the nearby sample. Although the High-z Supernova Search Team (HZT; Schmidt et al. 1998; Riess et al. 1998; Tonry et al. 2003; Barris 2004) and the Supernova Cosmology Project (SCP; Perlmutter et al. 1999; Knop et al. 2004) make this assumption in different forms, it is essentially identical. The SCP derives the corrections from all supernovae in their sample, i.e. nearby and distant ones, while the HZT derives the correlations from the (large) nearby sample and applies it to the distant objects (cf. Leibundgut & Suntzeff 2003). It is also interesting to note that the SCP claims that both light curve shape correction and correction for host galaxy reddening affect their result rather little Perlmutter et al. 1999; Knop et al. 2004. On the other hand, the normalisation and absorption correction done by the HZT (in three different implementations) are important for the cosmological result. This discrepancy between the two teams will need to be resolved at some point.

4. Universal acceleration according to Type Ia supernovae

We will describe here the current status of the supernova research and outline ongoing projects to distinguish between a cosmological constant or a vacuum density contribution to the energy-momentum tensor in the Einstein equation.

Type Ia supernovae measure luminosity distances to objects out to about a redshift of 1. These distances are the most accurate currently available to astronomers for cosmological purposes, i.e. beyond the Coma cluster distance. Since the luminosity distances depend on the evolution of the Hubble parameter and this in turn depends on the energy content of the universe through the Einstein equation (e.g. Carroll et al. 1993) one can derive the energy

sources dominating over the lookback time covered by the observations (see ; Leibundgut 2001; Perlmutter et al. 2003 for a detailed review). Once the luminosity distances are derived from the supernova data a likelihood calculation provides the most statistically suitable values for the complete supernova data under certain assumptions, like the neglect of dust and evolution. It is pointless to divide the supernova data into subsamples that do not cover the complete redshift range as the effect is not detectable on smaller scales. Figure 12.1 shows that the current data by far do not warrant such a treatment (as proposed by Padmanabhan & Choudhury 2003; Choudhury & Padmanabhan T. 2004; Wang & Mukherjee 2004).

The largest available data set is provided by Barris et al. Barris 2004, which includes a significant set of supernovae at redshifts near 1. This data set confirms the earlier results of the HZT (cf. Fig. 12.2) and is consistent with the most recent result of the SCP Knop et al. 2004. All astrophysical effects, like dust or evolution of the supernovae, have been ignored in this derivation. The HZT applies a correction for dust in the Milky Way and the host galaxy of the supernova directly. Only if dust at high redshift is systematically different from the one in our galaxy, is this correction biased. Recent detection of $850\mu\text{m}$ emission from host galaxies at $z \approx 0.5$ shows that dust is present in some of these galaxies Farrah et al 2002, although the amount may be negligible for the supernova cosmology. These results are in contrast with the claim by Sullivan et al. Sullivan et al. 2003 that the reddening of distant supernovae in spiral galaxies is very small, when these objects are compared to the SN data from dust-free elliptical galaxies. The reddening derived by Tonry et al. Tonry et al. 2003 for distant SNe Ia is typically smaller than the one found for nearby objects. This is not surprising considering that the distant searches are mostly flux limited and will not find many heavily extincted objects, while the nearby supernovae are drawn from a large heterogeneous sample, which in several cases includes highly reddened objects. There might be secondary selection effects at work as well, like the fact that the distant supernovae often have larger projected distances from their hosts than nearby ones. A last indication that dust is not a severe problem is the fact that among the first distant SNe Ia of the HZT were very blue Leibundgut 2001 objects. In fact, six out of nine objects were bluer than their nearby counterparts. Although this has now been claimed to possibly be a selection effect Knop et al. 2004, it is not clear whether this indeed is the case, as the effect should be redshift dependent, which it seems not to be, judging from the small sample in Leibundgut 2001. The recent publication by the SCP Knop et al. 2004 does not find the same effect for the objects which have multi-wavelength light curves. Further analysis of the K-corrections and the dust properties is clearly required.

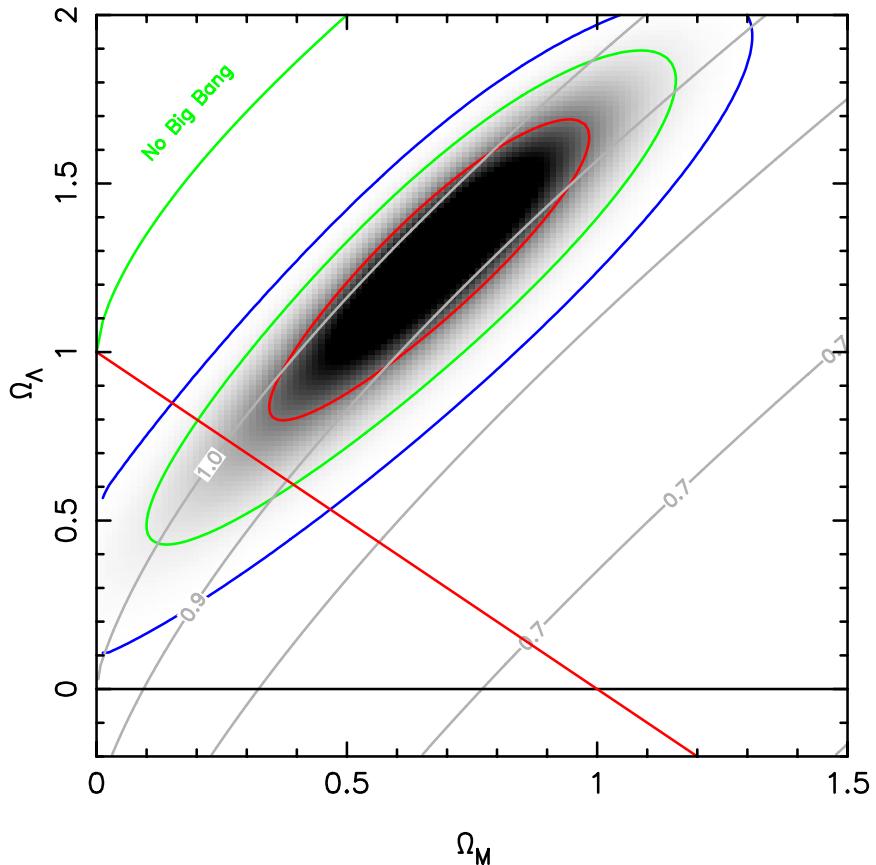
Evolution is another potential effect, which could mimic a cosmological signal. This is much harder to control. For reasonable predictions of how

progenitor metalicity or age could affect the brightness of SNe Ia one needs a detailed model of the explosion and the radiation escape from the explosion Hillebrandt & Niemeyer 2000. Both are unsolved problems. A detailed study of the properties of the SN host galaxies has not shown any correlation with supernovae distances or properties Williams et al. 2003. Progress can only be made through detailed observations of bright, nearby SNe Ia at all phases. Recent data sets are very encouraging Krisciunas et al. 2001; Krisciunas et al. 2003; Benetti 2004 (for a review see Leibundgut & Suntzeff 2003). In addition to the detailed spectral, light and colour curve data one can use bolometric light curves to derive the total emitted radiation from the explosion Contardo et al. 2000; Contardo 2001. The latter provides important information on the physical parameters that govern the explosion, like mass of synthesised nickel and the γ -ray escape fraction at late times.

With no clear indication of evolution, the simplest assumption is to disregard any evolutionary effects; a very dangerous approach, if it goes unchecked. This is the reason that the HZT has spectroscopically confirmed its distant SNe Ia. The spectra have been published together against the light curve data Riess et al. 1998; Tonry et al. 2003; Barris 2004 and separately for a few objects Coil et al. 2000; Leibundgut & Sollerman 2001. While the signal for some of the distant supernovae is not very good, and in a few early cases the SN classification may even be doubtful, there are no obvious strong deviations from the spectral appearance of the nearby supernovae. In some cases, the supernova spectra can be used to determine the phase of the distant SNe Ia and to check it with the light curves directly. This provides an independent consistency argument that the distant supernovae behave rather similar to their nearby counterparts. This means that the distant supernovae cannot be very different from the nearby ones. Yet, the colour of the distant objects appears to be systematically bluer. This could be the signature of evolution and will need to be worked out in more detail.

Luminosity distances over a limited redshift range result in degenerate likelihood distributions in the $\Omega_\Lambda vs. \Omega_M$ plane along a line corresponding roughly to $\Omega_\Lambda - 1.4\Omega_M = 0.35 \pm 0.14$ Perlmutter et al. 1999; Tonry et al. 2003 (cf. Fig. 12.2). These leads to an increased uncertainty along this direction. It should be noted that the most recent determinations of the cosmological parameters by the HZT favour values that are rather different from a flat universe solution Tonry et al. 2003; Barris 2004. If the universe indeed has a flat geometry, as suggested by the CMB data (e.g. Spergel et al. 2003) then this would be an indication of some unresolved systematic effect. The SCP has not observed a similar trend Knop et al. 2004, but the redshift range of their published data does not extend beyond $z \approx 0.8$ so far.

Figure 12.2. Likelihood distribution for Ω_Λ vs. Ω_M . The input data are from Tonry et al. 2003. This diagram should be compared to similar ones in Riess et al. 1998; Perlmutter et al. 1999; ; Leibundgut 2001; Perlmutter et al. 2003; Barris 2004; Knop et al. 2004. The degeneracy along $0.8\Omega_\Lambda - 0.6\Omega_M$ is obvious. The overlap with the flat universe model is not within the 68% likelihood area here. The grey contour lines show the dynamical age of the universe $H_0 \cdot t_0$. Clearly the SN data favour an age near 1.



5. Characterising dark energy

It has been generally accepted that a large fraction of the energy content of the universe is in a form very similar to the vacuum energy or a cosmological constant. Competing theories have been developed to explain the low, but non-zero, value of this energy form. An often used description is the equation of state parameter ($w = \frac{p}{\rho c^2}$), which in the case of dark energy has to be negative, i.e. contain negative pressure p , as the energy density ρ has to be

positive (c stands for the speed of light). With $w < -\frac{1}{3}$ the universe is actually accelerating. For field theories w is most likely variable with time and different from the value for a cosmological constant ($w = -1$). The transition from a matter dominated universe ($\Omega_M > \Omega_\Lambda$) happened sometime during the second half the history of the universe, $0.4 < z < 0.8$. It should hence be possible to determine this transition and then map the change as a function of redshift in the interval $0.2 < z < 0.8$. With a well-calibrated and controlled data set of SNe Ia in this redshift interval it should be possible to accurately map the transition and determine the strength of the dark energy and the (integrated) value of w . Several projects have embarked on such a project. The HZT has started the ESSENCE project with the search and photometry carried out with the CTIO Blanco 4m-telescope with the supporting spectroscopy from VLT, Gemini, Keck, Magellan and MMT. The goal is to have 200 spectroscopically confirmed SNe Ia with densely sampled light curves in at least two filters evenly distributed in redshift with $z < 0.8$ Smith et al. 2002. The CFHT Legacy Survey is aiming for about 900 SNe Ia out to a slightly larger redshift with spectroscopy from VLT, Keck, Gemini and Magellan. In the future the SNAP satellite, in the meantime renamed to Joint Dark Energy Mission (JDEM), should observe about 2000 SNe Ia out to $z < 1.7$.

The supernovae cannot do this alone. They will require an accurate determination of the matter density Ω_M from a different source. The required accuracy of this parameter should be a few percent (cf. Tonry et al. 2003).

In the meantime a survey for supernovae has been done within the GOODS collaboration. The goal was to find and follow supernovae at redshift larger than 1.2, which was achieved for about one third of the sample Riess et al. 2004a. Spectroscopic confirmation of 18 supernovae (nine with $z > 1$) is available Strolger et al. 2004; Riess et al. 2004b. Some of the distant objects could not be classified with a spectrum and rely on a spectroscopic redshift from the host galaxies only. These data constrain Ω_M more accurately than was possible so far, as the supernovae are in the deceleration portion of the Hubble diagram. They also show that evolutionary effects are not likely to explain the faintness of SNe Ia near $z=0.5$ and the change to more luminous objects at redshifts beyond $z=1$.

6. Conclusions

Supernovae at cosmological distances have provided some of the most accurate determinations of the Hubble constant. It is by now clear that the largest (systematic) uncertainties stem from the calibration of the local distance indicators, like the Cepheids. An intriguing result concerning the expansion history of the universe and its energy contents has emerged from the observations of very distant SNe Ia. The evidence for a cosmological constant or an energy

field acting very similarly is strong and generally accepted. The future supernova observations have to concentrate on two aspects. One is to determine the (integrated) equation of state parameter for the dark energy to possibly distinguish between a pure cosmological constant and a scalar particle field. Secondly to control any evolution of the supernovae, which potentially could affect the cosmological result, it is mandatory to constrain the models with exquisite data from nearby supernovae. These projects are under way and we can expect to make significant progress in both directions very soon.

References

- Baade, W. 1926 AN, 228, 359
 Barris, B., et al. 2004, ApJ, 602, 571
 Benetti, S., et al. 2004, MNRAS, 348, 261
 Branch, D. & Tammann, G.A., 1992, ARA&A, 30, 359
 Branch, D., 1998, ARA&A, 36, 17
 Candia, P., et al. 2003, PASP, 115, 277
 Carroll, S.M., Press, W.H. & Turner, E.L 1992, ARA&A 30, 499
 Choudhury, T. R. & Padmanabhan, T. 2004, MNRAS, in press (astro-ph/0311-622)
 Coil, A.L., et al. 2000, ApJ, 544, L111
 Contardo, G., Leibundgut, B. & Vacca, W. D., 2000, A&A, 359, 876
 Contardo, G., PhD Thesis, Technical University Munich
 Drell, P.S., Loredo, T.J. & Wasserman, I., 2000, ApJ, 530, 593
 Eastman, R. G., Schmidt, B. P. & Kirshner, R. P. 1996, ApJ, 466, 911
 Elmhamdi, A., et al. 2003, MNRAS, 338, 939
 Farrah, D, Meikle, W. P. S., Clemens, D., Rowan-Robinson, M. & Mattila, S.,
 2002, MNRAS, 336, L17
 Freedman, W.L., et al. 2001, ApJ, 553, 47
 Goldhaber, G., et al. 2001, ApJ, 558, 359
 Hamuy, M. 2001, PhD Thesis, Tucson: University of Arizona
 Hamuy, M. 2003, ApJ, 582, 905
 Hamuy, M., et al. 2002, AJ, 124, 417
 Hamuy, M. & Pinto, P. A. 2002, ApJ, 566, L63
 Hamuy, M., et al. 1996, AJ, 112, 2438
 Hillebrandt, W. & Niemeyer, J. C. 2000, ARA&A, 38, 191
 Hubble, E., The Realm of the Nebulae, 1936, (New Haven: Yale University
 Press)
 Knop, R., et al. 2004, ApJ, 598, 102
 Kowal, C. T. 1968, AJ, 73, 1021
 Krisciunas, K., et al. 2001, AJ, 122, 1616
 Krisciunas, K., et al. 2003, AJ, 125, 166
 Krisciunas, K., Phillips, M.M. & Suntzeff, N.B., 2004, ApJ, 602, L81

- Leibundgut, B. 2000, A&AR, 10, 179
 Leibundgut, B. 2001, ARA&A, 39, 67
 Leibundgut, B. & Suntzeff, N. B., in *Supernovae and Gamma-Ray Bursts*, ed. K. Weiler, (Heidelberg: Springer), 77
 Leibundgut, B. & Pinto, P.A. 1992, ApJ, 401, 49
 Leibundgut, B. & Sollerman, J., 2001, Europhysics News, 32, 4
 Leonard, D. C., et al. 2001, PASP, 114, 35
 Leonard, D. C., et al. 2003, ApJ, 594, 247
 Li, W., et al. 2001, PASP, 113, 1178
 Li, W., et al. 2003, PASP, 115, 453
 Padmanabhan, T. & Choudhury, T. R. 2003, MNRAS, 344, 823
 Peacock, J.A. 1999, *Cosmological Physics*, Cambridge: Cambridge University Press
 Peebles, P.J.E. 1993, *Principles of Physical Cosmology*, Princeton: Princeton University Press
 Perlmutter, S. & Schmidt, B. P. 2003, in *Supernovae and Gamma-Ray Bursts*, ed. K. Weiler, (Heidelberg: Springer), 195
 Perlmutter, S., et al. 1997, ApJ, 483, 565
 Perlmutter, S., et al. 1999, ApJ, 517, 565
 Phillips, M.M. 1993, ApJ, 413, L105
 Phillips, M.M., Lira, P., Suntzeff, N.B., Schommer, R.A., Hamuy, M., & Maza, J. 1999, AJ, 118, 1766
 Riess, A. G. 2000, PASP, 112, 1284
 Riess, A. G., Press, W. M. & Kirshner, R. P. 1996, ApJ, 473, 88
 Riess, A. G., et al. 1998, AJ, 116, 1009
 Riess, A. G., et al. 2004a, ApJ, 600, L163
 Riess, A. G., et al. 2004a, ApJ, in press (astro-ph/0402512)
 Saha, A, Sandage, A., Tammann, G.A., Labhardt, L., Macchett, F.D. & Parana, N. 1999, ApJ, 522, 802
 Sandage, A. 1961, ApJ, 133, 355
 Sandage, A. 1988, ARA&A, 26, 561
 Schmidt, B.P., et al. 1998, ApJ, 507, 46
 Smith, R.C., et al. 2002, BAAS, 78.08
 Spergel, D.N., et al. 2003, ApJS, 148, 175
 Strolger, L., et al. 2004, AJ, submitted
 Sullivan, M., et al. 2003, MNRAS, 340, 1057
 Tammann, A.G & Leibundgut, B., 1990, A&A, 236, 9
 Tonry, J.L., et al. 2003, ApJ, 594, 1
 Wang, Y. & Mukherjee, P. 2004, ApJ, submitted (astro-ph/0312192)
 Weinberg, S., 1972. *Gravitation and cosmology: principles and applications of the general theory of relativity*, New York: Wiley
 Williams, B. F., et al. 2003, AJ, 126, 2608

Chapter 13

GRAVITATIONAL LENSING: FROM μ -LENSING TO COSMIC SHEAR EXPERIMENTS

Francis Bernardeau

Service de Physique Théorique, CEA/DSM/SPhT, Unité de recherche associée au CNRS, CEA/Saclay 91191 Gif-sur-Yvette cédex
fbernard@sph.t.saclay.cea.fr

Abstract These notes present the physics of gravitational lensing in various cosmological contexts. The equations and approximations that are commonly used to describe the displacement field or the amplification effects are presented. Several observational applications are discussed. They range from micro-lensing effects to cosmic shear detection that is a weak lensing effect induced by the large-scale structure of the Universe. The scientific perspectives of this latter application are presented in some details.

Keywords: Cosmology; Lensing; Large-scale structure; Cosmic Shear; Dark matter; Dark energy; Perturbation Theory

1. Introduction

Gravitational lens effects offer a precious mean for probing the matter distribution in the Universe. From the search of matter in the Galactic halo to the study of the large-scale structure of the Universe, gravitational lens surveys represent a unique alternative to galaxy catalogues and have been widely used till now.

The early reports of sound detections of gravitational effects induced by the large-scale structure of the Universe acting on background galaxies date back the beginning of year 2000 Van Waerbeke et al. (2000); Kaiser et al. (2000); Bacon et al. (2000); Wittman et al. (2000). These detections have marked the beginning of a new era for the use of lens effects to probe the distribution of matter in the Universe. Unlike galaxy surveys that have been used mainly in the past two decades, cosmic shear surveys directly map the mass, not the light distribution, therefore avoiding the so-called bias problem. Cosmic shear

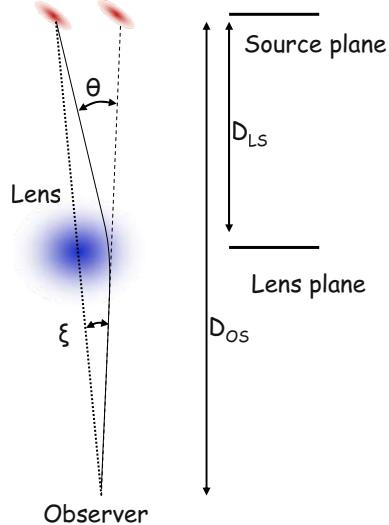


Figure 13.1. The geometrical relationship between the deflection angle θ and the displacement angle ξ .

surveys are thus expected to fruitfully complement CMB observations with projects such as CFHTLS and SNAP satellite.

The aim of these lectures is to study the effects of gravitational lenses in those different astrophysical contexts. These notes are voluntarily focused on the fundamental mechanisms and the basic concepts that are useful to describe these effects. A particular emphasis will be put on the subtleties of the scientific exploitation of the cosmic shear surveys.

These notes are organized the following way. In the first section I describe of the basic mechanisms and concepts of gravitational lensing calculations. The computations are illustrated in case of a very simple deflector, a point-like mass distribution. The next section is devoted to aspects of gravitational lenses that are more specific to a cosmological context with an application to galaxy clusters. The remaining sections are devoted to the weak lensing regime and its use in cosmology, the cosmic shear. This is a rapidly developing area that should eventually allow us to map the mass distribution in the Universe. Hints on how this can be used to constrain the cosmological models will be given in the last section.

2. Physical mechanisms

The physical mechanisms of gravitational lenses are well known since the foundation of General Relativity. Any mass concentration is going to deflect photons that are passing by with a fraction angle per unit length, $\delta\theta/\delta s$, given

by

$$\frac{\delta \vec{\theta}}{\delta s} = -2 \vec{\nabla}_x \frac{\phi}{c^2} \quad (13.1)$$

where the spatial derivative is taken in a plane that is orthogonal to the photon trajectory and ϕ is the Newtonian potential¹.

Born approximation and thin lens approximation

In practice, the total deflection angle is at most about an arcmin. This is the case for the most massive galaxy clusters. It implies that in the subsequent calculations it is possible to ignore the bending of the trajectories and calculate the lens effects as if the trajectories were straight lines. *This is the Born approximation.*

Eventually, one can do another approximation by noting that in general the deflection takes place along a very small fraction of the trajectory between the sources and the observer. One can then assume that the lens effect is instantaneous and is produced through the crossing of a plane, the lens plane. *This is the thin lens approximation.*

The induced displacement

The direct consequence of this bending is a displacement of the apparent position of the background objects. This apparent displacement depends on the distance of the source plane, D_{OS} , and on the distance between the lens plane and the source plane D_{LS} . More precisely we have (see Fig. 13.1),

$$\vec{\beta} = \vec{\alpha} - \frac{2}{c^2} \frac{D_{LS}}{D_{OS} D_{OL}} \vec{\nabla}_\alpha \left(\int ds \phi(s, \alpha) \right) \quad (13.2)$$

where $\vec{\alpha}$ is the position in the image plane, $\vec{\beta}$ is the position in the source plane. The gradient is taken here with respect to the angular position (this is why a D_{OL} factor appears). The total deflection is obtained by an integration along the line of sight, assuming the lens is thin. In a cosmological context the exact expressions of the angular distances are not trivial, they depend on the local curvature of the background.

The case of a point-like mass distribution

Multiple images and displacement field. The potential of a point-like mass distribution is given by,

$$\phi(r) = \frac{-GM}{r}, \quad (13.3)$$

¹We will see in section 13.3 what is its meaning in a cosmological context.

for an object of mass M . Let me calculate the instantaneous deflection angle at an apparent distance r . We suppose that the impact parameter of the trajectory is r and x is the abscissa to the point of the trajectory that is the closest to the lens. Along the trajectory the potential is given by,

$$\phi(x) = \frac{-G M}{\sqrt{r^2 + x^2}}. \quad (13.4)$$

Then the deflecting angle is given by,

$$\frac{\delta\theta}{\delta x} = -2 \frac{G M r}{c^2 (r^2 + x^2)^{3/2}}. \quad (13.5)$$

The total deflection angle θ is given by the result of the integration of this quantity with respect to x . It gives,

$$\theta = \frac{4 G M}{r c^2}. \quad (13.6)$$

It implies that the true position of an object on the sky, β , is related to its apparent position, α , with

$$\vec{\beta} = \vec{\alpha} - \frac{R_E^2}{\alpha^2} \vec{\alpha} \quad (13.7)$$

where $\vec{\beta}$ and $\vec{\alpha}$ are 2D angular position vectors (taken from the center of the lens) and R_E is the Einstein radius,

$$R_E = \sqrt{\frac{4 G M}{c^2} \frac{D_{LS}}{D_{OS} D_{OL}}}. \quad (13.8)$$

One can see that when $\alpha^2 = R_E^2$ the lens and the background objects are necessarily aligned. It implies that, since the optical bench is symmetric around its axis, the observed object appears as a perfect ring. It is worth noting that for this potential, except for this particular position, all background objects have two images. This is however quite specific to a point like mass distribution which has a singular gravitational potential.

The problem is that none of these features are observable when the lens is a star. Let for example assume that we have a one solar mass star in the halo of our galaxy (therefore at a distance of about 30 kpc). The apparent size of such a star is about 10^{-8} arcsec. Its Einstein ring is about 10^{-4} arcsec². None of these dimensions are accessible to the observations (the angular resolution of telescope is at best a few tens of arcsec). The Einstein radius is therefore much too small to be actually seen!

²To do this calculation it is useful to know that the horizon of a one solar mass black hole, $r = 2 G M_\odot / c^2$, is about 3 km.

Note however that these numbers show that the point-like approximation is entirely justified for a star (the Einstein ring is much more bigger than the apparent size of a star). Simple examination of the scaling in those relations shows that this would not be true for massive astrophysical objects such as galaxies or galaxy clusters.

The detection of gravitational effects due to stars should then be done by another mean: the amplification effect.

The amplification matrix. The case of circular lenses has already given us a clue: when the source is precisely aligned with the lens, the image is no more a point but a circle. One consequence is that the observed total luminosity is much larger than what would have been observed without lenses. The effect is basically due to the variations of the displacement field with respect to the apparent position. These variations induce a change of both the size and shape of the background objects. To quantify this effect one can compute the amplification matrix A which describes the linear transform between the source plane and the image plane,

$$A = \left(\frac{\partial \alpha_i}{\partial \beta_j} \right). \quad (13.9)$$

Its inverse, A^{-1} , is actually directly calculable in terms of the gravitational potential. It is given by the derivatives of the displacement with respect to the apparent position,

$$A^{-1} \equiv \frac{\partial \beta_i}{\partial \alpha_j} = \delta_{ij} - 2 \frac{D_{LS}}{D_{OS} D_{OL}} \phi_{,ij} \quad (13.10)$$

In case of a point-like mass distribution it is easy to see that (coordinates are taken from the center of the lens),

$$A^{-1} = \left(\delta_{ij} \left[1 - \frac{R_E^2}{\alpha^2} \right] + 2 \frac{\alpha_i \alpha_j}{\alpha^4} R_E^2 \right). \quad (13.11)$$

The amplification effect for each image is given by the inverse of the determinant of the amplification matrix computed at the apparent position of the image. The amplification factor is usually noted μ ,

$$\mu = 1 / \det(A^{-1}). \quad (13.12)$$

In case of the point-like distribution we have (the calculation is simple at the position $\alpha_1 = \alpha$, $\alpha_2 = 0$),

$$\mu = \left| \frac{\alpha^4}{\alpha^4 - R_E^4} \right|, \quad (13.13)$$

for each image. The total amplification effect is given by the summation of the two effects for the 2 images,

$$\mu_{\text{tot}} = \frac{u^2 + 2}{u(u^2 + 4)^{1/2}} \quad \text{with} \quad u = \frac{\beta}{R_E}, \quad (13.14)$$

where β appears to be the impact parameter of the background object in the source plane. The amplification effect is obviously dependent on the impact parameter. If it is changing with time, this effect is detectable. This is the basis of the microlensing experiments : when a compact object of the halo of our galaxy reaches, because of its proper motion, the vicinity of the light path of a background star (from the SMC or the LMC) the impact parameter is changing with time and can be small enough to induce a detectable amplification (when u is about unity, the amplification is about 30%). In practice one observes changes in the magnitude of the remote stars that obey specific properties,

- the time dependence of the amplification is symmetric and has a specific shape;
- the amplification effect is unique;
- the magnitude of the amplification effect is the same in all wavelengths.

The time scale of such an event is about a few days to a few month depending on the mass of the deflectors. Currently a fair number of such events have been recorded and constraints on the content of our halo with low massive compact objects have been put.

3. Gravitational lenses in Cosmology

The extension of the lens equations to a cosmological context raises some technical difficulties because the background in which the objects are embedded is not flat. In particular one must pay attention to the distances that should be employed. The other change is that lenses will now have extended mass distribution.

Cosmological distances and gravitational potential

In cosmology one distinguishes between the radial distance, χ (distance actually travelled on a light cone) and the angular distance D_0 (defined from the apparent size of distant objects). Reminding that

$$ds^2 = dt^2 - a^2(t) \left(\frac{dx^2}{1 - kx^2} + x^2 d\theta^2 + x^2 \sin^2 \theta d\varphi^2 \right) \quad (13.15)$$

The distance χ is directly related to the time dependence of the expansion factor. It is given by,

$$\chi(z) = \int_{t_0}^{t_1} \frac{c dt}{a} = \int_0^{x(z)} \frac{dx}{\sqrt{1 - kx^2}} \quad (13.16)$$

for the distance between $z = 0$ and a plane at redshift z . The distance between two planes at redshifts z_1 and z_2 is simply,

$$\chi(z_2, z_1) \equiv \int_{t_1}^{t_2} \frac{c dt}{a} = \chi(z_2) - \chi(z_1) \quad (13.17)$$

The angular distances can be obtained from the radial ones and are given by,

$$D_0(z) = \frac{1}{\sqrt{-k}} \sinh[\sqrt{-k} \chi(z)] \quad (13.18)$$

and

$$D_0(z_1, z_2) = \frac{1}{\sqrt{-k}} \sinh[\sqrt{-k} (\chi(z_2, z_1))] \quad (13.19)$$

where k is the (constant) curvature of the metric (Note that the two distances coincide when $k \rightarrow 0$). The exact expressions of these distances as a function on redshifts depend on the time dependence of a and thus on the content of the Universe (and particularly on the vacuum energy density). In case of an Einstein-de Sitter Universe we simply have,

$$D_0(z) = \chi(z) = \frac{c}{H_0} \left(2 - \frac{2}{\sqrt{1+z}} \right) \quad (13.20)$$

In the lens equation the distance to use is the angular distance. This can be unambiguously established by the geometric optic equations.

And the source term of the deflection is the gravitational potential, ϕ , which in a cosmological context is given by a slightly modified Poisson equation,

$$\Delta_x \phi(\mathbf{x}) = 4\pi G \bar{\rho} a^2 \delta_{\text{mass}}(\mathbf{x}) \quad (13.21)$$

where δ_{mass} is the local mass density contrast,

$$\delta_{\text{mass}}(\mathbf{x}) = \rho(\mathbf{x})/\bar{\rho} - 1 \quad (13.22)$$

Galaxy clusters as gravitational lenses

The study of galaxy clusters has become a very active field since the discovery of the first gravitational arc in Abell cluster A370 Soucail et al. (1998). Galaxy clusters give the most dramatic example of gravitational lens effects in a cosmological context. The difficulty is however to describe the shape of their mass distribution.

The isothermal profile

For an isothermal profile we assume that the local density $\rho(r)$ behaves like,

$$\rho(r) = \rho_0 \left(\frac{r}{r_0} \right)^{-2} \quad (13.23)$$

With such a density profile the total mass is not finite. So this is not a realistic description but it is a good starting point for the central part of clusters. It is actually more convenient to parameterize the depth of a potential well with the velocity dispersion it induces. Such a dispersion is in principle measurable with the observed galaxy velocity dispersion. It is related to the mass $M(< r)$ included within a radius r ,

$$\sigma^2(r) \sim \frac{G M(< r)}{r} \quad (13.24)$$

In case of a isothermal profile, the velocity dispersion is *independent* of the radius and we have

$$\sigma^2 = 2\pi G \rho_0 r_0^2 \quad (13.25)$$

The integrated potential along the line-of-sight is given by,

$$\varphi(r) = 2\pi \sigma^2 r \quad (13.26)$$

As a consequence, the amplitude of the displacement is independent of the distance to the cluster center and

$$\vec{\beta} = \vec{\alpha} - \frac{4\pi}{c^2} \frac{D_{LS}}{D_{OS}} \sigma^2 \frac{\vec{\alpha}}{\alpha} \quad (13.27)$$

For

$$\alpha = R_E = \frac{4\pi}{c^2} \frac{D_{LS}}{D_{OS}} \sigma^2 \quad (13.28)$$

$\alpha^S = 0$ is solution so that R_E corresponds to the radius of a possible ‘‘Einstein ring’’. Note that its size is directly proportional to the square of velocity dispersion (in units of c^2) and to the ratio D_{LS}/D_{OS} . Generically the number of images depends on the value of the impact parameter. If it is smaller than R_E , there are two solutions, otherwise only one. For a galaxy cluster of a typical velocity dispersion of 500 km/s, and for a source plane situated at twice the distance of the lens, the size of the Einstein ring is about 0.5 arcmin that well within observational limits.

Note that the amplification matrix reads,

$$A^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 - \frac{1}{x} \end{pmatrix} \quad (13.29)$$

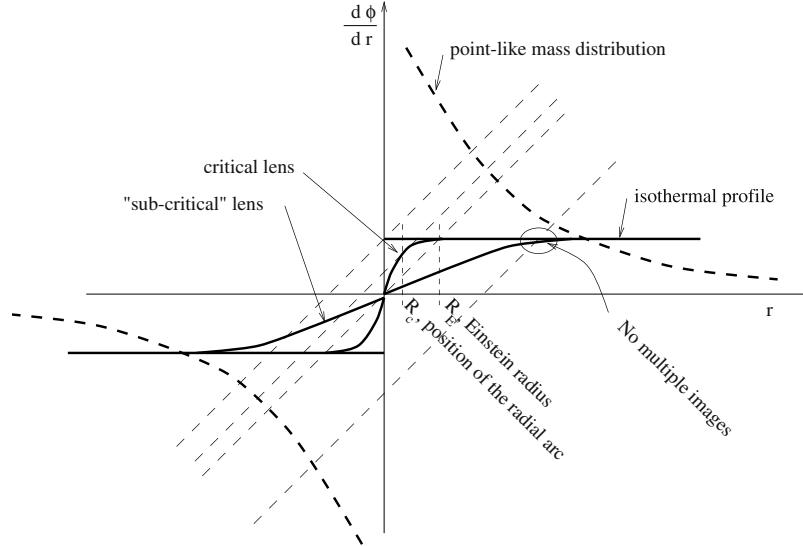


Figure 13.2. Graphical determination of the position and number of images from the shape of the potential.

where $x = r/R_E$. As a result the amplification factor is given by, $\mu = x/(1-x)$. It becomes infinite when $x \rightarrow 1$. It expresses the fact that at the position of the Einstein ring the background objects are infinitely stretched out.

The case of a spherically symmetric mass distribution. Let us consider a more general spherically symmetric profile for the mass distribution. The displacement is then given by,

$$\vec{\beta} = \vec{\alpha}^I - \frac{d\psi}{d\alpha} \frac{\vec{\alpha}}{\alpha}, \quad (13.30)$$

where ψ is proportional to the line-of-sight projected potential. It is interesting to visualize this relation with a graphic representation as shown on Fig. 13.2. The number and position of the images of a given background object are given by the number of intersection points between the curve and a straight line of slope unity. This is a direct consequence of the relation,

$$b - a = - \frac{d\varphi(a)}{da} \quad (13.31)$$

when the potential is computed *along a given axis* that crosses the cluster through the center and b and a are the abscissa on this axis of one given object in respectively the source and the image plane.

In this case the amplification matrix reads,

$$A^{-1} = \begin{pmatrix} 1 - \frac{\partial^2 \varphi}{\partial r^2} & 0 \\ 0 & 1 - \frac{1}{r} \frac{\partial \varphi}{\partial r} \end{pmatrix} \quad (13.32)$$

when it is written in the basis $(\vec{e}_r, \vec{e}_\theta)$. Then the amplification is infinite in two cases, when

$$\frac{\partial^2 \varphi}{\partial r^2} = 1 \text{ or } \frac{1}{r} \frac{\partial \varphi}{\partial r} = 1 \quad (13.33)$$

The second eigenvalue corresponds to an Einstein ring. The first eigenvalue, however, is associated with an eigenvector that is along the x direction, that is along the radial direction. It can lead to the formation of “radial arc”. It graphically corresponds to the case of two merging roots associated to positions close to the origin. That such arcs have been observed gives precious indication on the behavior of the potential near the central part of clusters.

Critical lines and caustics in realistic mass distributions. In realistic reconstructions of lens potential however, it is very rare that the lens has circular shape. Most of the time the mass distribution of the lens is much more complicated. It induces complex features and series of multiple images.

The simplest assumption beyond the spherically symmetric models is to introduce an ellipticity ϵ in the mass distribution,

$$\varphi = \varphi_0 \sqrt{1 + r_{\text{em}}^2/r_c^2} \text{ with } r_{\text{em}}^2 = \frac{x^2}{(1-\epsilon)^2} + \frac{y^2}{(1+\epsilon)^2} \quad (13.34)$$

To understand the physics it induces one should introduce the caustics and critical lines. The *critical lines* are the location on the image plane of the points of infinite magnification. The *caustics* are the location of these points on the source plane. These points are determined by the lines on which $\det(A^{-1}) = 0$. For instance the Einstein ring is the critical line for an isothermal profile. Its caustic reduces to one point that can be covered by a background object thus forming a genuine Einstein ring. In general critical lines will be associated with the formation of arcs produced by background galaxies that happen to be located on the caustics.

In practice the reconstructions of the basic phenomenological effects, position of giant arcs and of multiple images, seen in images of galaxy clusters can be obtained through the superposition of one or very few such elliptical potentials. However the reconstruction of complete galaxy cluster mass maps may eventually require the use of more complicated models and it can be necessary to perform non-parametric mass reconstructions. A number of important results have been obtained from such observations – see Mellier (1999) and references therein.

The weak lensing regime

In this section the possibility of using lens effects to probe the large-scale structure of the Universe is considered. The difficulty lies in that the distortion induced by the lenses can be very small. The projected potential should then be reconstructed with a statistical analysis on the deformation effects measured on a lot of background objects. For slightly extended objects such as background galaxies the deformation is described by the amplification matrix, the components of component of which are usually written

$$A^{-1} = \begin{pmatrix} 1 - \kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix} \quad (13.35)$$

taking advantage of the fact that it is a symmetric matrix. The components of this matrix are expressed in terms of the convergence, κ , (a scalar field) and the shear, γ (a pseudo vector field) with

$$\kappa = \frac{1}{2} \nabla^2 \psi; \quad \gamma_1 = \frac{1}{2} (\psi_{,11} - \psi_{,22}); \quad \gamma_2 = \psi_{,12} \quad (13.36)$$

with

$$\psi = 2 \frac{D_{LS}}{D_{OS} D_{OL}} \varphi \quad (13.37)$$

The convergence describes the linear change of size and the shear describes the deformation. *The weak lensing regime corresponds to small values of κ and γ .*

The consequences of such a transform matrix can be decomposed in two aspects:

- The magnification effect Broadhurst (1995); Broadhurst et al. (1995). Lenses induce a change of size of the objects. As the surface brightness is not changed by this effect, the change of surface induces a direct magnification effect that causes changes in the apparent number density of detected objects. This effect will be ignored in the following.
- The distortion effect. Lenses also induce a change of shape of the background objects. The eigenvalues of the matrix A^{-1} determine the direction and amplitude of such a deformation.

Shapes of background galaxies can then be used to map the distortion field. This is the first step towards the reconstruction of the actual projected mass density of the Universe. The elaboration of methods for reconstructing mass maps from distortion fields is itself not a trivial issue. Following the pioneering work of Kaiser and Squires Kaiser & Squires (1993) many methods have been proposed to address this point Kaiser (1995); Kaiser (1995); Bartelmann et al. (1996); Seitz & Schneider (1995); Seitz & Schneider (1996); Seitz & Schneider (1998) and shown it can be done now with great accuracy.

4. Cosmic Shear: weak lensing as a probe of the large-scale structure

In this section the possibility of using weak lensing analysis for mapping the large-scale structure of the universe is investigated.

Single lens and multiple lenses

At the end of the previous section, it has been stressed that the convergence and shear components all derived from a single scalar field directly proportional to the projected potential. To be more precise if the elements of the amplification matrix are dominated by one single lens then

$$A_{ij}^{-1} = \delta_{ij} - \frac{2}{c^2} \frac{D_{LS} D_{OL}}{D_{OS}} \int_0^{\chi_s} d\chi' \phi_{,ij}(\chi') \quad (13.38)$$

where $\phi_{,ij}(\chi')$ are the derivatives of the gravitational potential ϕ in comoving coordinates at distance χ' to the observer. The line of sight integral is done to the source plane distance χ_s ³. The expression of the convergence is then particularly interesting. It is obtained from the trace of the expression which makes appear the Laplacian of the potential field, which in turn is proportional to the local density contrast (e.g. Eq. 13.21) so that,

$$\kappa = \frac{3}{2} \Omega_0 \frac{1}{(c/H_0)^2} \frac{D_{LS} D_{OL}}{D_{OS}} \int_0^{\chi_s} \frac{d\chi'}{a(\chi)} \delta(\chi') \quad (13.39)$$

Note that the cosmological distances are proportional to c/H_0 so that the κ - δ relation does not depend on the value of H_0 .

Before further exploration of the phenomenological consequences of this relation, it should be extended to cases where multiple lenses are along the line of sight; a generic situation in a cosmological context.

As on Fig. 13.3 when 2 lenses contribute to the deflection the total displacement field, ξ , is obtained from the sum of the 2 contributions $\xi^{(1)}$ and $\xi^{(2)}$,

$$\xi_i^{\text{tot}} = \xi_i^{(1)} + \xi_i^{(2)} \quad (13.40)$$

$$= -\frac{2}{c^2} \frac{D_{1S}}{D_{OS} D_{O1}} \varphi_{,i}^{(1)} - \frac{2}{c^2} \frac{D_{2S}}{D_{OS} D_{O2}} \varphi_{,i}^{(2)} \quad (13.41)$$

where $\varphi_i^{(n)}$ is the gradient of the integrated potential of each lens (assuming each can be approximated by a thin lens). Note however that while the gradient

³Note that the source plane is not necessarily thin. The result is then obtained by a convolution of this expression by the source plane profile. It does not change the discussions presented here so that it will be ignored in the following.

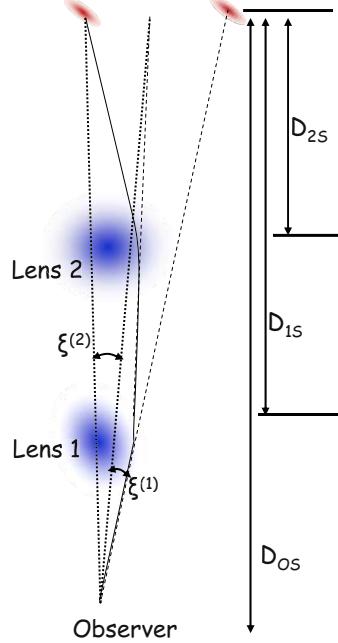


Figure 13.3. The case of a double deflection.

of lens 1 is taken at the image plane position, α^I , the one of lens 2 is taken on the lens 2 plane deformed by the first lens which then read:

$$\alpha_i^{L2} = \alpha_i^I - \frac{2}{c^2} \frac{D_{12}}{D_{O2} D_{O1}} \varphi_{,i}^{(1)} \quad (13.42)$$

As a result the total amplification matrix reads,

$$\begin{aligned} (A_{\text{tot.}}^{-1})_{ij} &= \delta_{ij} - \frac{2}{c^2} \frac{D_{1S}}{D_{OS} D_{O1}} \varphi_{,ij}^{(1)} \\ &\quad - \frac{2}{c^2} \frac{D_{2S}}{D_{OS} D_{O2}} \varphi_{,ik}^{(2)} \left(\delta_{kj} - \frac{2}{c^2} \frac{D_{12}}{D_{O2} D_{O1}} \varphi_{,kj}^{(1)} \right) \end{aligned} \quad (13.43)$$

which implies that,

$$A_{\text{tot.}}^{-1} - \text{Id} = A_1^{-1} - \text{Id} + A_2^{-1} - \text{Id} + \frac{D_{12}}{D_{O2}} \frac{D_{OS}}{D_{1S}} (A_2^{-1} - \text{Id}) (A_1^{-1} - \text{Id}) \quad (13.44)$$

The nontrivial part of the combination effects appear to be quadratic in the convergence and the shear. In the weak lensing regime, by construction we keep only linear terms which implies that

$$A_{\text{tot.}}^{-1} - \text{Id} = \sum_{\text{lenses}} (A_i^{-1} - \text{Id}) . \quad (13.45)$$

The amplification matrix is given by the superposition of lens effects of the different mass layers. Generalizing this result to a continuous field⁴, the component of the amplification matrix can then be inferred from the line of sight integration of the gravitational potential angular derivatives,

$$A_{ij}^{-1} = \delta_{ij} - \frac{2}{c^2} \int_0^{\chi_s} d\chi' \phi_{,ij} \frac{D(\chi', \chi_s)}{D(\chi_s) D(\chi')} \quad (13.46)$$

In particular the projected convergence reads,

$$\kappa = \frac{3}{2} \Omega_0 \frac{1}{(c/H_0)^2} \int_0^{\chi_s} \frac{d\chi'}{a(\chi')} \frac{D(\chi', \chi_s) D(\chi')}{D(\chi_s)} \delta_{\text{mass}}(\chi') \quad (13.47)$$

Basics of phenomenology

The previous relation basically tells us that

$$\kappa = \int d\chi w(\chi) \delta_{\text{mass}} \quad (13.48)$$

where $w(\chi)$ is a line-of-sight weight function proportional to Ω_0 .

From the previous equation it can be readily noted that the amplitude of the convergence field is expected to be both proportional to Ω_0 and to the amplitude of the 3D mass density fluctuation. Extra dependence with the other cosmological parameters (Ω_0, Ω_Λ) can take place through the growth rate of the fluctuation and the cosmological distances if the source plane is at large enough redshift (comparable to unity).

To a very crude approximation we then have

$$\sigma_\kappa \sim \Omega_0 \sigma_8 \quad (13.49)$$

where σ_8 is the amplitude of the matter density fluctuations at $8h^{-1}\text{Mpc}$ scale.

In the following we will try to examine in more details the κ - δ relation.

Convergence and shear correlation functions and power spectra

It can be easily shown that

$$\partial_1 \kappa = \partial_1 \gamma_1 + \partial_2 \gamma_2 \quad (13.50)$$

$$\partial_2 \kappa = -\partial_2 \gamma_1 + \partial_1 \gamma_2 \quad (13.51)$$

and which also gives,

$$\Delta \kappa = (\partial_1^2 - \partial_2^2) \gamma_1 + 2\partial_1 \partial_2 \gamma_2 \quad (13.52)$$

⁴A rigorous derivation of these formulae can be found in Sachs, Sachs (1961), where the equations of the geometric optics applied to a light bundle are derived in the context of a weakly perturbed FRW metric.

These real space relations can be transformed into Fourier space relations. The Fourier modes of the convergence field is defined as

$$\kappa(\mathbf{x}) = \int \frac{d^2\mathbf{l}}{2\pi} \kappa(\mathbf{l}) \exp(i\mathbf{l}\cdot\mathbf{x}) \quad (13.53)$$

where the survey is assumed to be small enough so that the decomposition can be made into plane waves⁵. The above relations imply that

$$\gamma_1(\mathbf{l}) = \frac{l_1^2 - l_2^2}{l^2} \kappa(\mathbf{l}) \quad (13.54)$$

$$\gamma_2(\mathbf{l}) = \frac{2l_1 l_2}{l^2} \kappa(\mathbf{l}). \quad (13.55)$$

According to our current cosmological principles, any angular survey is expected to be statistically isotropic so that ensemble averages of product of 2 Fourier modes $\kappa(\mathbf{l})$ are expected to behave like,

$$\langle \kappa(\mathbf{l}) \kappa(\mathbf{l}') \rangle = \delta_{\text{Dirac}}(\mathbf{l} + \mathbf{l}') P_\kappa(l) \quad (13.56)$$

and then the 2-point correlation function of the field κ is

$$\langle \kappa(\mathbf{x}) \kappa(\mathbf{x}') \rangle = \int \frac{d^2\mathbf{l}}{(2\pi)^2} P_\kappa(l) \exp[i\mathbf{l}\cdot(\mathbf{x} - \mathbf{x}')] \quad (13.57)$$

And from (13.54) it is not too difficult to show that,

$$\langle \kappa(\mathbf{x}) \kappa(\mathbf{x}') \rangle = \langle \gamma_1(\mathbf{x}) \gamma_1(\mathbf{x}') \rangle + \langle \gamma_2(\mathbf{x}) \gamma_2(\mathbf{x}') \rangle \quad (13.58)$$

$$\langle \gamma_1(\mathbf{x}) \gamma_2(\mathbf{x}') \rangle = 0 \quad (13.59)$$

The convergence power spectrum in the small angle approximation

The aim of this paragraph is to explicit the relation between the convergence power spectrum and the 3D matter power spectrum. To do that we will explicit the computation of the 2-point correlation of the convergence power spectrum. The 2-point κ correlation function reads,

$$\langle \kappa(\mathbf{x}) \kappa(\mathbf{x}') \rangle = \int d\chi d\chi' w(\chi) w(\chi') \langle \delta(\chi, D\mathbf{x}) \delta(\chi', D'\mathbf{x}') \rangle \quad (13.60)$$

which makes intervene the 2-point matter correlation function between points at radial distance respectively χ and χ' and at angular distance $D\mathbf{x}$ and $D'\mathbf{x}'$.

⁵If not, the convergence field has to be decomposed into spherical harmonics and the shear components into mathematical extensions of those.

This expression can be written in terms of the 3D matter density power spectrum P_δ ,

$$\begin{aligned} \langle \kappa(\mathbf{x})\kappa(\mathbf{x}') \rangle &= \int d\chi d\chi' w(\chi) w(\chi') \\ &\int \frac{d^3\mathbf{k}}{(2\pi)^3} P_\delta(k) \exp[i k_r(\chi - \chi') + i \mathbf{k}_\perp \cdot (\mathbf{D}\mathbf{x} - \mathbf{D}'\mathbf{x}')] \end{aligned} \quad (13.61)$$

where the wave vector \mathbf{k} has been decomposed into its radial component k_r and its tangential components \mathbf{k}_\perp . Unless the very long wave fluctuations dominate the matter power spectrum, the typical values for k_r and k_\perp that will dominate this integral are about respectively 1/Depth and 1/(| $\mathbf{x} - \mathbf{x}'$ |Depth) where Depth is the depth of the survey. In the small angle approximation it means that $k_r \ll k_\perp$. The integral over k_r can be done explicitly and it leads to $\delta_{\text{Dirac}}(\chi - \chi')$. As a result, we have

$$\langle \kappa(\mathbf{x})\kappa(\mathbf{x}') \rangle = \int d\chi w^2(\chi) \int \frac{d^2\mathbf{k}}{(2\pi)^2} P_\delta(k) \exp[i D\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}')] \quad (13.62)$$

which can be rewritten as:

$$\langle \kappa(\mathbf{x})\kappa(\mathbf{x}') \rangle = \int \frac{d^2\mathbf{k}}{(2\pi)^2} \exp[i l \cdot (\mathbf{x} - \mathbf{x}')] \int d\chi \frac{w^2(\chi)}{D^2} P_\delta\left(\frac{l}{D}\right) \quad (13.63)$$

By identification we obtain the expression of the convergence power spectrum

$$P_\kappa(l) = \int d\chi \frac{w^2(\chi)}{D^2} P_\delta\left(\frac{l}{D}\right) \quad (13.64)$$

It is therefore possible to relate the amplitude of the convergence field to the amplitude and shape of the matter fluctuations. The result can be expressed the r.m.s. of the convergence fluctuation in a disc of radius θ_0 ,

$$\sigma_\kappa \approx 0.01 \left(\frac{\theta_0}{1 \text{ deg}} \right)^{-(n+2)/2} \sigma_8 \Omega_0^{0.8} z_s^{0.75} \quad (13.65)$$

where σ_8 is the rms of the matter density fluctuations at $8 h^{-1} \text{Mpc}$ scale and n is the power spectrum index.

The question now is whether it is observable or not. This signal has to be above the systematics (that depends on what you use to do the measurements) and above the (white) noise due to the intrinsic shape fluctuations of the source galaxies. In practice the measurement is made in the following way. Galaxy ellipticity, 25000ϵ , are measured and a "super" pixel (say of the order of a few arcmin size) and averaged out. In the weak lensing regime we have,

$$\epsilon^I = \epsilon^S + \gamma \quad (13.66)$$

Gravitational len.

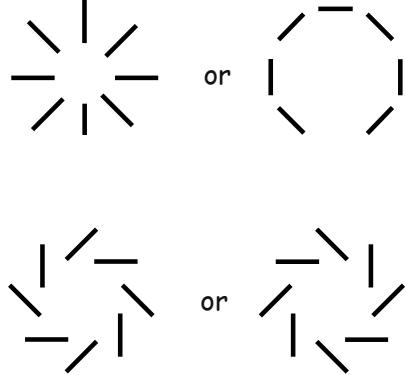


Figure 13.4. Shear configuration for pure E modes (top) and pure B modes (bottom).

so that the average ellipticity in the image plane reduces to the shear value if, as expected, the average ellipticity is zero in the source plane.

One important practical question is then the level of systematics. It is possible to do a diagnosis of it by taking advantage of the geometrical properties of the shear field, e.g. the fact that it is a potential field.

The E and B fields

As was stressed before, what is observed is the shear field. In practice it is likely that the measured shear field is not purely of cosmological origin due to contamination or systematics effects. The geometrical properties of the cosmological shear field can however be used to derive means for testing the amount of systematics in the data. The idea is the following. Any 2D spin 2 vector field can be decomposed into a scalar and a pseudo scalar parts. They are defined through their Laplacian,

$$\Delta E = (\partial_1^2 - \partial_2^2)\gamma_1 + 2\partial_1\partial_2\gamma_2 \quad (13.67)$$

$$\Delta B = -2\partial_1\partial_2\gamma_1 + (\partial_1^2 - \partial_2^2)\gamma_2 \quad (13.68)$$

In case the components γ are truly those of a cosmological shear field there would be no B components and E identifies with κ . E is a scalar field: it is invariant under parity change. B is a pseudo scalar field: it changes sign under parity change. These properties can be appreciated on Fig. 13.4 where shear configuration for respectively E and B modes are presented. The E -mode shear field configurations are unaffected by parity changes; the ones for the B modes are changed (in one-another).

The decomposition in E and B components is however not trivial because the relation between E and B and the shear field is non local. It is however possible to have a quasi-localized relation if one considers those field when

they are filtered with a compensated filter - that is a filter with 0 integral that therefore smoothes out the long wavelength modes⁶. Let me denote \hat{E} and \hat{B} such quantities in the following. Then formally \hat{E} (at the origin) can be written as

$$\hat{E} = \int d^2\mathbf{x}' E(\mathbf{x}') W_c(\mathbf{x}') \quad (13.69)$$

where W_c is a compensated filter. Because it is a compensated filter its Fourier transform can be written $k^2 \tilde{W}(k)$ so that

$$\hat{E} = \int d^2\mathbf{x}' \Delta E(\mathbf{x}') \tilde{W}(x') \quad (13.70)$$

where $\tilde{W}(x')$ is the real-space Fourier transform of $\tilde{W}(k)$. Then

$$\hat{E} = \int d^2\mathbf{x}' [(\partial_1^2 - \partial_2^2)\gamma_1(\mathbf{x}') + 2\partial_1\partial_2\gamma_2(\mathbf{x}')] \tilde{W}(x') \quad (13.71)$$

which after integration by parts can be written

$$\hat{E} = \int d^2\mathbf{x}' \begin{pmatrix} \cos(2\theta') \\ \sin(2\theta') \end{pmatrix} \cdot \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \left(\tilde{W}''(x') - \frac{\tilde{W}'(x')}{x'} \right). \quad (13.72)$$

where θ' is the angle of \mathbf{x}' . What we obtain here is an integral of the radial component of γ on an area with a given profile $p(x')$, e.g.

$$\hat{E} = \int dr p(r) \int d\theta \gamma_r(\theta). \quad (13.73)$$

Similarly \hat{B} corresponds to an integral of tangential component of the shear,

$$\hat{B} = \int dr p(r) \int d\theta \gamma_T(\theta). \quad (13.74)$$

Consistency relations for the shear correlation functions.

The observational data allow the estimation of two different correlation functions at finite distance. The one between the radial components, $\xi_{++}(x)$ and the one between the transverse component $\xi_{\times\times}(x)$ (the cross-correlation between the components being 0 for symmetry reasons).

However in the absence of B field, these two functions should be related together, e.g. they all come from a single power spectrum. Although it is possible to obtain a functional relation from a direct approach, I present here

⁶When filtered in such a way the convergence field is called the *mass aperture*.

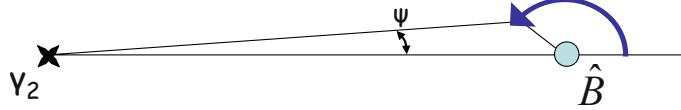


Figure 13.5. Sketch of the integration path for Eq. 13.74

a way to get such a relation from simple physical considerations. Indeed a vanishing B implies that both $\langle \gamma_1(\mathbf{x}_1) \hat{B}(\mathbf{x}_2) \rangle$ and $\langle \gamma_2(\mathbf{x}_1) \hat{B}(\mathbf{x}_2) \rangle$ are zero. This is true for parity reasons for the first one if $\mathbf{x}_2 - \mathbf{x}_1$ is on the horizontal axis. Imposing that the second one vanishes will lead to a non trivial relation between ξ_{++} and $\xi_{\times\times}$.

From the expression of \hat{B} we have

$$\begin{aligned} \langle \gamma_2(\mathbf{x}_1) \hat{B}(\mathbf{x}_2) \rangle &= \int dr p(r) \int_0^{2\pi} d\theta (-\sin(2\theta) \langle \gamma_2(\mathbf{x}_1) \gamma_1(\mathbf{x}') \rangle + \\ &\quad \cos(2\theta) \langle \gamma_2(\mathbf{x}_1) \gamma_2(\mathbf{x}') \rangle) \end{aligned} \quad (13.75)$$

where \mathbf{x}' is $\mathbf{x}' = \mathbf{x}_2 + r(\cos\theta, \sin\theta)$. The correlations $\langle \gamma_2(\mathbf{x}_1) \gamma_1(\mathbf{x}') \rangle$ and $\langle \gamma_2(\mathbf{x}_1) \gamma_2(\mathbf{x}') \rangle$ can be expressed in terms of ξ_{++} and $\xi_{\times\times}$,

$$\langle \gamma_2(\mathbf{x}_1) \gamma_1(\mathbf{x}') \rangle = \sin(2\psi) \cos(2\psi) [\xi_{++}(x') - \xi_{\times\times}(x')] \quad (13.76)$$

$$\langle \gamma_2(\mathbf{x}_1) \gamma_2(\mathbf{x}') \rangle = \sin^2(2\pi) \xi_{++}(x') + \cos^2(2\psi) \xi_{\times\times}(x') \quad (13.77)$$

where ψ is defined on Fig. 13.5 and $x' = |\mathbf{x}' - \mathbf{x}_1|$. The above integral should vanish for any function $p(r)$, therefore in particular when the contributing values of r are much smaller than $|\mathbf{x}_2 - \mathbf{x}_1|$. It is then possible to expand all quantities with respect to r (up to second order):

$$\begin{aligned} x' &= x + r \cos\theta + \frac{r^2}{2x^2} \sin^2\theta + \dots \\ \xi(x') &= \xi(x) + r \cos\theta \xi'(x) + \frac{r^2}{2x} \sin^2\theta \xi'(x) + \frac{r^2}{2} \cos^2\theta \xi''(x) + \dots \\ \sin\psi &= \frac{r}{x} \sin\theta - \frac{r^2}{x^2} \sin\theta \cos\theta + \dots \end{aligned}$$

After integration over θ we have,

$$\xi''_{\times\times} + 3 \frac{\xi'_{\times\times}}{x} + 4 \frac{\xi_{\times\times}}{x^2} - 4 \frac{\xi'_{++}}{x} - 4 \frac{\xi_{++}}{x^2} = 0 \quad (13.78)$$

$(\int_0^{2\pi} d\theta \cos(2\theta) \sin^2\theta = -(2\pi)/4$, $\int_0^{2\pi} d\theta \cos(2\theta) \cos^2\theta = (2\pi)/4$). This expression can be written in an integral form. It allows to make important consistency checks in the data samples.

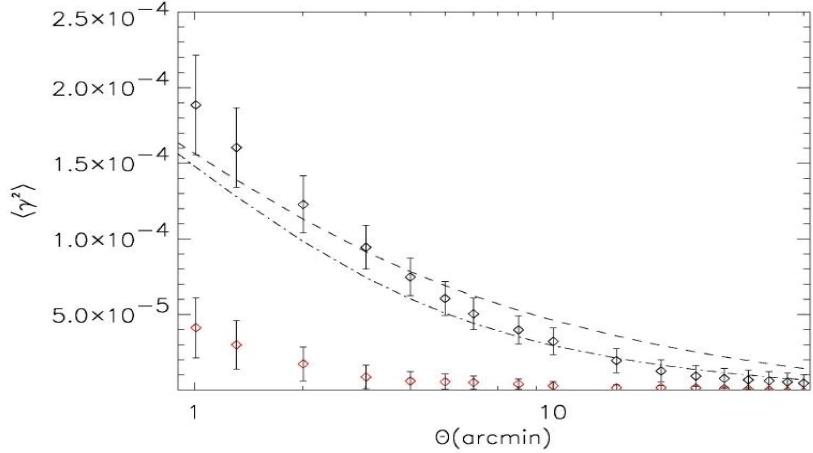


Figure 13.6. Measured behavior of the E (upper black diamonds) and B (lower red diamonds) mode components in the DESCART cosmic shear surveys (van Waerbeke, private communication).

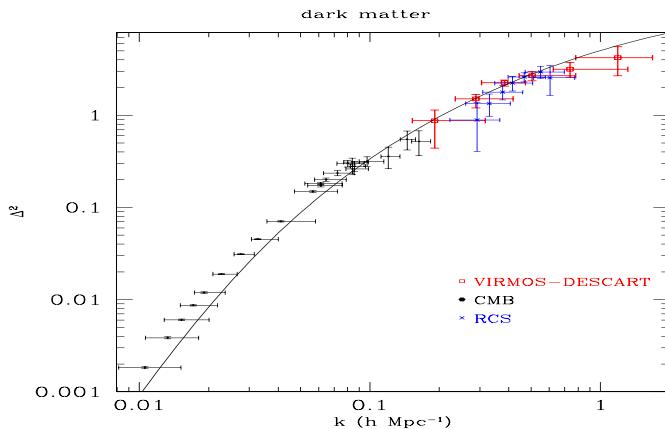


Figure 13.7. Shape of the dark matter power spectrum as reconstructed from weak lensing data (red squares, Virmos-Descart survey). It is confronted to what is expected from CMB anisotropy power spectrum (small black squares).

On Fig. 13.6 is presented what can be actually observed in data sets. B modes are shown to be detectable at very small angular scale. This could be due to the fact that close pairs may actually have gone through a physical interaction. At large scale however the B modes are found to be negligible, opening the way to a far reaching use of such catalogues.

The resulting dark matter power spectrum is shown on Fig. 13.7 where results from weak lensing surveys are checked against CMB observations.

5. Conclusions and perspectives: cosmic shear in a precision cosmology era

Because cosmic shear surveys probe the dark matter distribution up to significant redshift, it is a probe of the details of the gravitational dynamics. With large scale surveys it then becomes possible to test the details of large-scale structure growth with unprecedented accuracy.

It is possible for instance to test the mode couplings as the gravitational instabilities develop. The amount of non-Gaussianities embedded in cosmic shear surveys then betrays the details of the gravitational instability scenario, allowing for instance the possibility of measuring the mass density of the universe in a totally independent way. That would contribute in the consolidation of the current concordance model of cosmology.

Actually the crucial issue raised by this model is the dark energy, a putative energy density associated with the cosmological vacuum, whose existence is strongly suggested from the current cosmological observations. This is a major issue from a theoretical physics point of view. At this moment it is not clear that its solution lies in a change of the gravity law at very large scale (as could the existence of extra spatial dimension imply) or to the existence of a new form of matter in the universe. In all cases it is to be stressed that it is essentially a low redshift phenomenon that cosmic shear observations should allow to address. In such surveys precise quantitative tests can be designed where the very details of the model of the structure growth can be scrutinized, Bernardeau (2003).

References

- Bacon, D., Refregier, A. & Ellis, R., 2000, MNRAS318, 625
- Bartelmann, M., Narayan, R., Seitz, S. & Schneider, P., 1996, ApJ464, 115
- Benabed, K. & Bernardeau, F. 2001, Phys. Rev. D 64, 083501
- Bernardeau, F. 1998, in “Theoretical and Observational Cosmology”, proc. of Cargèse summer school, August 17-29 1998, ed. M. Lachièze-Rey
- Bernardeau, F., Colombi, S., Gaztañaga, E., Scoccimaro, R. 2002, Phys. Rep. 367, 1
- Bernardeau, F. 2003, Rep. Prog. Phys. 66, 691
- Bernardeau, F., van Waerbeke, L. & Mellier, Y. 1997, A&A 324, 15
- Bertschinger, E. 1996, in “Cosmology and Large Scale Structure”, Les Houches Session LX, August 1993, NATO series, eds. R. Schaeffer, J. Silk, M. Spiro, J. Zinn-Justin, Elsevier Science Press
- Blandford, R. D., Saust, A. B., Brainerd, T. G. & Villumsen, J. V. 1991, MNRAS251, 600

- Brax, Ph. & Martin, J. 2000, Phys. Rev. D 61, 103502
 Brax, Ph. & Martin, J. 1999, Phys. Letter B 468, 40
 Broadhurst, T. astro-ph/9511150.
 Broadhurst, T., Taylor, A.N. & Peacock, J. 1995, ApJ438, 49
 CFHLS web site, <http://www.cfht.hawaii.edu/Science/CFHLS/>
 Hamilton, A. J .S., Kumar, P., Lu, E. & Matthews, A. 1991, ApJL, 374, L1
 Hui, L. 1999, ApJL, 519, L9
 Jain, B. & Seljak, U. 1997, ApJ484, 560
 Kaiser, N., Squires, G. & Broadhurst, T. 1995, ApJ449, 460
 Kaiser, N. 1992, ApJ388, L72
 Kaiser, N. 1995, ApJ439, 1
 Kaiser, N., Wilson, G., Luppino, G. A., 2000, astro-ph/0003338
 Kaiser, N. & Squires, G. 1993, ApJ, 404, 441
 Mellier, Y. 1999, ARA&A37, 127
 Miralda-Escudé, J. 1991, ApJ, 380, 1
 Misner, C. W. Thorne, K. & Wheeler, J. A. 1973, Gravitation, San Francisco,
 Freeman.
 Peacock, J. A. & Dodds, S. J. 1994, MNRAS267, 1020
 PeacockDodds:96 Peacock J. A. & Dodds S. J., 1996, MNRAS280, L19
 Peebles, P. J. E. 1980, The Large-Scale Structure of the Universe, Princeton
 University Press, Princeton, N.J., USA;
 Sachs, R. K. 1961, Proc. Roc. Soc. London, A264, 309
 Schneider, P., Ehlers, J., Falco, E. E. 1992, *Gravitational Lenses*, Springer.
 Seitz, C. & Schneider, P. 1995, A&A 297, 247
 Seitz, S. & Schneider, P., 1996 A&A 305, 388
 Seitz, S. & Schneider, P., Bartelmann, M. 1998, A&A 337, 325
 Soucail, G., Mellier, Y., Fort, B., Mathez, G. & Cailloux, M. 1988, A&A 191,
 L19
 Van Waerbeke, L., Mellier, Y., Erben, T., Cuillandre, J.C., Bernardeau, F.,
 Maoli, R., Bertin, E., Mc Cracken, H., Le Fèvre, O., Fort, B., Dantel-Fort,
 M., Jain, B. & Schneider, P. 2000, A&A 358, 30
 Villumsen, J. V. 1996, MNRAS281, 369
 Wittman, D. M., Tyson, J. A., Kirkman, D., Dell'Antonio, I. & Bernstein, G.
 2000, Nature 405, 143

Chapter 14

DARK MATTER: EARLY CONSIDERATIONS

Jaan Einasto

Tartu Observatory, 61602 Tõravere, Estonia

einasto@aai.ee

Abstract A review of the study of dark matter is given, starting with earliest studies and finishing with the establishment of the standard Cold Dark Matter paradigm in mid 1980-s. Particular attention is given to the collision of the classical and new paradigms concerning the matter content of the Universe. Also the amount of baryonic matter, dark matter and dark energy is discussed using modern estimates.

Keywords: Dark matter; galaxies; clusters of galaxies

1. Introduction

Dark matter in the Universe can be described as the matter which has practically zero luminosity and its presence can be detected only by its gravity. Historically, the first modern study of the possible presence of dark matter goes back to 1915, when Öpik (1915) determined the dynamical density of matter in our Galaxy in the Solar vicinity. The same problem was investigated by Oort (1932, 1960), Kuzmin (1952a, 1955) and more recently by Bahcall (1985) and Gilmore, Wyse & Kuijen (1989). Modern data suggest that there is little evidence for the presence of a large amount of local dark matter in the Solar vicinity. If some invisible matter is there, then it should be in the form of brown dwarfs, jupiters or similar compact baryonic objects.

A different type of dark matter is found around galaxies and in clusters of galaxies. The first evidence for the presence of such global dark matter was given by Zwicky (1933) from the dynamics of galaxies in the Coma cluster. The presence of dark matter in clusters was questioned, and an alternative solution to explain large velocities of galaxies in clusters was suggested by Ambartsumian (1958) – the instability of clusters of galaxies. However, the evi-

dence for the presence of invisible matter in systems of galaxies accumulated, first for our Local Group of galaxies (Kahn & Woltjer 1959), and thereafter for all giant galaxies (Einasto, Kaasik & Saar 1974, Ostriker, Peebles & Yahil 1974). These results were questioned by Burbidge (1975), Materne & Tammann (1976). Independent determination of rotation velocities of galaxies at large galactocentric distances (Rubin, Ford & Thonnard 1978, 1980) confirmed previous results on the presence of dark halos or coronas around galaxies. The nature of dark matter around galaxies is not clear. Initially it was assumed that it consists of hot gas (Kahn & Woltjer 1959, Einasto 1974b). Modern data favour the hypothesis that dark matter around galaxies is non-baryonic, either neutrinos or some weakly interacting massive particles, such as axions. The neutrino-dominated dark matter is called hot, since neutrinos move with very high velocities. The other type of dark matter is called cold, as particle velocities are moderate. The cosmological model with cold dark matter (CDM) was suggested by Blumenthal et al. (1984). This model is presently accepted as the standard. With the establishment of the cold dark matter concept the early period of the study of dark matter was completed.

Excellent reviews of the dark matter problem have been given by Faber & Gallagher (1979), Trimble (1987), Turner (1991) and Silk (1992), alternatives to dark matter have been discussed by Sanders (1990). In this report I describe how astronomers developed step-by-step the concept of dark matter. Such process is typical for the formation of a new paradigm in our understanding of the Universe. Particular attention is given to the work on galactic modelling which has lead us to the understanding of the structure of stellar populations and the need for a new invisible population of dark matter in galaxies. The Power-Point version of the present report is available on the web-site of Tartu Observatory, <http://www.aai.ee/~einasto>.

2. Local Dark Matter

Ernst Öpik started his studies, being a student of the Moscow University. One of the first problems he was curious about was the absorption of light in the Galaxy and the possible presence of absorbing (invisible) matter in it. He developed a method to determine the density of matter near the Galactic plane using vertical oscillations of stars. He concluded that there is no evidence for large amounts of invisible matter near the Galactic plane (Öpik 1915).

The dynamical density of matter in the Solar vicinity was investigated again by Oort (1932), who arrived at a different answer. According to his analysis the total density exceeds the density of visible stellar populations by a factor of up to 2. This limit is often called the Oort limit. This result means that the amount of invisible matter in the Solar vicinity could be approximately equal to the amount of visible matter.

The work on galactic mass modelling in Tartu Observatory was continued by Grigori Kuzmin. He developed a new method for galactic mass modelling using ellipsoids of variable density, and applied the theory to the Andromeda galaxy (Kuzmin 1943), using the recently published rotation data by Babcock (1939). Next Kuzmin turned his attention to our own Galaxy. Here the central problem was the density of matter in the Solar vicinity. The mass density can be calculated from the Poisson equation, where the dominating term is the derivative of the gravitational potential in the vertical direction. He found that this derivative can be expressed through the ratio of dispersions of velocities and coordinates in the vertical direction, $C = \sigma_z/\zeta_z$; here C is called the Kuzmin constant. Kuzmin (1952a, 1955) used data on the distribution of A and gK stars and analysed the results obtained in earlier studies by Oort (1932) and others. He obtained a weighted mean value $C = 68 \text{ km s}^{-1} \text{ kpc}^{-1}$, which leads to the density estimate $\rho = 0.08 M_{\text{sun}} \text{ pc}^{-3}$, in good agreement with direct density estimates of all known stellar populations (including estimates for the mass in invisible low-mass stars and white dwarfs). Two students of Kuzmin made independent analyses, using different methods and observational data (Eelsalu 1959, Jõeveer 1972, 1974) and confirmed Kuzmin results.

The local density problem was studied again by Hill (1960) and Oort (1960); both obtained considerably higher local densities of matter, and argued that there exist large amounts of dark matter in the Galactic disk. More recently Bahcall (1984) constructed a new multicomponent model of the Galaxy and determined the density of matter in the Solar vicinity, in agreement with the Oort's (1932, 1960) results. The discrepancy between various determinations of the matter density in the Solar vicinity was not solved until recently. Modern data have confirmed the results by Kuzmin and his collaborators (Gilmore, Wyse & Kuijken 1989). Thus we came to the conclusion that *there is no evidence for the presence of large amounts of dark matter in the disk of the Galaxy*. If there is some invisible matter near the galactic plane, then it is probably baryonic (low-mass stars or jupiters), since non-baryonic matter is dissipationless and cannot form a highly flattened population. Spherical distribution of the local dark matter (in quantities suggested by Oort and Bahcall) is excluded since in this case the total mass of the dark population would be very large and would influence also the rotational velocity.

3. Clusters and Groups of Galaxies

The mass discrepancy in clusters of galaxies was found by Zwicky (1933). He measured redshifts of galaxies in the Coma cluster and found that the total mass of the cluster calculated from the velocity dispersion using the virial theorem exceeds the sum of masses of visible galaxies more than tenfolds. He concluded that the cluster contains large amounts of invisible dark matter.

For some reasons the work of Zwicky escaped the attention of the astronomical community. The next step in the study of mass of systems of galaxies was made by Kahn and Woltjer (1959). They paid attention to the fact that most galaxies have positive redshifts as a result of the expansion of the Universe, only the Andromeda galaxy M31 has a negative redshift of about 120 km/s. This fact can be explained, if both galaxies, M31 and our Galaxy, form a physical system. A negative radial velocity indicates that these galaxies have already passed the apogalacticon of their relative orbit and are presently approaching each other. From the approaching velocity, mutual distance and time since passing the perigalacticon (taken equal to the present age of the Universe) the authors calculated the total mass of the double system. They found that $M_{tot} \geq 1.8 \times 10^{12} M_{sun}$. The conventional mass of the Galaxy and M31 is of the order of $2 \times 10^{11} M_{sun}$, in other words, the authors found evidence for the presence of additional mass in the Local Group of galaxies. The authors suggested that the extra mass is probably in the form of hot ionised gas; most of the paper was devoted to the analysis of the physical state of the gas. Using modern data Einasto & Lynden-Bell (1982) made a new estimate of the total mass of the Local Group, the result was $4.5 \pm 0.5 \times 10^{12} M_{sun}$ for present age of the Universe 14 Gyr. This estimate is in good agreement with new determinations of total masses of M31 and the Galaxy including their dark halos (see below).

The conventional approach for the mass determination of pairs and groups of galaxies is statistical. The method is based on the virial theorem and is almost identical to the procedure used to calculate masses of clusters of galaxies. Instead of a single pair or group a synthetic group is used consisting of a number of individual pairs or groups. These determinations yield for the mass-to-luminosity (in blue light) ratio the values $M/L_B = 1 \dots 20$ for spiral galaxy dominated pairs and $M/L_B = 5 \dots 90$ for elliptical galaxy dominated pairs (Page 1960, Burbidge & Burbidge 1961, van den Bergh 1961, Karachentsev 1976, Faber & Gallagher 1979).

The stability of clusters of galaxies was discussed in a special meeting during the IAU General Assembly (Neyman, Page & Scott 1961). Here the hypothesis of Ambartsumian on the expansion of clusters was discussed in detail. Van den Bergh (1961) drew attention to the fact that the dominating population in elliptical galaxies is the bulge consisting of old stars, indicating that cluster galaxies are old. It is very difficult to imagine how old cluster galaxies could form an unstable and expanding system. These remarks did not find attention and the problem of the age and stability of clusters remained open.

4. Masses of Galaxies

Galactic Models

The classical models of spiral galaxies were constructed using rotation velocities. In contrast, the models of elliptical galaxies were found from luminosity profiles and calibrated using central velocity dispersions or motions of companion galaxies. An overview of classical methods to construct models of galaxies is given by Perek (1962).

Problems of the structure of galaxies were a major issue at the Tartu Observatory since Öpik's (1922) work on the distance of the M31, where a simple hydrostatic model of this galaxy was constructed. This work was continued by Kuzmin who developed the major principles of galactic modelling, and applied these to calculate models of M31 and the Galaxy (Kuzmin 1943, 1952b, 1953, 1956a, b). These were first models with a continuous change of the spatial density (earlier sums of ellipsoids of constant density were used). However, individual populations of galaxies were not represented in these models, in contrast to the Schmidt (1956) model of the Galaxy where different populations were included with ellipsoids of constant density. The study of kinematic and physical properties of stellar populations was made independently. For a review of the early views on the structure of galactic populations see Oort (1958), in Tartu this problem was investigated by Rootsmaäe (1961).

A natural generalisation of classical and Kuzmin models was the explicit use of major stellar populations, such as the bulge, the disk, and the halo, as well as the flat population in spiral galaxies (consisting of young stars and interstellar gas). I did my PhD work on stellar kinematics in 1955 and turned thereafter my interest to galactic modelling. My goal was twofold: first, to get more accurate mass distributions in galaxies, and second, to find physical parameters of main stellar populations in both spiral and elliptical galaxies. My assumption was that similar stellar populations (say bulges) in galaxies of different morphological type should have similar physical parameters if their constituent stars have similar age and metallicity distribution. The methodical aspects of the new multicomponent models were discussed in a series of papers in Tartu Observatory Publications (in Russian with an English summary in Einasto 1969a). The spatial (or surface) density of practically all stellar populations can be expressed by a generalised exponential law (Einasto 1970b, 1974b, a similar expression has been used independently elsewhere)

$$\rho(a) = \rho(0) \exp \left[- \left(\frac{a}{ka_0} \right)^{1/N} \right], \quad (14.1)$$

where $\rho(0) = hM/(4\pi\epsilon a_0^3)$ is the central density, $a = \sqrt{R^2 + z^2/\epsilon^2}$ is the semi-major axis of the isodensity ellipsoid, a_0 is the effective (mean) radius

of the population, h and k are normalising constants, M is the mass of the population, ϵ is the axial ratio of isodensity ellipsoids, and N is a structural parameter, determining the shape of the density profile. Here we assume that isodensity ellipsoids are concentric and axially symmetric with a constant axial ratio for a given population. The case $N = 4$ corresponds to the de Vaucouleurs (1953) density law for spheroidal populations (halo), $N = 1$ corresponds to the classical exponential density law, and $N = 1/2$ to a Gaussian density law. The practical procedure of the model construction is the following. First, using photometric data for galaxies the structural parameters N of all major stellar populations are found. Next, using colorimetric and other data mass-to-luminosity ratios of populations are derived. Thereafter a preliminary mass distribution model is found and the rotation (actually circular) velocity is calculated and compared with observations. From the difference of the calculated and observed velocity corrections to model parameters are found. Initially these corrections were found using a trial-and-error procedure, later an automatic computer program was developed by our young collaborator Urmas Haud (Einasto & Haud 1989).

Mass-to-luminosity Ratios and Models of Physical Evolution of Stellar Populations

The method was applied to the Andromeda galaxy (Einasto 1969b, 1970a, Einasto & Rümmel 1970a), and to our Galaxy (Einasto 1970b). In the case of the Andromeda galaxy the mass distribution model found from the rotational data did not agree with the data on physical properties of populations. If we accepted the rotational velocities, based mostly on radio observations (Roberts 1966), then the mass-to-luminosity ratio, M/L , of central stellar populations became very low, of the order of 1 in Solar units. On the other hand, the spectral data (Spinrad 1966) suggested a much higher value, $M/L \approx 17$.

The next problem was to find internally constituent values of physical parameters of stellar populations of different age and composition. For this purpose I developed a model of physical evolution of stellar populations (Einasto 1971). When I started the modelling of physical evolution of galaxies I was not aware of similar work by Beatrice Tinsley (1968). When my work was almost finished I had the opportunity to read the PhD thesis by Beatrice. Both studies were rather similar, in some aspects my model was a bit more accurate (evolution was calculated as a continuous function of time whereas Beatrice found it for steps of 1 Gyr, also some initial parameters were different). Both models used the evolutionary tracks of stars of various composition (metallicity) and age, and the star formation rate by Salpeter (1955). I accepted a low-mass limit of star formation, $M_0 \approx 0.03 M_{\text{sun}}$, whereas Beatrice used a much lower mass limit to get higher mass-to-luminosity ratio for elliptical galaxies. My model

yields a continuous sequence of population parameters (colour, spectral energy distribution, M/L) as a function of age. The calculated parameters of stellar populations were compared with observational data by Einasto & Kaasik (1973). The available data supported relatively high values ($M/L \approx 10 - 30$) for old metal-rich stellar populations near centres of galaxies; moderate values ($M/L \approx 3 - 10$) for disks and bulges; and low values ($M/L \approx 1 - 3$) for metal-poor halo-type populations. Modern data yield slightly lower values, due to more accurate measurements of velocity dispersions in central regions of galaxies, and more accurate input data for models.

These calculations suggest that the rotation data by Roberts (1966) are biased. To find the reason for this biasing, I analysed the velocity field obtained from the radio observations. My analysis suggested that low rotational velocities in the central regions are due to a low spatial resolution of the radio beam (Einasto & Rümmel 1970b,c). The corrected velocity field was in agreement with a higher value of M/L in the central regions of M31, suggested by direct spectral data and models of physical evolution.

Mass Discrepancy on the Periphery of Galaxies

The second problem encountered in the modelling of M31 was the rotation and density distribution on the periphery. If the rotation data were taken at face value, then it was impossible to represent the rotational velocity with the sum of gravitational attractions by known stellar populations. The local value of M/L increases toward the periphery of M31 very rapidly if the mass distribution is calculated directly from the rotation velocity. All known old metal-poor halo-type stellar populations have a low $M/L \approx 1$; in contrast, on the basis of rotation data we got $M/L > 1000$ on the periphery of the galaxy, near the last point with a measured rotational velocity.

There were two possibilities to solve this controversy: to accept the presence of a new population with a very high M/L (a very uncommon property for an old stellar population), or to assume that on the periphery of galaxies there exist non-circular motions. We found that the first alternative had several serious difficulties. If the hypothetical population is of stellar origin, it must be formed much earlier than conventional populations, because all known stellar populations form a continuous sequence of kinematical and physical properties (Oort 1958, Rootsmaäe 1961, Einasto 1974a), and there is no place where to include this new population in this sequence. Secondly, the star formation rate is proportional to the square of the local density (Schmidt 1959, Einasto 1972), thus stars of this population should have been formed during the contraction phase of the formation of the population near its central more dense regions (where the density is largest), and later expanded to the present distance. The only source of energy for expansion is the contraction of other stellar popula-

tions. The estimated total mass of the new population exceeded the summed mass of all previously known populations. Estimates of the energy needed for the expansion demonstrated that the mass of the new population is so large that even the contraction of all other stellar populations to a zero radius would not be sufficient to expand the new population to its present size. And, finally, it is known that star formation is not an efficient process (usually in a contracting gas cloud only about 1 % of the mass is converted to stars); thus we have a problem how to convert, in an early stage of the evolution of the Universe, a large fraction of primordial gas into this population of stars. Taking into account all these difficulties I accepted the second alternative – the presence on non-circular motions (Einasto 1969b), similar to many other astronomers (see Materne & Tammann 1976). As I soon realised, this was a wrong decision.

Galactic Coronas

In spring 1972 I was asked to give an invited review on Galactic models at the First European Astronomy Meeting in Athens. At this time population models of galaxies had been calculated already for 5 galaxies of the Local Group and the giant elliptical galaxy M87 in the Virgo cluster. New rotation velocities suggested the presence of almost flat rotation curves on the periphery of galaxies, thus it was increasingly difficult to accept the previous concept of large non-circular motions. On the other hand, recently finished calculations of the physical evolution of stellar populations confirmed our previous view that it is extremely difficult to accept a stellar origin of the hypothetical population. My collaborator Enn Saar suggested to abandon the idea that only stellar populations exist in galaxies, to accept an idea that there may exist a population of unknown nature and origin and to look which properties it should have using available data on known stellar populations. Quickly a second set of models for galaxies was calculated, and parameters for the new dark population were found. To avoid confusion with the conventional halo population I suggested to call the new population “corona” (Einasto 1974b). The available data were insufficient to determine the outer radii and masses of coronas. Rough estimates indicated that in some galaxies the mass and radius of the corona may exceed considerably the mass and radius of all stellar populations, taken together.

To determine the parameters of galactic coronas more accurately distant test bodies are needed. After some period of thinking I realised how it is possible to check the presence of dark coronas around galaxies. If coronas are large enough, then in pairs of galaxies the companion galaxies move inside the corona, and their relative velocities can be used instead of galaxy rotation velocities to find the distribution of mass around giant galaxies. This test showed that the radii and masses of galactic coronas exceeded the radii and masses of

parent galaxies by an order of magnitude! Together with A. Kaasik and E. Saar we calculated new models of galaxies including dark coronas.

In those years Soviet astronomers had the tradition to attend Caucasus Winter Schools. Our results of galactic mass modelling were reported in a Winter School in 1972. The next School was held near the Elbrus mountain in a winter resort, in January 1974. The bottom line of my report was: *all giant galaxies have massive coronas, therefore dark matter must be the dominating component in the whole universe (at least 90 % of all matter)*. In the Winter School prominent Soviet astrophysicists as Zeldovich, Shklovsky, Novikov and others participated. In the discussion after the talk two questions dominated: What is the physical nature of the dark matter? and What is its role in the evolution of the Universe? A detailed report of this study was sent to "Nature" (Einasto, Kaasik & Saar 1974).

The need for massive halos had been already suggested by Ostriker and Peebles (1973) to stabilise galaxies against bar formation. Soon after our "Nature" paper Ostriker, Peebles and Yahil (1974) published similar results using similar arguments. They used the conventional term "halo" for the dark population apparently not realising that this population cannot be of stellar origin.

Dark Matter Conferences 1975

The importance of dark matter for cosmological studies was evident, thus Tartu astronomers organised in January 1975 a conference in Tallinn devoted solely to dark matter. Historically this was the first conference on dark matter. This conference is not well known, so I give here the list of major talks:

Zeldovich: "Deuterium nucleosynthesis in the hot Universe and the density of matter";
Einasto: "Dynamical and morphological properties of galaxy systems";
Ozernoy: "The theory of galaxy formation";
Zasov: "The masses of spiral galaxies";
Fessenko: "Difficulties of the study of dynamics of galaxy systems";
Novikov: "The physical nature of galactic coronas";
Saar: "Properties of stellar halos";
Doroshkevich: "Problems of the origin of galaxies and galaxy systems";
Komberg: "Properties of the central regions of clusters of galaxies";
Vorontsov-Velyaminov: "New data on fragmenting galaxies".

As we see, the emphasis of the conference was on the discussion of the physical nature of dark matter and its role in the formation of galaxies. These preliminary studies demonstrated that both suggested models for coronas had difficulties. It is very difficult to explain the physical properties of the stellar corona, also no fast-moving stars as possible candidates for stellar coronas were found.

Stellar origin of dark matter in clusters was discussed by Napier & Guthrie (1974); they find that this is possible if the initial mass function of stars is strongly biased toward very low-mass stars. Thorstensen & Partridge (1974) discussed the suggestion made by Cameron & Truran (1971) that there may have been a pregalactic generation of stars (called now population III), all of them more massive than the Sun, which are now present as collapsed objects. They conclude that the total mass of this population is negligible, thus collapsed stars cannot make up the dark matter.

The gaseous corona of galaxies and clusters was discussed by Field (1972), Silk (1974), Tarter & Silk (1974) and Komberg & Novikov (1975). The general conclusion from these studies is that coronas of galaxies and clusters cannot consist of neutral gas (the intergalactic hot gas would ionise the coronal gas), but a corona consisting of ionised gas would be observable. Modern data show that part of the coronal matter in groups and clusters of galaxies consists of X-ray emitting hot gas, but the amount of this gas is not sufficient to explain flat rotation curves of galaxies.

The dark matter problem was discussed also during the Third European Astronomical Meeting in summer 1975. In contrast to the Tallinn Meeting now the major dispute was between the supporters of the dark matter concept and the older paradigm with conventional mass estimates of galaxies. The major arguments against the dark matter concept were summarised by Materne & Tammann (1976). They were as follows (see also Burbidge 1975):

- The dark halo hypothesis is based on the assumption that companions are physical; if they are not then they do not measure the mass of the main galaxy, but characterise mean random velocities of galaxies;
- Groups of galaxies are bound with conventional masses; the mean mass-to-luminosity ratios of groups are 4 and 30 for spiral and elliptical dominated groups, respectively;
- The high masses of clusters may be explained by the high masses of the dominant cD galaxies; in other words – there is no extra mass in clusters;
- Big Bang nucleosynthesis suggests a low-density Universe with the density parameter $\Omega \approx 0.05$; the smoothness of the Hubble flow also favours a low-density Universe.

It was clear that by sole discussion only the presence and nature of dark matter cannot be solved, new data and more detailed studies were needed.

Are Pairs of Galaxies Physical?

In mid 1970s the main arguments for the presence of dark halos (coronas) of galaxies and clusters of galaxies were statistical. In particular, the masses of

double galaxies were determined by statistical methods. If companion galaxies used in mass determination are not real physical companions but random interlopers, then the mean velocity dispersion reflects random velocities of field galaxies and no conclusions on the mass distribution around giant galaxies can been made.

The difficulties connected with the statistical character of our arguments were discussed already during the Caucasus Winter School. Immediately after the school we started a study of properties of companion galaxies. The main question was: are companions true members of the satellite systems, which surround giant galaxies. Soon we discovered that companion galaxies are segregated morphologically: elliptical (non-gaseous) companions lie close to the primary galaxy whereas spiral and irregular (gaseous) companions of the same luminosity have larger distances from the primary galaxy; the distance of the segregation line from the primary galaxy depends on the luminosity of the primary galaxy (Einasto et al. 1974a). This result shows, first of all, that the companions are real members of these systems – random by-fliers cannot have such properties. Second, this result demonstrated that diffuse matter can have a certain role in the evolution of galaxy systems. The role of diffuse matter in galactic coronas was discussed in detail by Chernin, Einasto & Saar (1976). Morphological properties of companion galaxies can be explained, if we assume that at least part of the corona is gaseous. On the other hand, Komberg & Novikov (1975) demonstrated that coronas cannot be fully gaseous. Thus the nature of coronas remained unclear. Also we found that dynamical and morphological properties of primary galaxies are well correlated with properties of their companions (Einasto et al. 1976c). Brighter galaxies have companions which move with larger relative velocities than companions of fainter primaries. A further evidence of the large mass of the corona of our Galaxy came from the study of the dynamics of the Magellanic Stream (Einasto et al. 1976a).

The status of the dark matter problem in galaxies was discussed during the Commission 33 Meeting of the IAU General Assembly in Grenoble, 1976. Here arguments for the presence of dark halos and its non-stellar nature were again presented by Einasto, Jõeveer & Kaasik (1976b). But there remained two problems:

- If the massive halo (or corona) is not stellar nor gaseous, of what stuff is it made of?
- And a more general question: in Nature everything has its purpose. If 90 % of matter is dark, then this must have some purpose. What is the purpose of dark matter?

Additional Evidence for Dark Halos

In mid 1970s Vera Rubin and her collaborators developed new sensitive detectors to measure rotation curves of galaxies at very large galactocentric distances. Their results suggested that practically all spiral galaxies have extended flat rotation curves (Rubin, Ford & Thonnard 1978, 1980, see also a review by Rubin 1987). Now, for the first time, it was possible to determine the mass distribution in individual galaxies out to distances far superior to previous data. The internal mass of galaxies rised with distance almost linearly up to the last measured point (see Fig. 6 of Rubin et al. 1978). The concept of the presence of dark matter halos around galaxies was confirmed with a high confidence.

Another very important measurement was made by Faber et al. (1977). They measured the rotation velocity of the Sombrero galaxy, a S0 galaxy with a massive bulge and a very weak population of young stars and gas clouds just outside the main body of the bulge. Their data yielded for the bulge a mass-to-luminosity ratio $M/L = 3$, thus confirming our previous estimates based on less accurate data, and calculations of the physical evolution of galaxies. Velocity dispersion measurements of high accuracy also confirmed lower values of mass-to-luminosity ratios of elliptical galaxies (Faber & Jackson 1976). These results showed that the mass-to-luminosity ratios of stellar populations in spiral and elliptical galaxies are similar for a given colour (the assumption used in our model calculations), and the ratios are much lower than accepted in most earlier studies.

More recently the masses of clusters of galaxies have been determined using the temperature of hot X-ray emission gas in clusters, and by gravitational lensing. These data are discussed in other reports during this School.

By the end of 1970s most objections against the dark matter hypothesis were rejected. In particular, luminous populations of galaxies have found to have lower mass-to-luminosity ratio than expected previously, thus the presence of extra dark matter both in galaxies and clusters has been confirmed. However, the nature of dark matter and its purpose was not yet clear. Also it was not clear how to explain the Big Bang nucleosynthesis constraint on the low density of matter, and the smoothness of the Hubble flow.

5. The Nature of Dark Matter

Neutrino-dominated Universe

Already in 1970s suggestions have been made that some sort of non-baryonic elementary particles may serve as candidates for dark matter particles. Gunn et al. (1978) considered heavy stable neutral leptons as possible candidates for dark matter particles, however in a later study Tremaine & Gunn (1979)

rejected this possibility. Cowsik & McClelland (1973), Szalay & Marx (1976) and Rees (1977) noticed that neutrinos can be considered as dark matter particles. Doroshkevich et al. (1980a, b), Chernin (1981) and Doroshkevich & Khlopov (1984) showed that, if dark matter consists of heavy neutrinos (or, more generally, is non-baryonic), then this helps to explain the paradox of small temperature fluctuations of the cosmic microwave background radiation. Density perturbations of non-baryonic dark matter start growing already during the radiation-dominated era whereas the growth of baryonic matter is damped by radiation. If non-baryonic dark matter dominates dynamically, the total density perturbations can have an amplitude of the order 10^{-3} at the recombination epoch, which is needed for the formation of the observed structure of the Universe. This problem was discussed in a conference in Tallinn in April 1981. Here all prominent Soviet cosmologists and particle physicists participated (this conference was probably the birth of the astro-particle physics). The central problem was the nature of dark matter. In the conference banquet Zeldovich held an enthusiastic speech: “*Observers work hard in sleepless nights to collect data; theorists interpret observations, are often in error; correct their errors and try again; and there are only very rare moments of clarification. Today it is one of such rare moments when we have a holy feeling of understanding the secrets of Nature.*” Non-baryonic dark matter is needed to start structure formation early enough. This example illustrates well the attitude of theorists to new observational discoveries – the Eddington’s test: “*No experimental result should be believed until confirmed by theory*” (cited after Turner 2000). Now, finally, the presence of dark matter was accepted by leading theorists.

The search of dark matter can be illustrated with the words of Sherlock Holmes “*When you have eliminated the impossible, whatever remains, however improbable, must be the truth*” (cited by Binney & Tremaine 1987).

Dark Matter and the Structure of the Universe

After my talk at the Caucasus Winter School Zeldovich offered me collaboration in the study of the universe. He was developing a theory of formation of galaxies (the pancake theory, Zeldovich 1970); an alternative whirl theory was suggested by Ozernoy (1971), and a third theory of hierarchical clustering by Peebles (1971). Zeldovich asked for our help in solving the question: Can we find some observational evidence which can be used to discriminate between these theories?

Initially we had no idea how we can help Zeldovich. But soon we remembered our previous experience in the study of galactic populations: kinematical and structural properties of populations hold the memory of their previous evolution and formation (Rootsmäe 1961, Eggen, Lynden-Bell & Sandage 1962).

Random velocities of galaxies are of the order of several hundred km/s, thus during the whole lifetime of the Universe galaxies have moved from their place of origin only by about $1 h^{-1}$ Mpc (we use in this paper the Hubble constant in the units of $H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}$). In other words – if there exist some regularities in the distribution of galaxies, these regularities must reflect the conditions in the Universe during the formation of galaxies. Actually we had already some first results: the study of companion galaxies had shown that dwarf galaxies are located almost solely around giant galaxies and form together with giant galaxies systems of galaxies. In other words – the formation of galaxies occurs in larger units, not in isolation.

Thus we had a leading idea how to solve the problem of galaxy formation: *We have to study the distribution of galaxies on larger scales.* The three-dimensional distribution of galaxies, groups and clusters of galaxies can be visualised using wedge-diagrams, invented just when we started our study. My collaborator Mihkel Jõeveer prepared relatively thin wedge diagrams in sequence, and plotted in the same diagram galaxies, as well as groups and clusters of galaxies. In these diagrams regularity was clearly seen: *isolated galaxies and galaxy systems populated identical regions, and the space between these regions was empty.* This picture was quite similar to the distribution of test particles in a numerical simulation of the evolution of the structure of the Universe prepared by Doroshkevich et al. (1980) (preliminary results of this simulation were available already in 1975). In this picture a system of high- and low-density regions was seen: high-density regions form a cellular network which surrounds large under-dense regions.

We reported our results (Jõeveer & Einasto 1978) at the IAU symposium on Large-Scale Structure of the Universe in Tallinn 1977, the first conference on this topic. The main results were: (1) galaxies, groups and clusters of galaxies are not randomly distributed but form chains, converging in superclusters; (2) the space between galaxy chains contains almost no galaxies and forms holes (voids) of diameter up to $\approx 70 h^{-1}$ Mpc; (3) the whole picture of the distribution of galaxies and clusters resembles cells of a honeycomb, rather close to the picture predicted by Zeldovich. The presence of holes (voids) in the distribution of galaxies was reported also by other groups: Tully & Fisher (1978), Tifft & Gregory (1978), and Tarenghi et al. (1978) in the Local, Coma and Hercules superclusters, respectively. Theoretical interpretation of the observed cellular structure was discussed by Zeldovich (1978).

Our analysis gave strong support to the Zeldovich pancake scenario. This model was based essentially on the neutrino dominated dark matter model. However, some important differences between the model and observations were detected. First of all, there exists a rarefied population of test particles in voids absent in real data. This was the first indication for the presence of biasing in galaxy formation – there is primordial gas and dark matter in voids,

but due to low-density no galaxy formation takes place here (Jõeveer, Einasto & Tago 1978, Einasto, Jõeveer & Saar 1980). The second difference lies in the structure of galaxy systems in high-density regions: in the model large-scale structures (superclusters) have rather diffuse forms, real superclusters consist of multiple intertwined filaments (Zeldovich, Einasto & Shandarin 1982, Oort 1983, White, Frenk & Davis 1983, see also Bond, Kofman & Pogosyan 1996).

Cold Dark Matter

The difficulties of the neutrino-dominated model became evident in early 1980s. A new scenario was suggested by Blumenthal, Pagels & Primack (1982), Bond, Szalay & Turner (1982), and Peebles (1982); here hypothetical particles like axions, gravitinos or photinos play the role of dark matter. Numerical simulations of structure evolution for axion-gravitino-photino-dominated universe were made by Melott et al. (1983). All quantitative characteristics (the connectivity of the structure, the multiplicity of galaxy systems, the correlation function) of this new model fit the observational data well. This model was called subsequently the Cold Dark Matter (CDM) model, in contrast to the neutrino-based Hot Dark Matter model. Presently the CDM model with some modifications is the most accepted model of the structure evolution. The properties of the Cold Dark Matter model were analysed in detail in the classical paper by Blumenthal et al. (1984). With the acceptance of the CDM model the modern period of the study of dark matter begins.

Numerical simulations made in the framework of the Cold Dark Matter Universe (with and without the cosmological Λ -term) yield the distribution of galaxies, clusters and superclusters in good agreement with observations. These studies are too numerous to be cited here. Also the evolution of the structure can be followed by comparison of results of simulations at different epochs. During the School a movie was demonstrated showing the evolution of a central region of a supercluster (the movie was prepared at the Astrophysical Institute in Potsdam). Here the growth of a rich cluster of galaxies at the center of the supercluster could be followed. The cluster had many merger events and has “eaten” all its nearby companions. During each merger event the cluster suffers a slight shift of its position. As merger galaxies come from all directions, the cluster sets more and more accurately to the center of the gravitational well of the supercluster. This explains the fact that very rich clusters have almost no residual motion in respect to the smooth Hubble flow. According to the old paradigm galaxies and clusters form by random hierarchical clustering and could have slow motions only in a very low-density universe (an argument against the presence of large amount of dark matter by Materne & Tammann 1976).

The amount of dark matter

In early papers on dark matter the total density due to visible and dark matter was estimated to be 0.2 of the critical cosmological density (Einasto, Kaasik & Saar 1974, Ostriker, Peebles & Yahil 1974). These estimates were based on the dynamics of galaxies in groups and clusters. In subsequent years several new independent methods were suggested. A direct method is based on the distant supernova project, which yields (for a spatially flat universe) $\Omega_m = 0.28 \pm 0.05$ (Perlmutter et al. 1998, Riess 1998). Here and below density parameters are expressed in units of the critical cosmological density. Another method is based on X-ray data on clusters of galaxies, which gives the fraction of gas in clusters, $f_{gas} = \Omega_b/\Omega_m$. If compared to the density of the baryonic matter one gets the estimate of the total density, $\Omega_m = 0.31 \pm 0.05(h/0.65)^{-1/3}$ (Mohr et al. 2000). The evolution of the cluster abundance with time also depends on the density parameter (see Bahcall et al. 1999 for a review). This method yields an estimate $\Omega_m = 0.4 \pm 0.1$ for the matter density. The formal weighted mean of these independent estimates is $\Omega_m = 0.32 \pm 0.03$. This density value is close to the value $\Omega_m = 0.3$, suggested by Ostriker & Steinhardt (1995) as a concordant model.

More recently, the density parameter has been determined from clustering in the 2-degree Field Redshift Survey (Peacock et al. 2001), and from the angular power spectrum measurements of the cosmic microwave background radiation with the Wilkinson Microwave Anisotropy Probe (WMAP) (Spergel et al. 2003). The most accurate estimates of cosmological parameters are obtained using a combined analysis of the Sloan Digital Sky Survey and the WMAP data (Tegmark et al. 2003). According to this study the matter density parameter is $\Omega_m = 0.30 \pm 0.04$. This method yields for the Hubble constant the value $h = 0.70 \pm 0.04$ independent of other direct methods. From the same dataset the authors get for the density of baryonic matter, $h^2\Omega_b = 0.0232 \pm 0.0012$, which gives $\Omega_b = 0.047$ for the above value of the Hubble constant. Comparing both density estimates we get for the dark matter density $\Omega_{DM} = \Omega_m - \Omega_b = 0.25$.

6. Summary

People often ask: Who discovered dark matter? The dark matter story is a typical scientific revolution (Kuhn 1970, Tremaine 1987). As often in a paradigm shift, there is no single discovery, the new concept was developed step-by-step.

First of all, actually there are two dark matter problems – the local dark matter close to the plane of our Galaxy, and the global dark matter surrounding galaxies and clusters of galaxies. The milestones of the local dark matter problem solution are the studies by Öpik, Oort, Kuzmin, Bahcall and Gilmore.

Dark matter in the Galactic disk, if present, must be baryonic (faint stars or jupiters). The amount of local dark matter is low, it depends on the boundary between luminous stars and faint invisible stars.

The story of the global dark matter also spans many decades. It began with the work by Zwicky (1933) on the Coma cluster, was continued with the study by Kahn and Woltjer (1959) on the dynamics of the Galaxy-M31 system, and statistical determinations of masses and mass-to-luminosity ratios of pairs, groups and clusters of galaxies. For some reason, these studies did not awake the attention of the astronomical community. However, the awareness of the presence of a controversy with masses of galaxies and galaxy systems slowly increased.

Further development of the dark matter concept was influenced by the East-West controversy during the Cold War (on this controversy see Fairall 1998, p. 11 - 12). The dark matter puzzle was solved in 1974 by two independent studies of masses of galaxies by Tartu and Princeton astronomers. It was suggested that all giant galaxies are surrounded by massive halos (coronas), and that dark matter is dynamically dominant in the Universe. As usual in scientific revolutions, the general awareness of a crisis comes when the most eminent scientists in the field begin to concentrate on the problem. This happened when the Princeton group, Burbidge (1975) and Materne & Tammann (1976) published their contributions pro and contra the dark matter hypothesis. In the following years experimenters devoted themselves to finding new evidence in favour of (or against) the new paradigm. The work by Rubin and collaborators on galaxy rotation curves, our work on properties of satellite systems of galaxies and the Magellanic stream, X-ray studies of clusters, as well as investigation of gravitational lensing in clusters belong to this type of studies.

The word on the crisis spread more rapidly in the East: the first dark matter conference was held in Tallinn in 1975, the first official IAU dark matter conference was held only ten years later. The first popular discussions of the problem were given in "Priroda" and "Zemlya i Vselennaya" (the Russian counterparts of "Scientific American" and "Sky & Telescope") by Einasto (1975) and Einasto, Chernin & Jõeveer (1975), and also in the respective journal in Estonian. In USA the first popular discussions were given many years later (Bok 1981, Rubin 1983). However, most experimental studies confirming the dark matter hypothesis were made by US astronomers, and the cold dark matter concept was also suggested by Western astronomers.

The new paradigm wins when its theoretical foundation is established. In the case of the dark matter this was done by Blumenthal et al. (1984) with the non-baryonic cold dark matter hypothesis. Also the need for non-baryonic dark matter was clarified: otherwise the main constituents of the universe – galaxies, clusters and filamentary superclusters – cannot form.

In the following years main attention was devoted to detailed elaboration of the concept of the cold dark matter dominated Universe. Here a central issue was the amount of dark matter. Initially opinions varied from a moderate density of the order of 0.2 critical density up to the critical density. Only a few years ago it was clarified that dark matter constitutes only 0.25 of the critical density, and the rest is mostly dark energy, characterized by the cosmological constant or the Ω_Λ -term.

To conclude we can say that the story of dark matter is not over yet – we still do not know of what non-baryonic particles the dark matter is made of.

Acknowledgments

I thank M. Jõeveer and E. Saar for fruitful collaboration which has lasted over 30 years. This study was supported by the Estonian Science Foundation grant 4695.

References

- Ambartsumian, V. A. 1958, Solvay Conference Report, Brussels, p. 241
- Babcock, H.W. 1939, *Lick Obs. Bull.* 19, (498), 41
- Bahcall, J. N. 1984, ApJ287, 926
- Bahcall, N.A., Ostriker, J.P., Perlmutter, S., & Steinhardt, P.J., 1999, *Science* 284, 1482, astro-ph/9906463
- Binney, J. & Tremaine, S. 1987, *Galactic Dynamics*, Princeton, Princeton Univ. Press, p. 638
- Blumenthal, G.R., Faber, S.M., Primack, J.R. & Rees, M.J. 1984, Nature311, 517
- Blumenthal, G.R., Pagels, H., & Primack, J.R. 1982, Nature299, 37
- Bok, B.J. 1981, *Scientific American* 244, 92
- Bond, J.R., Kofman, L. & Pogosyan, D. 1996, Nature380, 603
- Bond, J.R., Szalay, A.S. & Turner, M.S., 1982, Phys. Rev. Lett.48, 1636
- Burbidge, E.M. & Burbidge, G.R. 1961, AJ66, 541
- Burbidge, G. 1975, ApJ196, L7
- Cameron, A.G.W. & Truran, J.W. 1971, *Astrophys. & Space Sci.* 14, 179
- Chernin, A.D. 1981, Astron. Zh.58, 25
- Chernin, A., Einasto, J. & Saar, E. 1976, *Astrophys. Space Sc.* 39, 53
- Cowsik, R. & McClelland, J. 1973, ApJ180, 7
- de Vaucouleurs, G. 1953, MNRAS113, 134
- Doroshkevich, A.G., Khlopov, M. Y. 1984, MNRAS211, 277
- Doroshkevich, A.G., Kotok, E.V., Poliudov, A.N., Shandarin, S.F., Sigov, Y.S., & Novikov, I.D. 1980, MNRAS192, 321
- Doroshkevich, A.G., Zeldovich, Y. B., Sunyaev, R. A. & Khlopov, M. Y. 1980a, Sov. Astr. Lett. 6, 252

- Doroshkevich, A.G., Zeldovich, Y. B., Sunyaev, R. A. & Khlopov, M. Y. 1980b,
Sov. Astr. Lett. 6, 257
- Eelsalu, H. 1959, Tartu Astr. Obs. Publ. 33, 153
- Eggen, O.J., Lynden-Bell, D. & Sandage, A. 1962, ApJ136, 748
- Einasto, J. 1969a, Astr. Nachr. 291, 97
- Einasto, J. 1969b, Astrofiz. 5, 137
- Einasto, J. 1970a, Astrofiz. 6, 149
- Einasto, J. 1970b, Tartu Astr. Obs. Teated No. 26, 1
- Einasto, J. 1971, Dr Habil. Thesis, Tartu University
- Einasto, J. 1972, *Astrophys. Let.* 11, 195
- Einasto, J. 1974a, in *Highlights of Astronomy*, ed. G. Contopoulos, Reidel, p. 419
- Einasto, J. 1974b, in *Proceedings of the First European Astr. Meeting*, ed. L.N. Mavrides, Springer: Berlin-Heidelberg-New York, 2, 291
- Einasto, J. 1975, Zemlya i Vselennaya No. 3, 32
- Einasto, J., Chernin, A.D. & Jõeveer, M. 1975, Priroda No. 5, 39
- Einasto, J. & Haud, U. 1989, A&A223, 89
- Einasto, J., Haud, U., Jõeveer, M. & Kaasik, A. 1976a, MNRAS177, 357
- Einasto, J., Jõeveer, M., & Kaasik, A. 1976b, *Tartu Astr. Obs. Teated*, 54, 3
- Einasto, J., Jõeveer, M., Kaasik, A. & Vennik, J. 1976c, A&A53, 35
- Einasto, J., Jõeveer, M. & Saar, E. 1980, MNRAS193, 353
- Einasto, J., & Kaasik, A., 1973, Astron. Tsirk. No. 790, 1
- Einasto, J., Kaasik, A., Kalamees, P. & Vennik, J. 1975, A&A40, 161
- Einasto, J., Kaasik, A. & Saar, E. 1974, Nature250, 309
- Einasto, J. & Lynden-Bell, D. 1982, MNRAS199, 67
- Einasto, J. & Rümmel, U., 1970a, Astrofiz. 6, 241
- Einasto, J. & Rümmel, U., 1970b, in *The Spiral Structure of Our Galaxy*, eds. W. Becker & G. Contopoulos, Reidel, p. 42
- Einasto, J. & Rümmel, U., 1970c, in *The Spiral Structure of Our Galaxy*, eds. W. Becker & G. Contopoulos, Reidel, p. 51
- Einasto, J., Saar, E., Kaasik, A. & Chernin, A.D. 1974a, Nature252, 111
- Faber, S.M., Balick, B., Gallagher, J.S. & Knapp, G.R. 1977, ApJ214, 383
- Faber, S.M., & Gallagher, J.S. 1979, ARA&A17, 135
- Faber, S.M., & Jackson, R.E. 1976, ApJ204, 668
- Fairall, A. 1998, *Large-scale Structures in the Universe*, Wiley, England
- Field, G.B. 1972, ARA&A10, 227
- Gilmore, G., Wyse, R.F.G. & Kuijken, K. 1989, ARA&A27, 555.
- Gunn, J.E., Lee, B.W., Lerche, I., Schramm, D.N. & Steigman, G. 1978, ApJ223, 1015
- Hill, E.R. 1960, Bull. Astr. Inst. Netherlands 15, 1.
- Jõeveer, M. 1972, Tartu Astr. Obs. Publ. 37, 3
- Jõeveer, M. 1974, Tartu Astr. Obs. Publ. 46, 35

- Jõeveer, M., & Einasto, J. 1978, in *The Large Scale Structure of the Universe*, eds. M.S. Longair & J. Einasto, Reidel, p. 241
- Jõeveer, M., Einasto, J., & Tago, E. 1978, MNRAS185, 35
- Kahn, F.D. & Woltjer, L. 1959, ApJ130, 705
- Karachentsev, I.D. 1976, *Stars and Galaxies from Observational Points of View*, ed. E.K. Kharadze, Mecniereba, Tbilisi, p. 439
- Komberg, B.V., & Novikov, I.D. 1975, Pisma Astron. Zh. 1, 3
- Kuhn, T.S. 1970, *The Structure of Scientific Revolutions*, Univ. of Chicago Press, Chicago
- Kuzmin, G.G. 1943, Tartu Astr. Obs. Kalender 1943, 85
- Kuzmin, G.G. 1952a, Tartu Astr. Obs. Publ. 32, 5
- Kuzmin, G.G. 1952b, Tartu Astr. Obs. Publ. 32, 211
- Kuzmin, G.G. 1953, Proc. Estonian Acad. Sc. 2, No. 3 (*Tartu Astr. Obs. Teated 1*)
- Kuzmin, G.G. 1955, Tartu Astr. Obs. Publ. 33, 3
- Kuzmin, G.G. 1956a, Astron. Zh.33, 27
- Kuzmin, G.G. 1956b, Proc. Estonian Acad. Sc. 5, 91 (*Tartu Astr. Obs. Teated 1*)
- Materne, J., & Tammann, G.A. 1976, in *Stars and Galaxies from Observational Points of View*, ed. E.K. Kharadze, Mecniereba, Tbilisi, p. 455
- Melott, A.L., Einasto, J., Saar, E., Suisalu, I., Klypin, A.A. & Shandarin, S.F. 1983, Phys. Rev. Lett.51, 935
- Mohr, J.J., Reese, E.D., Ellingson, E., Lewis, A.D., & Evrard, A.E., 2000, in *Constructing the Universe with Clusters of Galaxies*, (IAP 2000 Meeting, Paris), astro-ph/0004242
- Napier, W. McD. & Guthrie, B.N.G. 1975, MNRAS170, 7
- Neyman, J., Page, T. & Scott, E. 1961, AJ66, 533
- Oort, J.H. 1932, Bull. Astr. Inst. Netherlands 6, 249
- Oort, J.H. 1958, Ricerche Astron. Specola Vaticana 5, 415
- Oort, J.H. 1960, Bull. Astr. Inst. Netherlands 15, 45
- Oort, J.H. 1983, ARA&A21, 373
- Öpik, E. 1915, Bull. de la Soc. Astr. de Russie 21, 150
- Öpik, E. 1922, ApJ55, 406
- Ostriker, J.P., & Peebles, P.J.E. 1973, ApJ186, 467
- Ostriker, J.P., Peebles, P.J.E. & Yahil, A. 1974, ApJ193, L1
- Ostriker, J.P., & Steinhardt, P.J., 1995, Nature377, 600
- Ozernoy, L.M. 1971, Astron. Zh. 48, 1160
- Page, T.L. 1960, ApJ132, 910
- Peacock, J.A. et al. 2001, Nature410, 169
- Peebles, P.J.E. 1971, *Physical Cosmology*, Princeton Series in Physics, Princeton Univ. Press
- Peebles, P.J.E. 1982, ApJ263, 1

- Perek, L. 1962, Adv. Astron. Astrophys. 1, 165
- Perlmutter, S., et al. 1998, ApJ517, 565
- Rees, M. 1977, in *Evolution of Galaxies and Stellar Populations*, ed. B.M. Tinsley & R.B. Larson, New Haven, Yale Univ. Obs., 339
- Riess, A.G., 1998, AJ, 116, 1009
- Roberts, M.S., 1966, ApJ144, 639
- Rootsmäe, T. 1961, Tartu Astr. Obs. Publ. 33, 322
- Rubin, V.C. 1983, Scientific American 248, 88
- Rubin, V.C. 1987, in *Dark Matter in the Universe*, eds. J. Kormendy & G.R. Knapp, Reidel, Dordrecht, p. 51
- Rubin, V.C., Ford, W.K. & Thonnard, N. 1978, ApJ225, L107
- Rubin, V.C., Ford, W.K. & Thonnard, N. 1980, ApJ238, 471
- Salpeter, E.E. 1955, ApJ121, 161
- Sanders, R.H. 1990, A&ARev. 2, 1
- Schmidt, M. 1956, Bull. Astr. Inst. Netherlands 13, 14
- Schmidt, M. 1959, ApJ129, 243
- Silk, J. 1974, Comm. Astrophys. & Space Phys., 6, 1
- Silk, J. 1992, in *Stellar Populations*, eds. B. Barbuy & A. Renzini, Kluwer, Dordrecht, p. 367
- Spergel, D.N. et al. 2003, ApJS148, 175
- Spinrad, H., 1966, PASP78, 367
- Szalay, A.S. & Marx, G. 1976, A&A49, 437
- Tarter, J. & Silk, J. 1974, Q. J. Royal astr. Soc. 15, 122
- Tarenghi, M., Tifft, W.G., Chincarini, G., Rood, H.J. & Thompson, L.A. 1978, *The Large Scale Structure of the Universe*, eds. M.S. Longair & J. Einasto, Dordrecht: Reidel, p. 263
- Tegmark, M. et al. 2003, Phys. Rev. D(submitted), astro-ph/0310723
- Thornstensen, J.R. & Partridge, R.B. 1975, ApJ200, 527
- Tifft, W. G. & Gregory, S.A. 1978, *The Large Scale Structure of the Universe*, eds. M.S. Longair & J. Einasto, Dordrecht: Reidel, p. 267
- Tinsley, B.M., 1968, ApJ151, 547
- Tremaine, S. 1987, *Dark Matter in the Universe*, eds. J. Kormendy & G. R. Knapp, Dordrecht, Reidel, p. 547
- Tremaine, S., Gunn, J.E. 1979, Phys. Rev. Lett. 42, 407
- Trimble, V. 1987, ARA&A25, 425
- Tully, R.B. & Fisher, J.R. 1978, *The Large Scale Structure of the Universe*, eds. M.S. Longair & J. Einasto, Dordrecht: Reidel, p. 214
- Turner, M.S. 1991, Physica Scripta, T36, 167
- Turner, M.S. 2000, in *Type Ia Supernovae, Theory and Cosmology*, Edt. J.C. Niemeyer and J.W. Truran, Cambridge Univ. Press, p. 101 (astro-ph/9904049)
- van den Bergh, S. 1961, AJ66, 566
- White, S. D. M., Frenk, C. S. & Davis, M. 1983, ApJ274, L1

- Zeldovich, Ya.B. 1970, A&A5, 84
Zeldovich, Ya.B., 1978, *The Large Scale Structure of the Universe*, eds. M.S. Longair & J. Einasto, Dordrecht: Reidel, p. 409
Zeldovich, Ya.B., Einasto, J. & Shandarin, S.F. 1982, Nature300, 407
Zwicky, F. 1933, Helv. Phys. Acta 6, 110

Chapter 15

DARK MATTER AND GALAXY FORMATION

Joseph Silk

*Astrophysics, Denys Wilkinson Building,
Keble Road, Oxford OX1 3RH, UK*

silk@astro.ox.ac.uk

Abstract I describe recent challenges in dark matter. I review the budgets for baryonic and nonbaryonic dark matter. Problems with cold dark matter in the context of galaxy formation are summarized, and possible solutions are presented. I conclude with a description of the prospects for observing cold dark matter.

1. Challenges of dark matter

We are confronted by a paradox. Baryonic dark matter (BDM) exists and contributes to Ω_b . There are examples of baryonic dark matter, but we cannot reliably calculate the BDM mass fraction. On the other hand, cold dark matter (CDM) is motivated by theory and explains much of the large-scale structure of the universe. CDM dominates Ω_m . We can calculate the relic CDM mass fraction, but no CDM candidate particles are known to exist.

Current observations have attained considerable precision, as a result of surveys over much of the sky. These consist of cosmic microwave background maps, such as WMAP, and galaxy redshift surveys such as 2DF and SDSS. The combination of CMB and LSS experiments WMAP + CBI/ACBAR + 2DF, supplemented with the Lyman alpha forest analysis that probes the density correlations in the intergalactic medium, provides a measure of the homogeneity of the universe and the flatness of space: $\Omega_{total} = 1.02 \pm 0.02$. We infer that the age of the universe satisfies $t_0 = 13.7 \pm 0.2$ Gyr. There is known to be dominance of dark matter $\Omega_m = 0.27 \pm 0.02$ relative to baryonic matter. The baryon density amounts to $\Omega_b = 0.044 \pm 0.004$, and is about 15% of the dark matter. Dark energy with $\Omega_\Lambda = 0.70 \pm 0.03$ dominates over dark matter, but only contributes to the unclustered component of the density. The preceding results are sensitive to the adopted priors. The uncertainties can be increased

by an order of magnitude if more radical priors are adopted, such as inclusion of an admixture of primordial adiabatic and isocurvature density fluctuations.

Galaxy clustering also provides a powerful probe of the distribution of dark matter. One constrains the dark matter component by several independent techniques. These include virial theorem estimates with optical data on galaxy redshifts, x-ray emission and hydrostatic support of the hot intracluster gas, and gravitational lensing both of background galaxies and of cluster members. Strong lensing of remote galaxies via the formation of giant arcs probes the cluster core, and weak lensing of distortions to cluster galaxies via gravitationally-induced shear maps the outer region. The Sunyaev-Zeldovich effect supplements these studies by measuring the gas pressure.

2. Global baryon inventory

There is a significant but subdominant mass in dark baryons. Spheroid stars amount to 10% of the baryons or 0.004 in terms of Ω_b . Disk stars contribute 5% or 0.002 in Ω_b . Intracluster gas amounts to 5% or $\Omega_b = 0.002$. The Lyman alpha forest (at $z \sim 0$) contains $29 \pm 4\%$ of the baryons or $\Omega_b = 0.008$. This is all we actually observe in any quantifiable amount. In addition, intermediate temperature intergalactic gas, the so-called warm/hot intergalactic medium (WHIM) has been detected, at a temperature of $10^5 - 10^6$ K. It is estimated from simulations (at $z \sim 0$) to amount to 30% of the local baryons or $\Omega_b = 0.012$, with however a large uncertainty. Indeed the WHIM simulations do not resolve the Jeans mass at the resolution limit, and the existence of WHIM is purely a theoretical inference, at least in so far as its quantitative fraction is concerned.

The total baryonic contribution in the universe, including the hypothesised WHIM, sums to $\Omega_b = 0.028 \pm 0.005$. The corresponding baryon fraction Ω_b/Ω_m is 0.10 ± 0.02 . This is to be compared with primordial nucleosynthesis at $z \sim 10^9$: $\Omega_b = 0.04 \pm 0.004$. In addition the CMB peak heights at $z \sim 1000$ yield a similar value: $\Omega_b = 0.044 \pm 0.003$. Finally, Lyman alpha forest modelling at $z \sim 3$ suggests that $\Omega_b \approx 0.04$. There is also the indirect measurement of baryon fraction from the intracluster gas fraction of 15%. From this, combined with Ω_m , we also find that $\Omega_b \approx 0.04$.

I conclude that approximately $25 \pm 15\%$ of the baryons could be dark. If so, and this is far from a robust conclusion, one can pose the following question: where are they? Intriguingly, the possible shortfall is comparable to the mass observed in stars ($\sim \Omega_*$). Could there be a mass in *dark* stars comparable to that in visible stars? Or could early star formation and death have resulted in the ejection of a comparable mass of baryons from the galaxy and its halo? Were the situation to rest here, there would be little reason to take the issue of

dark baryons any further. However there is strong local evidence that there is a baryonic deficiency.

3. Confirmation via detailed census of MWG/M31

Observations of our galaxy and also of M31 give direct local measures of the baryon fraction. The virial mass is $M_{virial} \approx 10^{12} M_\odot$, whereas the mostly stellar baryon mass is $M_b \approx 6 \times 10^{10} M_\odot$. This yields a baryon fraction of $\sim 6\%$, comparable to the global baryon mass fraction *measured* in stars and gas at $z \sim 0$. A similar amount *may* be present in the intracluster medium. However it is the contents of a volume containing the galaxy, now and at formation, that concern us.

For comparison, the primordial baryon fraction from observations at high redshift is $\sim 15\%$. Presumably these baryons were present initially, when the galaxy formed. Indeed modelling of disk formation requires an initial baryon fraction of $\sim 10 - 15\%$ in order for sufficient cooling to have occurred to form the disk. The "missing" galactic baryons amount to a baryon fraction comparable to what is observed, namely around $\sim 5 - 10\%$ of the dark halo.

If these baryons are indeed in the halo, one possibility is that they are in the form of MACHOs, compact massive objects. The current limit on halo MACHOs is $\lesssim 20\%$ of the dark halo mass. The constrained mass range is $10^{-8} M_\odot - 10 M_\odot$. If the detection claimed by the MACHO collaboration is accepted, the preferred MACHO mass is $\sim 0.5 M_\odot$, which favours old white dwarfs: current searches for halo white dwarfs as faint, high proper motion red dwarfs are inconclusive. More generally, one might tolerate a wider mass range for the MACHOs. Theory does not exclude either primordial brown dwarfs ($0.01 - 0.1 M_\odot$), primordial black holes (mass $\gtrsim 10^{-16} M_\odot$) or cold dense H_2 clumps $\lesssim 1 M_\odot$ (invoked in the Milky Way halo to account, for example for extreme scattering events or unidentified Scuba objects). Such solutions are difficult to justify in any plausible theoretical framework. There is one intriguing counter-example to the subdominance of H_2 that is typical of spiral galaxies. Diffuse cold molecular gas observed in H_2 rotational emission in the outer disk of NGC 891 could conceivably account for the "missing" baryons, amounting to some 10 times the HI in mass, although there is no evidence for H_2 via H_2 absorption in another metal-poor galaxy nor in galaxies with normal abundances where constraints come from CO emission. The principle alternative is early ejection from the halo.

This could have occurred as a wind in the early, vigorously star-forming phase of galaxy evolution. This inference has additional credence from the facts that the mass in hitherto undetected baryons is comparable to that observed in stars and that the WHIM is observed to exist outside of rich clusters

where it is enriched. Moreover it could have a mass comparable to that both in stars and in the colder intergalactic medium.

Ejection via early winds is inferred in the enriched intracluster medium. Observations of Mpc-scale “holes” around Lyman break galaxies, detected via studying absorption of the IGM towards background quasars, support an explanation in terms of early winds from L_* galaxies. The so-called cooling catastrophe in galaxy formation theory, which results in overly luminous massive galaxies can be avoided if early winds eject an amount of baryons comparable to that retained in stars.

Hence the ““missing” baryons *could* be in the WHIM, which would be correspondingly enriched. Unfortunately such strong winds are not supported by hydrodynamical simulations. These use supernovae as the energy source that drives the wind. However the current multiphase simulations lack sufficient fine-scale resolution, as discussed below.

4. Hierarchical galaxy formation

Galaxy formation and dark matter are intimately related. Before I discuss the connection, I first review the current status of galaxy formation theory. The ab initio approach to large-scale structure has undergone a revolution in the past twenty years, with an improved understanding of the initial conditions of structure formation. This has come about as a confluence of theory and observation. Growth from inflation-boosted quantum fluctuations provides the current paradigm that sets the point of departure for virtually all theories of large-scale structure. The theory of structure growth made one notable prediction that has been verified with outstanding success. This was the existence of fossil cosmic microwave background temperature fluctuations imprinted on the last scattering surface of the cosmic microwave background. The fluctuations are on angular scales that correspond to the comoving scales of the observed large-scale structure in the galaxy distribution. The WMAP satellite, adding unprecedented precision to many earlier experiments, most notably those of BOOMERANG, MAXIMA and DASI, has verified to within a factor of order unity one of the most remarkable predictions of cosmology, thereby confirming the growth of structure via the gravitational instability of primordial density fluctuations

With the initial conditions specified, it became possible to simulate galaxy formation. Three distinct approaches have emerged: numerical, semi-analytical and hybrid. The fully numerical approach cannot yet cope with the complexities of star formation, but has been instrumental in guiding us towards an understanding of the dark matter distribution. The semi-analytical approach has had most success, because it can cope with a wide dynamic range via the extended Press-Schechter formalism, to which is added a prescription for star formation

based on baryonic dissipation and plausible but empirical rules. The hybrid approach, combining N-body simulations with a star formation prescription, is particularly useful for its predictive power in observational cosmology, as it is ideal for constructing mock catalogues of galaxies.

There have been some notable successes in the theory of semi-analytic galaxy formation. These include an understanding of the large-scale clustering of galaxies via the primordial density fluctuation power spectrum $P(k)$, including the two-point correlation function $\xi(r)$ and its higher moments, the predictions of the existence of filaments and sheets in the galaxy distribution and of the morphologies of galaxy clusters, the derivations of the cluster and galaxy mass functions, and the predictions of large-scale velocity fields and weak lensing optical depths. On smaller scales, the predictions of galaxy rotation curves and of strong lensing by massive galaxies and galaxy clusters are generally considered to be successes of the theory. Global results that have motivated many observations which are in general agreement with the theory include the cosmic star formation history and the distribution and evolution of HI clouds in the intergalactic medium.

There are complications, however, that demonstrate that we have not yet converged on the ultimate theory of galactic disk formation. There is no fundamental understanding of the Tully-Fisher correlation that relates galaxy luminosity to maximum rotational velocity. Models consistently give too high a normalisation of mass at a given rotation velocity, due to the predominance of dark matter in the model galaxies. There also is some question as to whether the slope is well understood both for samples of nearby disk galaxies which have been carefully corrected for inclination effects, and for distant disk galaxies when projected forward in time for comparison with current epoch samples. This may seem to be a detail, however the fundamental problems are twofold: most of the initial angular momentum in the theoretical models is lost to the dark halo as the disk forms, and the distribution of observed angular momentum is skewed towards high angular momentum in contrast to the initial distribution predicted by the simulations.

Nor is there any fundamental star formation theory for dynamically hot systems such as elliptical galaxies. Appeal must be made to phenomenology. Tidal interactions and mergers are found in simulations to be very effective at concentrating gas into the inner hundreds of parsecs. Ultraluminous infrared galaxies are observed to have star formation rates of hundreds or even thousands of solar masses per year, as inferred if the stars formed in a monolithic collapse of the system. Post-starburst near-infrared light profiles are also suggestive of forming spheroids. Since the ultraluminous infrared galaxies are almost inevitably associated with ongoing mergers or strong tidal interactions with nearby galaxies, it therefore seems entirely plausible that these conditions are capable of driving intense bursts of star formation at the prodigious star

formation rates that are observed. Measurements of the molecular gas masses in several such systems at high redshift demonstrate that a very high efficiency indeed of star formation is required, with some $10^{10} M_{\odot}$ of stars being inferred to form in 10^7 years. This is probably consistent with hierarchical galaxy formation provided that the efficiency of star formation was very high during the first major merger. Why star formation was so efficient is not understood, however.

Colours and spectra of elliptical galaxies at redshift of unity or beyond are suggestive of a very early formation epoch, at least for the stars if not for overall assembly. The cosmic star formation history is likely to be dominated by the precursors of today's ellipticals at $z \gtrsim 2$. Of course such a probe, which relies on galaxy surveys, is rest frame UV flux-limited. However the extra-galactic diffuse background light from FIR to optical/UV wavelengths provides a glimpse of all the star formation that ever occurred in the universe. It seems likely that forming dust-shrouded ellipticals dominate the far infrared background above $400\mu\text{m}$.

5. Unresolved issues in galaxy formation theory

One of the greatest puzzles in galaxy formation theory concerns the distribution of the dark matter. The cold dark matter concentration is predicted from N-body simulations to follow a density profile:

$$\rho = \frac{A}{r^{\gamma}(1+r/r_s)^{3-\gamma}}.$$

Here, r_s is a scale factor that is incorporated into the concentration parameter, $c \equiv r_v/r_s$, where r_v is either the virial radius or the radius at an overdensity, spherically-averaged, of 200. The profile slope parameter γ is measured in high resolution N-body simulations (,) to be $\gamma \approx 1.2 \pm 0.3$, and the normalisation parameter A reflects the epoch of formation, typically defined to be when half of the present mass was at overdensity of 200.

Unfortunately, observations seem to be in mild disagreement with this predicted profile merr. A low CDM concentration is observed in low surface brightness dwarf galaxies where the rotation curve is well measured. The predicted dark matter cusp is not usually seen; the typical profile has a soft core, although the interpretation is compounded by issues of disk inclination, of the HI distribution which is usually used to measure the rotation curve, and of the possible mismatch between baryon and CDM potential well depths.

The case for any discrepancy is weakened by the claim that the high resolution numerical simulations extrapolate to a core rather than to a cusp. This result however is disputed by other simulators who find a central slope $\gamma = 0.16 \pm 0.14$ that holds to 0.3% of the virial radius. In the Milky Way, a low concentration of nonbaryonic dark matter is inferred, with the argument

being forcefully made that no more than 10 percent of the total mass interior to the solar circle can be non-baryonic. Theory predicts something like 50 percent for a CDM-dominated universe. However the gravitational microlensing optical depth towards the bulge of the galaxy is used to assess the stellar contribution to the inner rotation curve, and this is uncertain by a factor of ~ 3 . This uncertainty has allowed modellers to fit the rotation curve with an NFW ($\gamma = 1$) halo initial density profile that is further concentrated by the adiabatic response of the halo to baryon dissipation.

Another issue is that of dark matter clumpiness. Large numbers of dwarf galaxy halos are predicted at masses comparable to those of the dwarf galaxies in the Local Group, exceeding the observed numbers by an order of magnitude or more. If these systems formed stars, they would be in gross disagreement with observations. If the angular momentum of the baryons is mostly lost to the dark halo as the baryons contract to form the disk, according to simulations, then disk sizes of spiral galaxies are predicted to be smaller by about a factor of 5 than observed. The baryons are clumped and lose angular momentum as a consequence of dynamical friction on the dark matter.

A related prediction is that of the galaxy luminosity function. If the mass in stars tracks that in dark matter, far too many small galaxies are predicted. Too many massive galaxies are also predicted. This has been noted both in isolated groups of galaxies at the L_* level and for the field luminosity function, where an excessive frequency of super- L_* galaxies is expected if a modern value for the initial baryon density is adopted. The problem arises because the baryons fall into the dark matter potential wells, cool and eventually form stars. There are simply too many cold baryons. If one begins with the baryon fraction predicted by primordial nucleosynthesis of about 15 percent, one ends up with about twice as many baryons as are seen even for the Milky Way galaxy. This issue has been aggravated by recent studies which show that many of the accreting baryons enter the disk cold, without shocking to the virial temperature. This appears to be the dominant form of accretion both for low mass galaxies and at high redshift.

6. Resurrecting CDM

It would seem that cold dark matter has certain difficulties to overcome. One approach is to tinker with the particle physics by modifying the dark matter, for example by introducing self-interacting or fluid dark matter. This approach is not only non-compelling from the physics perspective but it has also resulted in about as many new difficulties as it purports to resolve. Another strategy is to modify gravity. The less said about this the better: it seems to this author that one should only modify the laws of fundamental physics in the case of true desperation. We are not there yet.

A more promising approach is via astrophysics. The dark matter distribution is inevitably modified by the impact of astrophysical processes. These include dynamical feedback, such as via a massive, transient, rapidly rotating bar. Such gaseous bars are expected to form in the course of a major merger that preceded the first episode of star formation in the protogalaxy, and later would settle into the galactic disk. Indeed up to half of spiral galaxies have significant stellar bars. The initial tumbling of the bar is slowed by dynamical friction on the dark matter. This provides a substantial heat source that is capable of softening the CDM cusp into an isothermal core , but see for an independent appraisal of bar-halo angular momentum exchange. The converse consequence is that to explain the observed stellar bars that are generally in rapid rotation, one needs either a deficiency of dark matter, less than 10 percent of the total mass within the region where the bar is observed, or to argue that the observed bars are young. Cold gas infall to disks produces cold stellar disks that can subsequently become bar-unstable . The jury is still out on the history and secular evolution of bars.

A more radical astrophysical approach appeals to the formation of supermassive black holes in the protogalaxy. These must have formed contemporaneously with the oldest stars, as evidenced by the remarkable correlation between spheroid velocity dispersion and supermassive black hole mass that extends over more than 3 orders of magnitude. Gas accretion onto the supermassive black hole is inevitable in the gas-rich protogalaxy, and provokes violent outflows. It is these outflows that are viewed in the spectra of quasars, the most luminous objects in the universe, and which are powered by accretion onto supermassive black holes. These massive outflows of baryons could have provoked efficient star formation and preferentially expelled the low angular momentum gas. This is a promising, if largely unexplored, source of feedback into the protogalaxy, which offers a potential clue as to why disks and more generally galaxies are the sizes they are, why spheroids formed with great efficiency, why half of the baryons have apparently been expelled from massive galaxies, and why only high angular momentum gas remained to form the disk. As for the impact on the dark matter, rapid loss of more than half the mass in the inner core of the galaxy should leave an impact by softening the dark matter profile.

It is tempting to argue that the dark halo forms as a consequence of substructure mergers, thereby resolving part of the substructure problem if the dwarfs that do not merge are stripped of gas as they become tidally disrupted in the inner halo. The Sagittarius dwarf galaxy provides dramatic evidence of ongoing disruption: a substantial fraction of its mass has already been stripped . Dynamical and chemical evidence from studying tidal debris in the halo of our galaxy suggests that disrupted dwarfs might have contributed up to as much as $\sim 10\%$ of the stellar halo . Gravitational lensing provides evidence for the

survival of substructure in massive galaxy halos on $\sim 10^6 M_\odot$ mass scales, amounting to a few percent of the halo mass .

7. An astrophysical solution: early winds

Whether one opts for dynamical heating by tumbling bars or black hole-driven outflows, the impact on the protogalaxy is likely to be dramatic. Bars drive gas into the centre to form and eventually to fuel the SMBH. Mergers of smaller black holes play a subdominant role, especially in the early stages of black hole growth. Gas accretion onto the SMBH is responsible for fueling the jets and is inferred to be the dominant growth process. This conclusion follows from consideration of quasar outflows. Simulations demonstrate that twin radio jets interact with a clumpy interstellar medium to drive powerful winds that plausibly provide strong feedback into the protogalactic environment. Resolution of the overcooling problem for massive galaxies may require more energy input than is available with a normal initial stellar mass function. A more exotic solution is possible, and may even be necessary.

There are a number of reasons for believing that massive winds played an important role in galaxy formation. Enrichment of the intracluster gas is observed to $\sim Z_\odot/3$. This cannot be explained by current epoch star formation activity, or indeed by past activity unless substantial mass ejection occurred. Intracluster magnetic fields are observed at a level that is about 10 percent of the typical galactic value. Ejection of magnetic flux from galaxies early in the lifetime of the galaxies seems to be the most plausible explanation. At a redshift of about 3, the Lyman break galaxies are inferred to have outflows to velocities of $\sim 600 \text{ km/s}$. More indirectly, absorption against background quasars near these galaxies has revealed evidence for a proximity effect on the intergalactic medium. This is in the form of a deficiency of HI that is observed as an increase in the transparency to Ly α and possibly CIV absorption extending out to about $\sim 1 \text{ Mpc}$ from the Lyman break galaxies .

However numerical simulations of early supernova-driven winds fail to find any evidence for substantial gas ejection from luminous ($\sim L_*$) galaxies . One can ask what is wrong with the hydrodynamic simulations? Certainly, the simulations lack adequate resolution. Rayleigh-Taylor instabilities enhance wind porosity and Kelvin-Helmholtz instabilities enhance wind loading of the cold interstellar medium. Both effects are certain to occur and will enhance the wind efficacy. Yet another omission is that one cannot yet resolve the motions of massive stars before they explode. This means that energy quenching is problematic and the current results are inconclusive for typical massive galaxies.

An interesting from studies of nearby starbursts is that the mass outflow rate is of the same order of magnitude as the star formation rate. This result is also motivated by a multiphase interstellar medium in which supernova explosions

provide the main energy and momentum source , and naturally can account for why around half of the baryons are ejected. This mechanism should be effective for sub-L_{*} galaxies. It is unlikely however that there is sufficient energy available with a normal initial mass function to drive winds from the most massive forming galaxies. In fact, the simulation initial conditions assume that the winds are driven by supernovae produced by massive stars whose initial mass function is similar to that found in the solar neighbourhood. This is a dangerous assumption, given that we have no fundamental theory of the initial mass function, and that conditions both in massive starbursts and in the early universe may be quite different from anything sampled locally. A top-heavy initial mass function is one way to boost the specific energy and momentum input by up to an order of magnitude. It has been speculated that a top-heavy initial mass function is necessary to account for the high efficiencies of star formation observed in certain very high redshift ultraluminous infrared sources. This option has also been invoked in order to account for the surprisingly high redshift of reionisation found by the WMAP satellite and for the intracluster gas enrichment .

Another possibility is that some of the early supernovae may in fact be hypernovae. A hypernova has up to 10^{53} ergs of kinetic energy. If one supernova in 10 at high redshift is in fact a hypernova, the specific energy input is boosted by as much as an order of magnitude. The case for an enhanced hypernova fraction at high redshift is based on the nucleosynthetic evidence from abundances measured for the oldest stars in our halo. The enhancements of zinc and chromium and deficiency of iron in these stars can be explained in terms of hypernova yields. In hypernovae, the energy output is boosted by infall of the inner rotating core onto a black hole, and the corresponding ejecta mass cuts for precursors of $\sim 25M_{\odot}$ reflect the observed abundance anomalies relative to standard supernova yields . Some subset of hypernovae are also a possible source of the r-process nuclear enhancement seen in the oldest stars.

Finally, the ubiquitous AGN, as traced by the presence of supermassive black holes that amount to ~ 0.001 of the spheroid mass, would inevitably have been activated in the gas-rich protogalactic environment. The supermassive black hole is presumed to achieve most of its growth by gas accretion from a circumnuclear disk. This would inevitably have been accompanied by intense jets of relativistic plasma that provide a means of exerting strong positive feedback onto the protogalactic environment . Jet propagation into a clumpy interstellar medium destabilises the jet and generates high porosity in the relativistic jet fluid that eventually fills the entire diffuse interstellar medium . A likely consequence is the compression of massive clumps of cold gas and the subsequent triggering of star formation. Such positive feedback is precisely what may be required to account for the high efficiency of star formation inferred for massive ellipticals, as characterised by the observed enhancement in

$[\alpha/Fe]$ abundance ratios and by the predominance of red stellar populations at high redshift.

QSO observations attest to the role of AGN in star formation, with 30% of high redshift QSOs being luminous far infrared sources with substantial amounts of molecular gas either detected or inferred. The CO and FIR observations are indicative of high star formation rates of $\lesssim 1000 M_{\odot}$ per year, with the star formation efficiency rising for the most luminous FIR QSOs. The presence of large amounts of gas, heavy elements and dust in QSOs at $z \gtrsim 6$ argues for a causal connection between the AGN trigger and the high efficiency of star formation.

8. Observing CDM via the WIMP LSP

Assuming that the WIMPs once were in thermal equilibrium, one finds that the relic WIMP froze out at

$$n_x < \sigma_{ann} v > t_H \lesssim 1 \implies T \lesssim m_{\chi}/20k.$$

From this, one infers that the relic CDM density is $\Omega_x \sim \sigma_{weak}/\sigma_{ann}$. It is useful to know the mass range of the WIMPs in order to define search parameters. Minimal SUSY has many free parameters, and most of them are generally suppressed in parameter searches. For example, requiring the relic neutralino density to be within mSUGRA greatly reduces the parameter space for possible masses edsj. If the WIMP is a SUSY neutralino, simple scaling arguments yield

$$< \sigma_{ann} v > \propto m_{\chi}^2 \text{ for } m_{\chi} \ll Z^0$$

and

$$< \sigma_{ann} v > \propto m_{\chi}^{-2} \text{ for } m_{\chi} \gg Z^0,$$

thereby defining a window of opportunity for dark matter. Stability is assumed for the SUSY LSP to be a WIMP candidate, usually via R-parity conservation. From accelerator limits combined with model expectations, the allowed mass range is conservatively found to satisfy

$$50 \text{ GeV} \lesssim m_{\chi} \lesssim 1 \text{ TeV}.$$

Accelerator limits set a lower bound, and the inclusion of the extra degrees of freedom from coannihilations sets an upper bound. Direct searches may also independently set a model-dependent lower bound.

Indirect searches via halo annihilations of the LSP into $\gamma, \bar{p}, e^+, \nu$ have hitherto been inconclusive. There are hints of an anomalous feature in the high energy e^+ spectrum. However halo detection of e^+ requires clumpiness of order

$$< n^2 > / < n >^2 \sim 100,$$

both to get sufficient flux and to allow the possibility of a nearby clump which might allow the observed spectral feature to be reproduced . Such clumpiness could also boost the predicted gamma ray flux from annihilations into the range observable by EGRET. Clumpiness of this order is indeed predicted by galaxy halo simulations. However this generally applies in the outer halo. The γ -ray flux towards the galactic centre is observed to have a hard spectrum (as expected for annihilations), but the clumps would not survive the tidal disruptions that are inevitable in the inner galaxy . To account for the observed diffuse gamma ray flux from the direction of the galactic centre, one would need to have a very steep density profile ($\rho \propto r^{-1.5}$). It has been argued that this would conflict with microlensing observations and the inner rotation curve of the Galaxy. A detailed attempt at modelling the inner core including both rotation curve and microlensing constraints concludes however that the NFW profile combined with adiabatic compression of the dark matter leads to a consistent model with a potentially detectable gamma ray signal. Detection could be accomplished with an atmospheric Cerenkov telescope which has the advantage of good angular resolution and a threshold that could probe neutralinos with masses as low as $\sim 100\text{GeV}$.

9. The future

There are exciting prospects for addressing many of the challenges facing galaxy formation and dark matter. With regard to directly observing forming galaxies, we can look forward to sampling the galaxy luminosity function at redshifts beyond unity with both SIRTF and ground-based NIR spectroscopy. The theory of multiphase galaxy formation is certain to be greatly refined, incorporating dynamical feedback and the impact of supermassive black holes. We will probe scales down to $\sim 10^6 M_\odot$ via spectroscopic gravitational lensing. Baryonic dark matter will be mapped at UV/SXR wavelengths. In the area of indirect detection of CDM, new experiments will search for high energy halo annihilation signatures in the form of γ, e^+, \bar{p} and ν . Over the next 5 years, these experiments will include GLAST, HESS, MAGIC, VERITAS, ICECUBE, ANTARES, PAMELA and AMS. High energy neutrinos from annihilations in the sun (and earth) will be probed, thereby providing a measure of the cold dark matter density at the solar circle.

The Galactic Centre could provide a “smoking gun” with radio synchrotron, γ -ray and ν data: annihilations measure cold dark matter where Milky Way formation began “inside-out”, some 12 Gyr ago. Accretion models onto the central black hole fail to give sufficient low frequency radio or gamma ray emission to account for the observed fluxes from SagA*, and it is tempting to invoke a more exotic alternative. For example, the low frequency radio emission can be explained by spike-enhanced self-absorbed synchrotron emission,

contrary to recent claims. Even choice of equipartition magnetic fields is allowed with appropriate choice of a suitable initial dark matter density profile. In fact this choice depends on uncertain astrophysics that includes the past history of the formation of the SMBH, including mergers and associated heating and radiation recoil.

However even if the history of the supermassive black hole at the centre of the galaxy were to disfavour a significant cold dark matter spike, as a consequence of something as mundane as dynamical heating by stellar encounters , one might expect lesser spikes to survive around other relic massive black holes. The central supermassive black hole and the bulge of the galaxy most likely formed from the mergers of protogalactic dwarf galaxies that themselves contained smaller black holes. It is the seed black holes that formed by accretion which should retain initial CDM spikes. Dynamical mergers of black holes, if they occur, result in black hole ejection from shallow potential wells as a consequence both of formation of unstable 3-body systems and radiation recoil in 2-body mergers. This model suggests that there should be relic "naked" intermediate mass black holes in the inner halo . The adiabatic growth of these seed black holes should have generated local spikes in cold dark matter that could have survived and maintained a density profile

$$\rho \propto r^{-\gamma} \Rightarrow \rho \propto r^{-\gamma'}, \text{ with } \gamma' = \frac{9 - 2\gamma}{4 - \gamma}.$$

Annihilation fluxes would be enhanced, to a level where such sources could possibly account for a subset of the unidentified EGRET gamma ray sources.

The preceding interpretation rests heavily on the hypothesis that the dark matter consists primarily of the lightest $N = 1$ SUSY neutralinos. This is well motivated, but as has often been emphasized, the most compelling and elegant explanation of any natural phenomenon is often false. Of course, if accelerator evidence were found for SUSY, the odds in favour of a neutralino explanation of dark matter would be dramatically increased.

Another intriguing option is that of $N = 2$ SUSY, which could allow the possibility of two CDM relic candidate particles. One might take the dominant, heavier species to be the conventional WIMP, with a light, subdominant, counterpart with a correspondingly larger cross-section. The mass scaling of the gamma ray flux($\propto <\sigma v> m_\chi^{-2}$) allows the light dark matter candidate to be subdominant for a given flux and relic density (or equivalently, annihilation cross-section), if the cross-section is S-wave suppressed. A recent interpretation of 511 keV line emission from the Galactic centre and bulge region detected by the INTEGRAL gamma ray satellite appeals to MeV dark matter. These particles annihilate via e^+e^- pair production, and the positrons decelerate in the interstellar medium to generate a narrow annihilation line that simultaneously matches the observed flux and angular profile. There are al-

ternative explanations, and potential tests include searching for 511 keV line emission from the Sagittarius dwarf galaxy as well as from the brightest low mass x-ray binary stars. The former would confirm this idea, the latter would support an alternative, more conventional explanation in terms of astrophysical accelerators. Low mass x-ray binaries are accreting neutron stars that have a spatial distribution which is similar to that of Population II, and generate weak radio jets that are a possible injection source of energetic positrons into the interstellar medium.

Despite our failure to converge on a dark matter candidate, dark matter is here to stay. It is exceedingly difficult to construct a theory of galaxy formation without some compelling evidence for the nature of the dark matter. We assume that the dark matter is cold and stable, and this results in beautiful simulations of cosmic structure that meet many, but by no means all, of the observational challenges. Our hope is that with increasingly refined probes of galaxies near and far, we will be able to construct a strong inferential case for the required properties of the dark matter. Indeed, even now we are not far from this goal in so far as our modelling of large-scale structure is concerned.

On smaller scales, however, the picture, and the corresponding role of dark matter, is much less clear. It is particularly disconcerting that we know so little about the fundamental physics of star formation, despite decades of detailed observations. It is only too tempting to assume that conditions in the distant universe, while being far more extreme than those encountered locally, nevertheless permit us to adopt similar rules and inputs for star formation. We may be easily misled. Galaxy formation moreover rests on knowledge of the initial conditions that seeded structure formation, and that we measure in the cosmic microwave background. Here too it is worth recalling that our conclusions are only as robust as the initial priors. Change these substantially, and new modes of fluctuations are allowed that can, for example, permit a much earlier epoch of massive galaxy formation than in the standard model. It is clear that only increasingly refined and precise observations will guide us: if evidence were to be confirmed for a hypothesis that was far from our current prejudices, theory would rapidly adapt. We should bear in mind that Nature has more surprises than we can imagine, otherwise physics would be hopelessly dull.

I thank my colleagues especially at Oxford for their unwavering enthusiasm in discussions about many of the topics covered here. In particular I acknowledge the contributions of Reba Bandyopadhyay, Celine Boehm, Greg Bryan, Julien Devriendt, Ignacio Ferreras, Dan Hooper, Hugues Mathis, Adrienne Slyz and James Taylor.

References

- M. Bucher, J. Dunkley, P. G. Ferreira, K. Moodley & C. Skordis, PRL, submitted, astro-ph/0401417 (2004)
- S. Penton, J. Stocke & J. M. Shull, ApJ Suppl., in press, astro-ph/0401036 (2004)
- M. Walker & M. Wardle, ApJ, 498 (1998) L125
- A. Lawrence, MNRAS, 323 (2001) 147L
- E. Valentijn & P. van der Werf, ApJ, 522 (1999) L29
ApJ, 538 (2000) L77
- A. Boselli, J. Lequeux & G. Gavazzi, A&A, 384 (2002) 33
- D. Thilker et al., ApJ Letters, in press, astro-ph/0311571 (2003)
- G. Kauffmann et al., MNRAS, 341 (2003) 54
- T. Kranz, A. Slyz & H. Rix, ApJ, 586 (2003) 143
- C. Conselice et al., in IAU Symposium 220 "Dark Matter in Galaxies," Sydney, July 2003, to be published by ASP, astro-ph/0312352 (2003)
- F. Bertoldi et al., A&A Letters, in press, astro-ph/0307408 (2003)
- P. G. van Dokkum et al., ApJ, 587 (2003) L83 -L88
- C. Balland, J. E. G. Devriendt & J. Silk, MNRAS, 343 (2003) 107
- T. Fukushige, A. Kawai & J. Makino, ApJ, submitted, astro-ph/0306203 (2003)
- J. F. Navarro et al., MNRAS, submitted, astro-ph/0311231 (2003)
- M. R. Merrifield, in IAU Symposium 220 "Dark Matter in Galaxies," Sydney, July 2003, to be published by ASP, astro-ph/0310497 (2003)
- F. Stoehr, MNRAS, submitted, astro-ph/0403077 (2004)
- J. Diemand, B. Moore & J. Stadel, MNRAS, submitted, astro-ph/0402267 (2004)
- R. R. Islam, J. E. Taylor & J. Silk, MNRAS, in press, astro-ph/0307171 (2003)
- C. Maraston et al., A&A, 400 (2003) 823-840
- E. D'Onghia & G. Lake, ApJ Letters, submitted, astro-ph/0309735 (2003)
- A. J. Benson, R. G. Bower, C. S. Frenk, C. G. Lacey, C. M. Baugh & S. Cole, ApJ, submitted, astro-ph/0302450 (2003)
- Y. Birnboim & A. Dekel, MNRAS, in press, astro-ph/0302161 (2003)
- K. Holley-Bockelmann, M. Weinberg. & N. Katz, MNRAS, submitted, astro-ph/0306374 (2003)
- E. Athanassoula, in IAU Symposium 220 "Dark Matter in Galaxies", Sydney, July 2003, to be published by ASP, astro-ph/0312155 (2003).
- D. L. Block, F. Bournaud, F. Combes, I. Puerari & R. Buta, in IAU Symposium 217 "Recycling Intergalactic and Interstellar Matter", Sydney, July 2003, to be published by ASP (2003)
- J. Silk, MNRAS, 343 (2003) 249

- S. Majewski et al., Contribution to proceedings of “IAU Symposium 220: Dark Matter in Galaxies”, eds. S. Ryder, D.J. Pisano, M. Walker, & K. Freeman, to be published by ASP, astro-ph/0311522 (2003)
- E. Tolstoy & K. Venn, to be published in the Proceedings of Joint Discussion 15 of the 2003 IAU, eds. P. Nissen & M. Pettini, astro-ph/0402295 (2004)
- R. B. Metcalf, L. A. Moustakas, A. J. Bunker & I. R. Parry, ApJ, submitted, astro-ph/0309738 (2003)
- K.L. Adelberger, C.C. Steidel, A.E. Shapley & M. Pettini, ApJ, 584 (2003) 45-75
- V. Springel & L. Hernquist, MNRAS, 339 (2003) 289
- B. Ciardi, A. Ferrara & S. D. M. White, MNRAS, 344 (2003) L7
- L. Portinari, A. Moretti, C. Chiosi, J. Sommer-Larsen, ApJ, in press, astro-ph/0312360 (2003)
- H. Umeda & K. Nomoto, ApJ, submitted, astro-ph/0308029 (2003)
- W. van Breugel, C. Fragile, P. Anninos & S. Murray, in IAU Symposium 217 “Recycling Intergalactic and Interstellar Matter”, Sydney, July 2003, to be published by ASP (2003).
- G. Bicknell, C. Saxton, R. Sutherland, S. Midgley & S. Wagner, New Astronomy Reviews, 47 (2003) 537 (2004)
- F. Bertoldi et al., A&A, 406 (2003) L55-L58
- Y. Gao & P. Solomon, ApJ, in press, astro-ph/0310339 (2003)
- J. Edsjo, M. Schelke, P. Ullio & P. Gondolo, JCAP 0304 (2003) 001
- D. Hooper, J. E. Taylor & J. Silk, PRD, submitted, hep-ph/0312076 (2003)
- F. Stoehr, S. D. M. White, V. Springel, G. Tormen & N. Yoshida, MNRAS, 345 (2003) 1313
- D. Merritt, preprint, astro-ph/0311594 (2003)
- F. Prada et al., PRL, submitted, astro-ph/0401512 (2004)
- C. Boehm et al., PRL, in press, astro-ph/0309686 (2004)

Chapter 16

NON-BARYONIC DARK MATTER

Paolo Gondolo

Department of Physics, University of Utah, 115 South 1400 East, Suite 201, Salt Lake City, Utah 84112, USA

pao@physics.utah.edu

These lectures on non-baryonic dark matter are divided into two parts. In the first part, I discuss the need for non-baryonic dark matter in light of recent results in cosmology, and I present some of the most popular candidates for non-baryonic dark matter. These include neutrinos, axions, neutralinos, WIMPZILLAs, etc. In the second part, I overview several observational techniques that can be employed to search for WIMPs (weakly interacting massive particles) as non-baryonic dark matter. Among these techniques, I discuss the direct detection of WIMP dark matter, and its indirect detection through high-energy neutrinos, gamma-rays, positrons, etc. References cited in these lectures are intended mostly for further study, and no attempt has been made to provide a comprehensive list of original and recent work on the subject.

1. The need for non-baryonic dark matter

We live in a time of great observational advances in cosmology, which have given us a consistent picture of the matter and energy content of our Universe. Here matter and energy (which special relativity tells us are equivalent) are distinguished by their different dependence on the cosmic volume: matter density decreases with the inverse of the volume, while energy density remains (approximately) constant.

Nothing is known about the nature of the energy component, which goes under the name of dark energy. Of the matter component, less than 2% is luminous, and no more than 20% is made of ordinary matter like protons, neutrons, and electrons. The rest of the matter component, more than 80% of the matter, is of an unknown form which we call non-baryonic. Finding the nature of non-baryonic matter is referred to as the non-baryonic dark matter problem.

A summary of the current measurements of the matter density Ω_m and the energy density Ω_Λ are shown in Figure 1 (adapted from Verde et al.(2002)). Both are in units of the critical density $\rho_{\text{crit}} = 3H_0^2/(8\pi G)$, where G is the

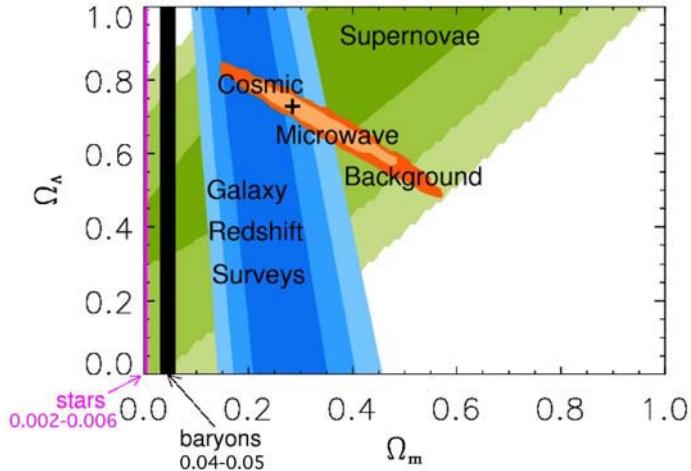


Figure 16.1. The concordance cosmology and the need for non-baryonic dark matter. Current cosmological measurements of the matter density Ω_m and energy density Ω_Λ give the value marked with a cross at $\Omega_m \simeq 0.27$, $\Omega_\Lambda \simeq 0.73$. The baryon density does not exceed 0.05 (black vertical band). The rest of the matter is non-baryonic. (Figure adapted from Verde et al.(2002).)

Newton's gravitational constant, and H_0 is the present value of the Hubble constant. Three types of observations – supernova measurements of the recent expansion history of the Universe, cosmic microwave background measurements of the degree of spatial flatness, and measurements of the amount of matter in galaxy structures obtained through big galaxy redshift surveys – agree with each other in a region around the best current values of the matter and energy densities $\Omega_m \simeq 0.27$ and $\Omega_\Lambda \simeq 0.73$ (cross in Figure 1). Measurements of the baryon density in the Universe using the cosmic microwave background spectrum and primordial nucleosynthesis constrain the baryon density Ω_b to a value less than ~ 0.05 (black vertical band in Figure 1). The difference $\Omega_m - \Omega_b \simeq 0.22$ must be in the form of non-baryonic dark matter.¹

A precise determination of the cosmological density parameters is able to give the matter and energy densities in physical units. For example, in units of $1.879 \times 10^{-29} \text{ g/cm}^3 = 18.79 \text{ yg/m}^3$, Spergel et al.(2003) have determined a total matter density

$$\Omega_m h^2 = 0.135^{+0.008}_{-0.009}, \quad (16.1)$$

¹The red vertical band labeled 'stars' in Figure 1 shows the density of luminous matter, corrected for all expected dim stars and gas ?, see, e.g., Fukugita:1998. The difference between the amount of luminous matter and the amount of baryons constitutes the dark baryon problem, which will not be addressed here ?, see, e.g., Silk:2003.

of which

$$\Omega_\nu h^2 < 0.0076 \quad (16.2)$$

is in the form of neutrinos (to 95% confidence level),

$$\Omega_b h^2 = 0.0224 \pm 0.0009 \quad (16.3)$$

is in the form of baryons (protons and nucleons in cosmological parlance), and

$$\Omega_{\text{CDM}} h^2 = 0.113^{+0.008}_{-0.009} \quad (16.4)$$

is in the form of cold dark matter (CDM), a non-baryonic component whose nature we are still trying to uncover. Some ideas of what it may be are presented in the next Section.

2. Popular candidates for non-baryonic dark matter

A new kind of elementary particle has been the dominant (exclusive?) candidate for non-baryonic dark matter.

A major classification of non-baryonic dark matter is based on its temperature at the time of galaxy formation, which occurs at a photon temperature of about 1 keV. Hot dark matter was relativistic at the time of galaxy formation, and as a consequence hindered the formation of the smallest objects by streaming out of the forming structures. An example of a hot dark matter particle is a light neutrino, much lighter than \sim keV. Cold dark matter was non-relativistic when galaxies formed, and thus was able to collapse effectively under the action of gravity because of its negligible pressure. Examples of cold dark matter particles are neutralinos, axions, WIMPZILLAs, solitons (B-balls and Q-balls), etc. Warm dark matter was semi-relativistic at the time of galaxy formation, and is therefore an intermediate case between hot and cold dark matter. Two examples of warm dark matter are keV-mass sterile neutrinos and gravitinos.

Another important classification of particle dark matter rests upon its production mechanism. Particles that were in thermal equilibrium in the early Universe, like neutrinos, neutralinos, and most other WIMPs (weakly interacting massive particles), are called thermal relics. Particles which were produced by a non-thermal mechanism and that never had the chance of reaching thermal equilibrium in the early Universe are called non-thermal relics. There are several examples of non-thermal relics: axions emitted by cosmic strings, solitons produced in phase transitions, WIMPZILLAs produced gravitationally at the end of inflation, etc.

For the sake of presentation, we find still another classification useful. We will divide candidates for particle dark matter into three categories: Type Ia, Type Ib, and Type II (following a common practice in superconductors and supernovas). Type Ia candidates are those known to exist, foremost among

them are the neutrinos. Type Ib candidates are candidates which are still undiscovered but are ‘well-motivated.’ By this we mean that (1) they have been proposed to solve genuine particle physics problems, a priori unrelated to dark matter, and (2) they have interactions and masses specified within a well-defined (and consistent) particle physics model. We are aware of the arbitrariness of this classification, and reserve the honor of belonging to the Type Ib category only to a sterile neutrino, the axion, and the lightest supersymmetric particle (which may be a neutralino, a gravitino, or a sneutrino) Finally, Type II candidates are all other candidates, some of which are examples of maybe fruitful ideas, such as WIMPZILLAs, solitons (B-balls, Q-balls), dark matter from extra-dimensions, self-interacting dark matter, string-inspired dark matter, string-perspired dark matter, etc. It goes without saying that a candidate may move up from Type II to Type Ib and even to Type Ia as our understanding of particle physics models progresses.

We now examine some of the current candidates.

Type Ia: candidates that exist

Dark matter candidates that are known to exist in Nature have an obvious advantage over candidates that have not been detected. The chief particles in this category are the neutrinos.

There are three known ‘flavors’ of neutrinos: the electron neutrino ν_e , the muon neutrino ν_μ , and the tau neutrino ν_τ . They are so named because they are produced or destroyed in concomitance with the electron, the muon, and the tau lepton, respectively.

“If neutrinos had a mass, they would be a good candidate for the dark matter,” said Steven Hawking. We now know that neutrinos, or at least some of the neutrinos, do have a mass. This was discovered indirectly through the observation of neutrino flavor oscillations, i.e. the spontaneous conversion of one neutrino flavor into another as a neutrino propagates from point to point. The connection between flavor oscillations and neutrino masses can be seen as follows.

Consider for simplicity two flavors of neutrinos instead of three, ν_e and ν_μ , say. Weak interactions produce the flavor eigenstates $|\nu_e\rangle$ and $|\nu_\mu\rangle$, which are associated with their respective charged leptons. However, these flavor eigenstates are *not* energy eigenstates. Let $|\nu_1\rangle$ and $|\nu_2\rangle$ denote the two energy eigenstates for the two-flavor system, with energies E_1 and E_2 respectively. Then the flavor and the energy eigenstates are connected by a unitary transformation,

$$\begin{cases} |\nu_e\rangle = \cos\theta|\nu_1\rangle + \sin\theta|\nu_2\rangle, \\ |\nu_\mu\rangle = -\sin\theta|\nu_1\rangle + \cos\theta|\nu_2\rangle. \end{cases} \quad (16.5)$$

Imagine that at time $t = 0$ we produce a ν_e , so that the initial wave function is

$$|\psi(0)\rangle = |\nu_e\rangle. \quad (16.6)$$

After a time t , the wave function evolves according to the system Hamiltonian \hat{H} as (we use natural units $\hbar = c = 1$)

$$|\psi(t)\rangle = e^{-i\hat{H}t}|\psi(0)\rangle = e^{-i\hat{H}t}|\nu_e\rangle. \quad (16.7)$$

To see the evolution explicitly, we expand $|\nu_e\rangle$ into energy eigenstates, and obtain

$$|\psi(t)\rangle = \cos \theta e^{-iE_1 t} |\nu_1\rangle + \sin \theta e^{-iE_2 t} |\nu_2\rangle. \quad (16.8)$$

We can now ask what is the probability of observing the neutrino in the state $|\nu_\mu\rangle$ after a time t , i.e. of observing a neutrino with flavor ν_μ instead of the initial ν_e . According to standard rules of quantum mechanics, this probability is

$$\text{Prob}(\nu_e \rightarrow \nu_\mu) = |\langle \nu_\mu | \psi(t) \rangle|^2 \quad (16.9)$$

$$= |\cos \theta e^{-iE_1 t} \langle \nu_\mu | \nu_1 \rangle + \sin \theta e^{-iE_2 t} \langle \nu_\mu | \nu_2 \rangle|^2 \quad (16.10)$$

$$= |- \cos \theta \sin \theta e^{-iE_1 t} + \sin \theta \cos \theta e^{-iE_2 t}|^2 \quad (16.11)$$

$$= \sin^2 2\theta \sin^2 \left[\frac{1}{2}(E_2 - E_1)t \right], \quad (16.12)$$

where we have used Eq. (16.5) and $\langle \nu_1 | \nu_1 \rangle = \langle \nu_2 | \nu_2 \rangle = 1$, $\langle \nu_1 | \nu_2 \rangle = 0$. For free relativistic neutrinos, with momentum p much larger than their mass m , we have

$$E = \sqrt{p^2 + m^2} \simeq p + \frac{m^2}{2p}, \quad (16.13)$$

and

$$E_2 - E_1 \simeq \frac{m_2^2 - m_1^2}{2p}. \quad (16.14)$$

Hence

$$\text{Prob}(\nu_e \rightarrow \nu_\mu) = \sin^2 2\theta \sin^2 \left[\frac{(m_2^2 - m_1^2)t}{4p} \right]. \quad (16.15)$$

This equation shows that the probability of conversion from flavor ν_e to flavor ν_μ oscillates in time with a frequency proportional to the difference of the squares of the neutrino masses $\Delta m_{12}^2 = m_2^2 - m_1^2$. The observation of neutrino flavor oscillations therefore implies that neutrino masses differ from each other, and in particular that at least one of them is different from zero.

Neutrino oscillations have up to now been detected in two systems. Atmospheric muon neutrinos, which originate from the collision of cosmic rays

with the Earth atmosphere, have been observed to oscillate into tau neutrinos (Fukuda et al.(1998)),

$$\nu_\mu \rightarrow \nu_\tau, \quad \Delta m_{23}^2 \sim 3 \times 10^{-3} \text{ eV}^2. \quad (16.16)$$

Solar neutrinos, produced in the nuclear reactions that make the Sun shine, also show oscillations (Ahmad et al.(2002)),

$$\nu_e \rightarrow \nu_\mu \text{ or } \nu_\tau, \quad \Delta m_{12}^2 \sim 7 \times 10^{-5} \text{ eV}^2. \quad (16.17)$$

These results can be used to set a lower limit on the mass of the heaviest neutrino. Indeed, the mass of the heaviest neutrino must be greater than or equal to the square root of the largest mass-squared difference (just take the mass of the other neutrino to vanish). This gives the lower limit

$$\text{mass of heaviest neutrino} \gtrsim 0.05 \text{ eV}. \quad (16.18)$$

Upper limits on neutrino masses come from laboratory experiments, such as tritium decay and high-energy accelerator experiments, and are (see Review of Particle Physics, Hagiwara et al.(2002))

$$m_1 < 2.8 \text{ eV}, \quad m_2 < 190 \text{ keV}, \quad m_3 < 18.2 \text{ MeV}. \quad (16.19)$$

However, the small mass differences implied by Eqs. (16.16) and (16.17) imply that the smallest of the three upper limits applies to all three active neutrino masses. Thus we have

$$m_i < 2.8 \text{ eV} \quad (i = 1, 2, 3). \quad (16.20)$$

It follows from this mass constraint that reactions such as $\nu_e \bar{\nu}_e \leftrightarrow e^+ e^-$ in the hot early universe were able to keep standard-model neutrinos in thermal equilibrium. The neutrino density then follows from a computation of the neutrino number density (see Section 2.2). The result is

$$\Omega_\nu h^2 = \sum_{i=1}^3 \frac{g_i m_i}{90 \text{ eV}}, \quad (16.21)$$

where $g_i = 1$ for a neutrino which is its own antiparticle (Majorana neutrino) and $g_i = 2$ for a neutrino which is not its own antiparticle (Dirac neutrino).

We already mentioned in Section 1 that cosmology provides an upper limit on the neutrino density $\Omega_\nu h^2$. This translates into a cosmological upper limit on the neutrino mass using Eq. (16.21). The cosmological limit is strictly speaking on the mass density in relativistic particles at the time of galaxy formation. An excessive amount of relativistic particles when galaxies form, i.e. of particles with mass $m \ll \text{keV}$, would erase too much structure at the smallest scales. A combination of cosmic microwave background measurements,

galaxy clustering measurements, and observations of the Lyman- α forest gives the upper limit quoted before Spergel et al.(2003)

$$\Omega_\nu h^2 < 0.0076 \quad (95\% \text{ C.L.}). \quad (16.22)$$

Eq. (16.21) then gives

$$g_1 m_1 + g_2 m_2 + g_3 m_3 < 0.7 \text{ eV}. \quad (16.23)$$

On the other hand, Eq. (16.21) can be used in conjunction with inequality (16.18) to obtain a lower bound on the cosmological density in neutrinos. Taking only one massive Majorana flavor,

$$\Omega_\nu h^2 \gtrsim 0.0006. \quad (16.24)$$

Thus neutrinos are definitely a form of dark matter, although perhaps a minor component of it.

The results for the known neutrinos as dark matter can be summarized by the constraints

$$0.05 \text{ eV} < m_1 + m_2 + m_3 < 0.7 \text{ eV}, \quad (16.25)$$

$$0.0006 < \Omega_\nu h^2 < 0.0076, \quad (16.26)$$

where the constraint in Eq. (16.23) has somewhat been relaxed by taking $g_i = 1$.

The upper limit on $\Omega_\nu h^2$ forbids currently known neutrinos from being the major constituents of dark matter. Moreover, since they are light and relativistic at the time of galaxy formation, the three neutrinos known to exist are hot, not cold, dark matter.

The three active neutrinos are our only known particle candidates for non-baryonic dark matter. Since they fail to be cold dark matter, we are lead to consider hypothetical particles.

Type Ib: ‘well-motivated’ candidates

We will discuss three cold dark matter candidates which are ‘well-motivated’, i.e. that have been proposed to solve problems in principle unrelated to dark matter and whose properties can be computed within a well-defined particle physics model. The three candidates we discuss are: (1) a heavy active neutrino with standard model interactions, (2) the neutralino in the minimal supersymmetric standard model, and (3) the axion. Examples of other candidates that can be included in this category are a sterile neutrino (See e.g. Abazajian, Fuller, & Patel (2001)) and other supersymmetric particles such as the gravitino (See e.g. Ellis et al.(1984)) and the sneutrino (see, e.g., Hall, Moroi & Murayama(1998)).

The first two candidates we discuss belong to a general class called weakly interacting massive particles (WIMPs).² WIMPs that were in thermal equilibrium in the early universe (thermal WIMPs) are particularly interesting. Their cosmological density is naturally of the right order of magnitude when their interaction cross section is of the order of a weak cross section. This also makes them detectable in the laboratory, as we will see later. In the early Universe, annihilation reactions that convert WIMPs into standard model particles were initially in equilibrium with their opposite reactions. As the universe expanded, and the temperature became smaller than the WIMP mass, the gas of WIMPs, still in equilibrium, diluted faster than the gas of standard model particles. This occurred because the equilibrium number density of non-relativistic particles is suppressed by a Boltzmann factor $e^{-m/T}$ with respect to the number density of relativistic particles. After a while, WIMPs became so rare that the WIMP annihilation reactions could no longer occur (chemical decoupling), and from then on the number density of WIMPs decreased inversely with volume (or in other words, the number of WIMPs per comoving volume remained constant). Chemical decoupling occurs approximately when the WIMP annihilation rate $\Gamma_{\text{ann}} = \langle \sigma_{\text{ann}} v \rangle n$ became smaller than the universe expansion rate H . Here σ_{ann} is the WIMP annihilation cross section, v is the relative velocity of the annihilating WIMPs, n is the WIMP number density, and the angle brackets denote an average over the WIMP thermal distribution. Using Friedmann's equation to find the expansion rate H gives

$$\Omega h^2 \approx \frac{3 \times 10^{-27} \text{ cm}^3/\text{s}}{\langle \sigma_{\text{ann}} v \rangle} \quad (16.27)$$

for the relic density of a thermal WIMP. An important property of this equation is that smaller annihilation cross sections correspond to larger relic densities (“The weakest wins.”) This can be understood from the fact that WIMPs with stronger interactions remain in chemical equilibrium for a longer time, and hence decouple when the universe is colder, wherefore their density is further suppressed by a smaller Boltzmann factor. Figure 2 illustrates this relationship.

It must be remarked here that in the non-relativistic limit $v \rightarrow 0$, the product $\sigma_{\text{ann}} v$ tends to a constant, because the annihilation cross section σ_{ann} diverges as $1/v$ as $v \rightarrow 0$. This is analogous to what happens for the scattering cross section of thermal neutrons.

Heavy neutrino The WIMP *par excellence* is a heavy neutrino. The example we consider is a thermal Dirac neutrino ν of the fourth generation with Standard Model interactions and no lepton asymmetry. Figure 3 summarizes

²Notice that according to the Merriam-Webster Dictionary of the English Language, a wimp is a weak, cowardly, or ineffectual person.

its relic density as a function of mass. Also shown in the Figure are the current constraints from accelerator experiments and dark matter searches.

A neutrino lighter than ~ 1 MeV decouples while relativistic. If it is so light to be still relativistic today ($m_\nu \lesssim 0.1$ meV), its relic density is $\rho_\nu = 7\pi^2 T_\nu^4 / 120$. If it became non-relativistic after decoupling, its relic density is determined by its equilibrium number density as $\rho_\nu = m_\nu 3\zeta(3) T_\nu^3 / 2\pi^2$. Here $T_\nu = (3/11)^{1/3} T_\gamma$, where $T_\gamma = 2.725 \pm 0.002$ K is the cosmic microwave background temperature. (We use natural units, $c = \hbar = 1$.)

A neutrino heavier than ~ 1 MeV decouples while non-relativistic. Its relic density is determined by its annihilation cross section, as for a general WIMP (see Eq. (16.27)). The shape of the relic density curve in Figure 3 is a reflection of the behavior of the annihilation cross section. The latter is dominated by the Z-boson resonance at $m_\nu \simeq m_Z/2$. This resonant annihilation gives the characteristic V shape to the relic density curve. Above $m_\nu \sim 100$ GeV, new annihilation channels open up, namely the annihilation of two neutrinos into two Z- or W-bosons. The new channels increase the annihilation cross section and thus lower the neutrino relic density. Soon, however, the perturbative expansion of the cross section in powers of the (Yukawa) coupling constant becomes untrustworthy (the question mark in Figure 3). An alternative unitarity argument limits the Dirac neutrino relic density to the dashed curve on the right in the Figure. Neutrinos heavier than 10 TeV ‘overclose’ the universe, i.e. have a relic density that corresponds to a universe which is too young.

The ‘dark matter’ band in Figure 3 indicates where the neutrino is a good dark matter candidate (the band is actually quite generous in light of the most recent measurements of Ωh^2). A thermal Dirac neutrino is a good dark matter candidate when its mass is around few eV, a few GeV or possibly a TeV. For

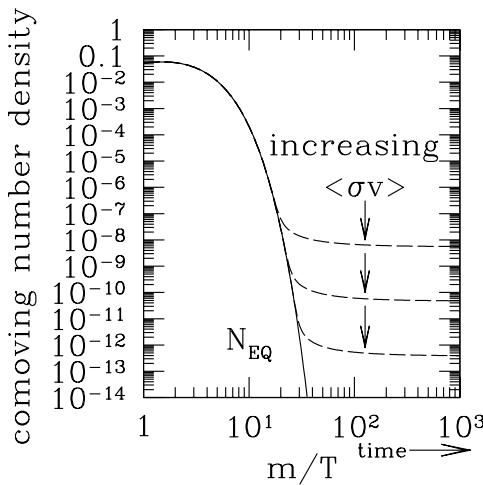
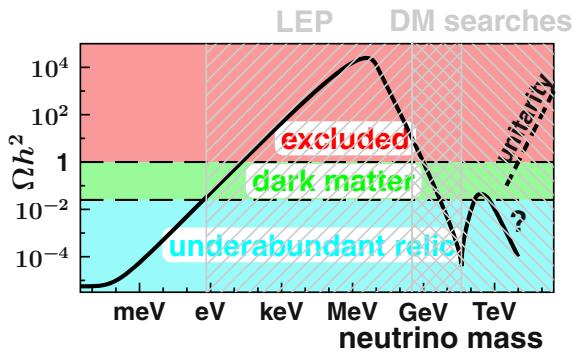


Figure 16.2. Evolution of a typical WIMP number density in the early universe. The number of WIMPs in a volume expanding with the universe (comoving density) first decreases exponentially due the Boltzmann factor $e^{-m/T}$ and then ‘freezes out’ to a constant value when the WIMP annihilation reactions cannot maintain chemical equilibrium between WIMPs and standard model particles. In the figure, $\langle\sigma v\rangle$ is the thermally averaged annihilation cross section times relative velocity. WIMPs with larger annihilation cross section end up with smaller densities.

Figure 16.3. Neutrinos as dark matter. Relic density of a thermal Dirac neutrino with standard-model interactions, together with current constraints from cosmology, accelerators (LEP), and dark matter searches. See text for explanations. (The ‘dark matter’ band is quite generous in light of the WMAP measurements.)



masses smaller than about an eV and between ~ 10 GeV and ~ 100 GeV, it is an underabundant relic from the Big Bang, too dilute to be a major component of the dark matter but nevertheless a cosmological relic. For other masses, it is cosmologically excluded.

Dark matter neutrinos with a mass around 1 eV would be relativistic at the time of galaxy formation (\sim keV), and would thus be part of hot dark matter. From the bounds on hot dark matter in the preceding Section, however, they cannot be a major component of the dark matter in the Universe.

Neutrinos can be cold dark matter if their masses are around few GeV or a TeV. However, fourth-generation heavy neutrinos lighter than 45 GeV are excluded by the measurement of the Z-boson decay width at the Large Electron-Positron collider at CERN. Moreover, direct searches for WIMP dark matter in our galaxy exclude Dirac neutrinos heavier than ~ 0.5 GeV as the dominant component of the galactic dark halo (see Figure 3). Thus although heavy Dirac neutrinos could still be a tiny part of the halo dark matter, they cannot solve the cold dark matter problem.

We need another non-baryonic candidate for cold dark matter.

Neutralino The WIMP *par default* is the lightest neutralino $\tilde{\chi}_1^0$, or sometimes simply χ , which is often the lightest supersymmetric particle in supersymmetric extensions of the Standard Model of particle physics. Supersymmetry is a new symmetry of space-time that has been discovered in the process of unifying the fundamental forces of nature (electroweak, strong, and gravitational). Supersymmetry also helps in stabilizing the masses of fundamental scalar particles in the theory, such as the Higgs boson, a problem, called the hierarchy problem, which basically consists in explaining why gravity is so much weaker than the other forces.

Of importance for cosmology is the fact that supersymmetry requires the existence of a new particle for each particle in the Standard Model. These su-

partners differ by half a unit of spin, and come under the names of sleptons (partners of the leptons), squarks (partners of the quarks), gauginos (partners of the gauge bosons) and higgsinos (partners of the Higgs bosons). Sleptons and squarks have spin 0, and gauginos and higgsinos have spin $\frac{1}{2}$.

If supersymmetry would be an explicit symmetry of nature, superpartners would have the same mass as their corresponding Standard Model particle. However, no Standard Model particle has a superpartner of the same mass. It is therefore assumed that supersymmetry, much as the weak symmetry, is broken. Superpartners can then be much heavier than their normal counterparts, explaining why they have not been detected so far. However, the mechanism of supersymmetry breaking is not completely understood, and in practice it is implemented in the model by a set of supersymmetry-breaking parameters that govern the values of the superpartners masses (the superpartners couplings are fixed by supersymmetry).

The scenario with the minimum number of particles is called the minimal supersymmetric standard model or MSSM. The MSSM has 106 parameters beyond those in the Standard Model: 102 supersymmetry-breaking parameters, 1 complex supersymmetric parameter μ , and 1 complex electroweak symmetry-breaking parameter $\tan \beta$ (see, e.g., the article by Haber in Review of Particle Physics, Hagiwara et al.(2002)). Since it is cumbersome to work with so many parameters, in practice phenomenological studies consider simplified scenarios with a drastically reduced number of parameters. The most studied case (not necessarily the one Nature has chosen) is minimal supergravity, which reduces the number of parameters to five: three real mass parameters at the Grand Unification scale (the scalar mass m_0 , the scalar trilinear coupling A_0 , and the gaugino mass $m_{1/2}$) and two real parameters at the weak scale (the ratio of Higgs expectation values $\tan \beta$ and the sign of the μ parameter). Other scenarios are possible and are considered in the literature. Of relevance to dark matter studies is, for example, a class of models with seven parameters specified at the weak scale: μ , $\tan \beta$, the gaugino mass parameter M_2 , the mass m_A of the CP-odd Higgs boson, the sfermion mass parameter \tilde{m} , the bottom and top quark trilinear couplings A_b and A_t . See the reviews by Jungman, Kamionkowski & Griest (1996) and Bergström (2000) for more details.

It was realized long ago by Goldberg (1983) and Ellis et al.(1984) that the lightest superposition of the neutral gauginos and the neutral higgsinos (which having the same quantum numbers mix together) is an excellent dark matter candidate. It is often the lightest supersymmetric particle, it is stable under the requirement that superpartners are only produced or destroyed in pairs (called R-parity conservation), it is weakly interacting, as dictated by supersymmetry, and it is massive. It is therefore a genuine WIMP, and it is among the most studied of the dark matter candidates. Its name is the lightest neutralino.

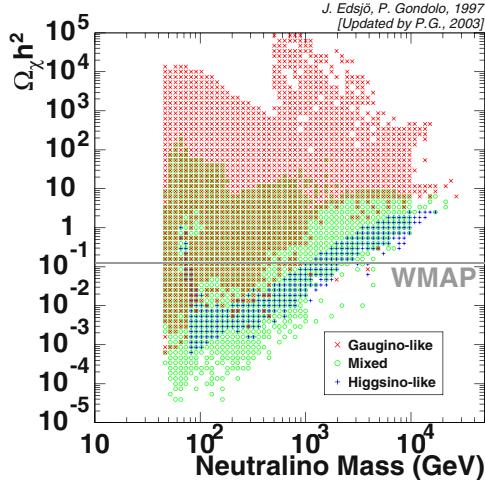


Figure 16.4. Relic density of the lightest neutralino as a function of its mass. For each mass, several density values are possible depending on the other supersymmetric parameters (seven in total in the scenario plotted). The color code shows the neutralino composition (gaugino, higgsino or mixed). The gray horizontal line is the current error band in the WMAP measurement of the cosmological cold dark matter density. (Figure adapted from Edsjö & Gondolo (1997).)

Several calculations exist of the density of the lightest neutralino. An example is given in Figure 4, which reproduces a figure from Edsjö & Gondolo (1997), updated with the WMAP value of the cold dark matter density. This figure was obtained in a scenario with seven supersymmetric parameters at the weak scale. The relic density is not fixed once the neutralino mass is given, as was the case for a Dirac neutrino, because the neutralino annihilation cross section depends on the masses and composition of many other supersymmetric particles, thus ultimately on all supersymmetric parameters. Therefore the density in Figure 4 is not a single-valued function of the neutralino mass, and the plot must be obtained through an extended computer scan in the seven-dimensional parameter space.³ It is clear from Figure 4 that it is possible to choose the values of the supersymmetric parameters in a way that the neutralino relic density satisfies the current determination of Ωh^2 . Although it may seem ridiculous to claim that the neutralino is *naturally* a good dark matter candidate, let us notice that the neutralino relic density in Figure 4 and the neutrino relic density in Figure 3 have a similar range of variation. It is the precision of the cosmological measurements that make us think otherwise.

Reversing the argument, the precision of the cosmological measurements can be used to select the regions of supersymmetric parameter space where the lightest neutralino is cold dark matter. With the current precision of cosmological measurements, these are very thin regions in supersymmetric parameter

³We may ask if there can be points in the empty regions, or in more general terms, what is the meaning of the density of points in Figure 4, and in similar figures in Section 3. For a discussion on this, see Bergström & Gondolo (1996).

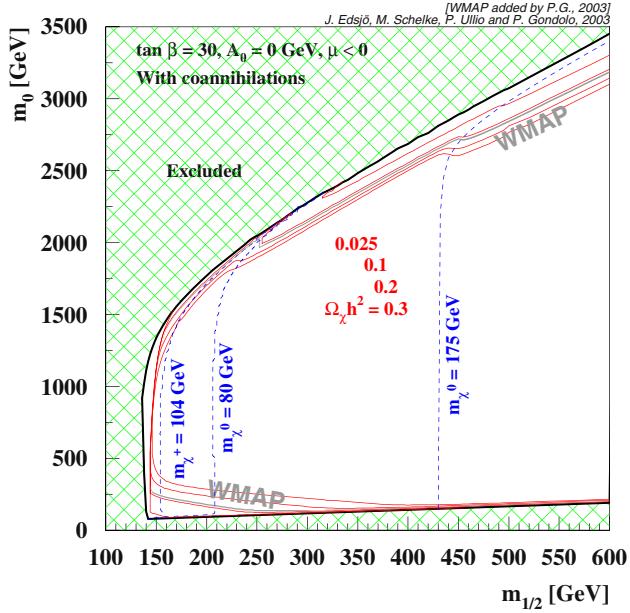


Figure 16.5. Illustration of the power of WMAP constraints on the minimal supergravity parameter space. The figure shows a slice in m_0 and $m_{1/2}$ with $\mu < 0$ at $\tan \beta = 30$ and $A_0 = 0$. The WMAP constraint is a very thin (grey) line that approximately follows the edges of the allowed region. (Figure adapted from Edsjö, Schelke, Ullio & Gondolo (2003).)

space. For this approach to be carried out properly, the theoretical calculation of the neutralino relic density should match the precision of the cosmological data. The latter is currently 7%, as can be gathered from Eq. (16.4), and is expected to improve to about 1% before the end of the decade with the launch of the Planck mission. A calculation of the neutralino relic density good to 1% now exists, and is available in a computer package called DarkSUSY Gondolo et al.(2000); ?. The 1% precision refers to the calculation of the relic density starting from the supersymmetric parameters at the weak scale. The connection with the parameters at the Grand Unification scale, which is vital for minimal supergravity, introduces instead large errors in important regions of parameter space, errors that some authors estimate to be as big as 50% Allanach, Kraml, & Porod (2003).

Given the importance of this calculation, we give now a rapid survey of the ingredients needed to achieve a precision of 1%. Firstly, the equation governing the evolution of the neutralino number density n ,

$$\dot{n} + 3Hn = -\langle \sigma_{\text{ann}} v \rangle (n^2 - n_{\text{eq}}^2), \quad (16.28)$$

has to be solved numerically. The equation being ‘stiff’ (i.e. its difference equation being stable only for unreasonably small stepsizes), a special numerical method should be used. The thermal average of the annihilation cross section $\langle \sigma_{\text{ann}} v \rangle$ at temperature T should be computed relativistically, since the

typical speed of neutralinos at decoupling is of the order of the speed of light, $v \sim c/3$. For this purpose, we can use the expression in Gondolo & Gelmini (1991):

$$\langle \sigma_{\text{ann}} v \rangle = \frac{\int_0^\infty dp p^2 W(p) K_1(\sqrt{s}/T)}{m^4 T [K_2(m/T)]^2}, \quad (16.29)$$

where $W(p)$ is the annihilation rate per unit volume and unit time (a relativistic invariant), $s = 4(m^2 + p^2)$ is the center-of-mass energy squared, and K_1 , K_2 are modified Bessel functions. In passing, let us remark that the common method of expanding in powers of v^2 , $\sigma_{\text{ann}} v = a + bv^2 + \dots$, and then taking the thermal average to give $\langle \sigma_{\text{ann}} v \rangle = a + b\frac{3T}{2m} + \dots$ is unreliable, since it gives rise to negative $\langle \sigma_{\text{ann}} v \rangle$, and thus negative Ωh^2 , near resonances and thresholds. These are nowadays the most important regions of parameter space. Finally, an essential ingredient in the calculation of the neutralino relic density is the inclusion of coannihilation processes. These are processes that deplete the number of neutralinos through a chain of reactions, and occur when another supersymmetric particle is close in mass to the lightest neutralino ($\Delta m \sim T$). In this case, scattering of the neutralino off a particle in the thermal ‘soup’ can convert the neutralino into the other supersymmetric particle close in mass, given that the energy barrier that would otherwise have prevented it (i.e. the mass difference) is easily overcome. The supersymmetric particle participating in the coannihilation may then decay and/or react with other particles and eventually effect the disappearance of neutralinos. We give two examples. Coannihilation with charginos $\tilde{\chi}^\pm$ (partners of the charged gauge and Higgs bosons) may proceed via, for instance,

$$\tilde{\chi}_1^0 e^- \rightarrow \tilde{\chi}_2^- \nu_e, \quad \tilde{\chi}_2^- \rightarrow \tilde{\chi}_2^0 d \bar{u}, \quad \tilde{\chi}_2^0 \tilde{\chi}_1^0 \rightarrow W^+ W^- \quad (16.30)$$

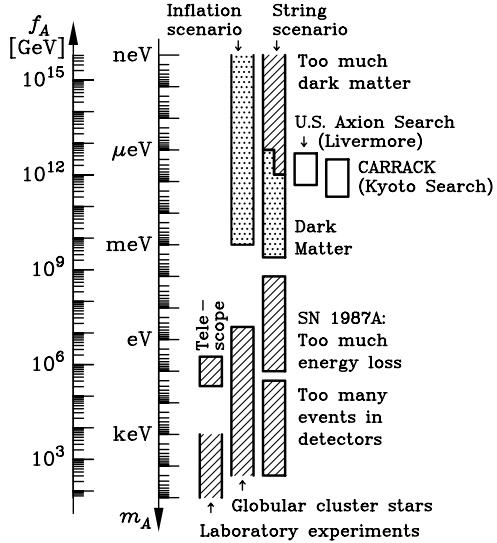
(subscripts on superpartner names indicate particles with different masses). Coannihilation with tau sleptons $\tilde{\tau}$ may instead involve the processes

$$\tilde{\chi}_1^0 \tau \rightarrow \tilde{\tau} \gamma, \quad \tilde{\tau} \tilde{\chi}_1^+ \rightarrow \tau W^+. \quad (16.31)$$

Coannihilations were first included in the study of near-degenerate heavy neutrinos by Binetruy, Girardi & Salati (1984) and were brought to general attention by Griest & Seckel (1991). The current state-of-the-art treatment of neutralino coannihilations, which involves several thousands of processes, is contained in the work by Edsjö & Gondolo (1997) and Edsjö, Schelke, Ullio & Gondolo (2003).

To illustrate the power of the cosmological precision measurements in selecting regions of supersymmetric parameter space, Figure 5 shows the WMAP constraint in one of the figures in Edsjö, Schelke, Ullio & Gondolo (2003). The constraint is as a very thin line that approximately follows the edges of the allowed region in this slice of parameter space (the $m_0 - m_{1/2}$ plane with $\mu < 0$, $\tan \beta = 30$ and $A_0 = 0$).

Figure 16.6. Laboratory, astrophysical, and cosmological constraints on the axion mass m_A . The inflation scenario and the string scenario are referred in the text as the vacuum alignment scenario and the string emission scenario, respectively. f_A is the axion decay constant, which is inversely related to m_A . The axion is a good dark matter candidate for $1 \mu\text{eV} \lesssim m_A \lesssim 1 \text{ meV}$. (Figure from Raffelt in Review of Particle Physics, Hagiwara et al.(2002).)



Cosmological constraints on supersymmetric models are very powerful, and may even serve as a guidance in searching for supersymmetry. This partially justifies the extensive literature on the subject. The neutralino as dark matter is certainly ‘fashionable.’

Axion Our third and last example of a ‘well-motivated’ cold dark matter candidate is the axion.

Axions were suggested by Peccei & Quinn (1977) to solve the so-called “strong CP problem”. Out of the vacuum structure of Quantum Chromodynamics there arises a large CP-violating phase, which is at variance with stringent measurements of the electric dipole moment of the neutron, for example. A possible solution to this problem is that the CP-violating phase is the vacuum expectation value of a new field, the axion, which relaxes dynamically to a very small value. The original axion model of Peccei and Quinn is today experimentally ruled out, but other axion models based on the same idea have been proposed. Among them are the invisible axions of Kim (1979) and Shifman, Vainshtein & Zakharov (1980) (KSVZ axion) and of Dine, Fischle, & Srednicki (1981) and Zhitnitsky (1980) (DFSZ). They differ in the strength of the axion couplings to matter and radiation.

In a cosmological context, axions, contrary to neutrinos and neutralinos, are generally produced non-thermally (although thermal axion production is sometimes considered, as are non-thermal neutrino and neutralino productions). The two main mechanisms for non-thermal axion production are vacuum alignment

and emission from cosmic strings. In the vacuum alignment mechanism, a potential is generated for the axion field at the chiral symmetry breaking, and the axion field, which can in principle be at any point in this potential, starts moving toward the minimum of the potential and then oscillates around it. Quantum-mechanically, the field oscillations correspond to the generation of axion particles. In the other main non-thermal mechanism for axion production, axions are emitted in the wiggling or decay of cosmic strings. In both cases, axions are produced with small momentum, \ll keV, and thus they are cold dark matter despite having tiny masses, between $1 \mu\text{eV}$ and 1 meV . This is in fact the range of masses in which axions are good dark matter candidates. Figure 6 shows the current constraints on the axion mass from laboratory, astrophysical, and cosmological data.

Searches for axions as galactic dark matter rely on the coupling of axions to two photons. An incoming galactic axion can become a photon in the magnetic field in a resonant cavity. For this to happen, the characteristic frequency of the cavity has to match the axion mass. Since the latter is unknown, searches for galactic axions use tunable cavities, and scan over the cavity frequency, a time-consuming process. The U.S. axion search at Livermore is currently exploring a wide range of interesting axion masses, and has put some constraint on the KSVZ axion as a dominant component of the galactic halo Asztalos et al.(2004). Figure 7 shows the constraints on the local galactic density in axions as a function of the axion mass. KSVZ axions with mass in the range $1.91\text{-}3.34 \mu\text{eV}$ cannot be the main component of galactic dark matter. The Livermore search is still continuing to a larger range of axion masses. It is fair to say that axion dark matter is either about to be detected or about to be ruled out.

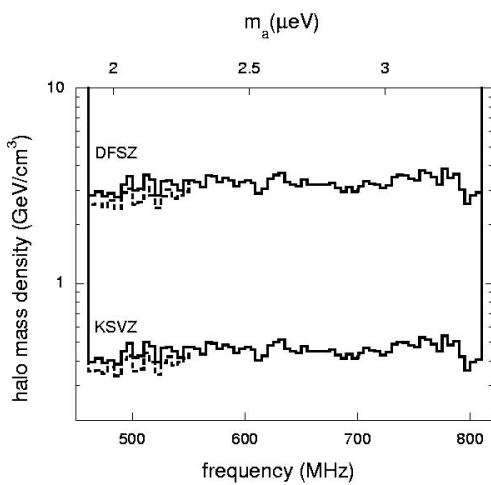


Figure 16.7. Experimental constraints on the density of axions in the galactic halo near the Sun as a function of the axion mass (upper scale) and cavity frequency (lower scale). The regions above the curves marked ‘DFSZ’ and ‘KSVZ’ are excluded from the respective axion models. The currently accepted value for the local dark halo density is $0.45 \text{ GeV}/\text{cm}^3$, which is approximately the extension of the excluded region for the KSVZ axion. (Figure from Asztalos et al.(2004).)

Type II: other candidates

In the Type II category we put all hypothetical cold dark matter candidates that are neither Type Ia nor Type Ib. Some of these candidates have been proposed for no other reason than to solve the dark matter problem. Others are examples of beautiful ideas and clever mechanisms that can provide good possibilities for non-baryonic dark matter, but in some way or another lack the completeness of the theoretical particle physics models of Types Ia and Ib. Although Type II candidates are not studied as deeply as others, it may well be that eventually the question of the nature of cold dark matter might find its answer among them.

Below we present the idea of self-interacting dark matter and gravitationally-produced WIMPZILLAs. Other interesting candidates have been proposed recently in models with extra dimensions, such as Kaluza-Klein dark matter (Cheng, Feng & Matchev (2002)) and branons (Cembranos, Dobado & Maroto (2003)).

There are several ways in which one may be able to come up with an *ad hoc* candidate for non-baryonic cold dark matter. A humorous flowchart on how to do this was put together around 1986 by a group of graduate students at Princeton (Lauer, Statler, Ryden & Weinberg(1986)). The flowchart involves multiple options, and one possibility runs as follows. "A new particle is envisioned which is cooked up just to make everything OK but violates federal law ... still, it doesn't prevent a paper being written by Spergel ..." We have now the proper setting to introduce a candidate suggested by (Spergel & Steinhardt (2000)), self-interacting dark matter.

Self-interacting dark matter The idea behind the introduction of self-interacting dark matter is to find a solution to the cusp and satellite problems of standard cold dark matter scenarios. These two problems are in effect discrepancies between observations and results of simulations of structure formation on the galactic scale. Namely, in the cusp problem, numerical simulations predict a dark matter density profile which increases toward the center of a galaxy like a power law $r^{-\gamma}$ with $\gamma \sim 1$ or higher. This sharp density increase is called a cusp. On the other hand, kinematical and dynamical determinations of the dark matter profile in the central regions of galaxies, especially of low surface brightness galaxies, tend not to show such a sharp increase but rather a constant density core. The observational situation is still rather confused, with some galaxies profiles being compatible with a cusp and others with a core. The theoretical situation is also not fully delineated, with higher resolution simulations showing a dependence of the slope γ on the mass of the galaxy. Although many ideas have been proposed for the resolution of the cusp problem, it is still not completely resolved.

Perhaps connected with the cusp problem is the satellite problem, which is a mismatch between the observed and the simulated numbers of satellites in a galaxy halo. Too many satellites are predicted by the simulations. In reality, observations can detect only the luminous satellites while simulations contain all satellites, including the dark ones. It may be that many dark satellites do not shine, thus solving the satellite problem, but how and which satellites become luminous is not understood yet.

Spergel & Steinhardt (2000) suggested another way to solve both problems. They realized that if dark matter particles would interact with each other with a mean free path of the order of the size of galactic cores, the dark matter interactions would efficiently thermalize the system and avoid the formation of both a central cusp and too many satellites. The requirement on the mean free path λ is roughly $\lambda \sim 10$ kpc. We can figure out the necessary cross section σ for dark matter self-interactions by recalling that the mean free path is related to the cross section and the number density n , or matter density $\rho = mn$, through the relationship $\lambda = 1/(n\sigma) = m/(\rho\sigma)$. Taking a typical $\rho \sim 0.3$ GeV/cm³ gives $\sigma/m \sim 60$ cm²/g. So Spergel & Steinhardt (2000) suggested a new self-interacting dark matter particle with σ/m in the range

$$1 \text{ cm}^2/\text{g} \lesssim \sigma/m \lesssim 100 \text{ cm}^2/\text{g}. \quad (16.32)$$

To understand the magnitude of this number, it is useful to compare it with the geometric cross section of a proton, which is one of the largest known cross sections for elementary particles. We have

$$(\sigma/m)_{\text{proton}} \sim 1 \text{ fm}^2/\text{GeV} \sim 0.006 \text{ cm}^2/\text{g}. \quad (16.33)$$

Thus the desired σ/m seems rather big. It is therefore not surprising that astrophysical constraints on self-interacting dark matter are rather stringent. Gnedin & Ostriker (2001) considered the evaporation of halos inside clusters and set the constraint $\sigma/m < 0.3\text{--}1$ cm²/g. Yoshida, Springel, White & Tormen (2000) considered the shape of cluster cores, which is rounder for self-interacting dark matter than for standard cold dark matter, and concluded that σ/m must be < 10 cm²/g. This bound was later strengthened by Miralda-Escudé(2002) to $\sigma/m < 0.02$ cm²/g. Markevitch et al.(2003) discovered a gas bullet lagging behind dark matter in the merging galaxy cluster 1E0657-56. They combined Chandra X-ray maps of the hot gas in the cluster with weak lensing maps of its mass distribution. From estimates of the column mass densities and of the distance between the gas and the dark matter, Markevitch et al. were able to set the upper limit $\sigma/m < 10$ cm²/g with direct observations of the dark matter distribution. All these bounds leave little room, if any, to self-interacting dark matter a la Spergel & Steinhardt.

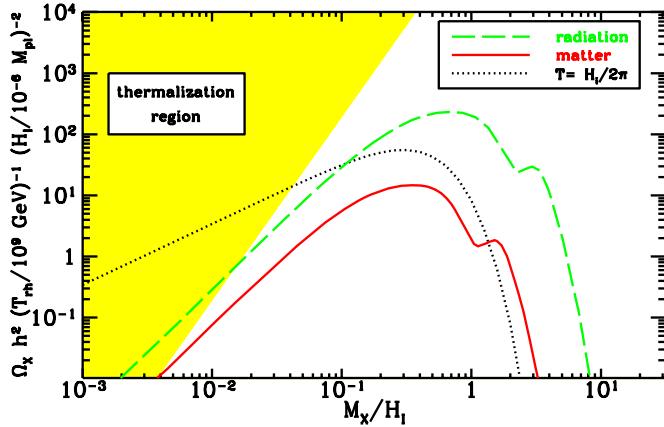


Figure 16.8. Relic density of gravitationally-produced WIMPZILLAs as a function of their mass M_X . H_I is the Hubble parameter at the end of inflation, τ_{rh} is the reheating temperature, and $M_{pl} \approx 3 \times 10^{19}$ GeV is the Planck mass. The dashed and solid lines correspond to inflationary models that smoothly end into a radiation or matter dominated epoch, respectively. The dotted line is a thermal distribution at the temperature indicated. Outside the ‘thermalization region’ WIMPZILLAs cannot reach thermal equilibrium. (Figure from Chung, Kolb & Riotto (1998).)

WIMPZILLAs Our last example of cold dark matter candidates illustrates a fascinating idea for generating matter in the expanding universe: the gravitational creation of matter in an accelerated expansion. This mechanism is analogous to the production of Hawking radiation around a black hole, and of Unruh radiation in an accelerated reference frame.

WIMPZILLAs Chung, Kolb & Riotto (1998); Chung, Kolb & Riotto (1999); Kuzmin & Tkachev (1998) are very massive relics from the Big Bang, which can be the dark matter in the universe if their mass is $\approx 10^{13}$ GeV. They were produced at the end of inflation through a variety of possible mechanisms: gravitationally, during preheating, during reheating, in bubble collisions. It is possible that their relic abundance does not depend on their interaction strength but only on their mass, giving great freedom in their phenomenology. To be the dark matter today, they are assumed to be stable or to have a lifetime of the order of the age of the universe. In the latter case, their decay products may give rise to the highest energy cosmic rays, and solve the problem of cosmic rays beyond the GZK cutoff.

Gravitational production of particles is an important phenomenon that is worth describing here. Consider a scalar field (particle) X of mass M_X in the expanding universe. Let η be the conformal time and $a(\eta)$ the time dependence of the expansion scale factor. Assume for simplicity that the universe is flat.

The scalar field X can be expanded in spatial Fourier modes as

$$X(\vec{x}, \eta) = \int \frac{d^3k}{(2\pi)^{3/2}a(\eta)} \left[a_k h_k(\eta) e^{i\vec{k}\cdot\vec{x}} + a_k^\dagger h_k^*(\eta) e^{-i\vec{k}\cdot\vec{x}} \right]. \quad (16.34)$$

Here a_k and a_k^\dagger are creation and annihilation operators, and $h_k(\eta)$ are mode functions that satisfy (a) the normalization condition $h_k h_k^* - h'_k h_k^* = i$ (a prime indicates a derivative with respect to conformal time), and (b) the mode equation

$$h_k''(\eta) + \omega_k^2(\eta) h_k(\eta) = 0, \quad (16.35)$$

where

$$\omega_k^2(\eta) = k^2 + M_X^2 a^2 + (6\xi - 1) \frac{a''}{a}. \quad (16.36)$$

The parameter ξ is $\xi = 0$ for a minimally-coupled field and $\xi = \frac{1}{6}$ for a conformally-coupled field. The mode equation, Eq. (16.35), is formally the same as the equation of motion of a harmonic oscillator with time-varying frequency $\omega_k(\eta)$. For a given positive-frequency solution $h_k(\eta)$, the vacuum $|0_h\rangle$ of the field X , i.e. the state with no X particles, is defined as the state that satisfies $a_k|0_h\rangle = 0$ for all k . Since Eq. (16.35) is a second order equation and the frequency depends on time, the normalization condition is in general not sufficient to specify the positive-frequency modes uniquely, contrary to the case of constant frequency ω_0 for which $h_k^0(\eta) = e^{-i\omega_0\eta}/(2\omega_0)^{1/2}$. Different boundary conditions for the solutions $h_k(\eta)$ define in general different creation and annihilation operators a_k and a_k^\dagger , and thus in general different vacua.⁴ For example, solutions which satisfy the condition of having only positive-frequencies in the distant past,

$$h(\eta) \sim e^{-i\omega_k^- \eta} \quad \text{for } \eta \rightarrow -\infty, \quad (16.37)$$

contain both positive and negative frequencies in the distant future,

$$h(\eta) \sim \alpha_k e^{-i\omega_k^+ \eta} + \beta_k e^{+i\omega_k^+ \eta} \quad \text{for } \eta \rightarrow +\infty. \quad (16.38)$$

Here $\omega_k^\pm = \lim_{\eta \rightarrow \pm\infty} \omega_k(\eta)$. As a consequence, an initial vacuum state is no longer a vacuum state at later times, i.e. particles are created. The number density of particles is given in terms of the Bogolubov coefficient β_k in Eq. (16.38) by

$$n = \frac{1}{(2\pi a)^3} \int d^3k |\beta_k|^2. \quad (16.39)$$

⁴The precise definition of a vacuum in a curved space-time is still subject to some ambiguities. We refer the interested reader to Fulling (1979); Fulling (1989); Birrell & Davis (1982); Wald (1994) and to the discussion in Chung, Notari & Riotto (2003) and references therein.

These ideas have been applied to gravitational particle creation at the end of inflation by Chung, Kolb & Riotto (1998) and Kuzmin & Tkachev (1998). Particles with masses M_X of the order of the Hubble parameter at the end of inflation, $H_I \approx 10^{-6} M_{Pl} \approx 10^{13}$ GeV, may have been created with a density which today may be comparable to the critical density. Figure 8 shows the relic density Ωh^2 of these WIMPZILLAs as a function of their mass M_X in units of H_I . Curves are shown for inflation models that have a smooth transition to a radiation dominated epoch (dashed line) and a matter dominated epoch (solid line). The third curve (dotted line) shows the thermal particle density at temperature $T = H_I/2\pi$. Also shown in the figure is the region where WIMPZILLAs are thermal relics. It is clear that it is possible for dark matter to be in the form of heavy WIMPZILLAs generated gravitationally at the end of inflation.

3. Neutralino dark matter searches

The second part of these lectures is an introduction to several methods to detect non-baryonic dark matter. We will use the lightest neutralino as our guinea pig, because of the variety of techniques that can be employed to detect it, but the discussion is more general and can be applied to a generic WIMP. Thus in this second part we assume that non-baryonic dark matter is made of WIMPs (in particular, neutralinos), and we examine several observational ways to test our assumption.

Neutralino dark matter searches are traditionally divided into two main categories: (1) direct detection of Galactic dark matter in laboratory experiments, and (2) indirect detection of neutralino annihilation products. For the sake of exposition, indirect searches are further subdivided into: (2a) searches for high-energy neutrinos from the center of the Sun or of the Earth; (2b) searches for anomalous cosmic rays and gamma-rays from galactic halos, especially our own; and (2c) searches for neutrinos, gamma-rays, and radio waves from the Galactic Center. We now examine each of them in turn.

Direct detection

The idea here is that neutralino dark matter is to be found not only in the halo of our galaxy and in our solar system, but also here on Earth and in the room we are in. Thus if we could set up a detector that records the passage of dark matter neutralinos, we could hope of detecting neutralino dark matter.

A process that can be used for this purpose is the elastic scattering of neutralinos off nuclei. Inelastic scattering could also be used in principle, as could scattering off electrons, but the rate of these processes are expected to be (much) smaller.

Dozens of experiments worldwide, too numerous to be all listed here, are using or plan to use elastic scattering to search for neutralino dark matter, or WIMP dark matter in general. The small expected detection rate, and the necessity of suppressing any ionizing radiation passing through the detector, are reasons to shelter these experiments from cosmic rays, e.g. by placing them in mines or underground laboratories.

Generally, with the notable exception of directional detectors described below, only the energy deposited in the detector during the elastic scattering can be measured. This energy is of the order of a few keV, for typical neutralino masses and speeds in the galactic halo. The kinetic energy of the recoiling nucleus is converted partly into scintillation light or ionization energy (giving an electric current) and partly into thermal energy (heating up the detector).

In cryogenic detectors, a simultaneous measurement of both ionization and thermal energy allows the discrimination of nuclear recoils from electrons produced in radioactive decays or otherwise. This discrimination, however, cannot tell if the nuclear recoil was caused by a WIMP or an ambient neutron. The detector, most often a germanium or silicon crystal, needs to be cooled at liquid helium temperature so that its low heat capacity converts a small deposited energy into a large temperature increase. Only relatively small crystals can be currently used in these cryogenic detectors, with relatively low detection rates.

Detection rates can be increased by using bigger detectors operated at room temperature, at the expense of giving up a measurement of the thermal energy and loosing discrimination power against electrons. The biggest dark matter detector is currently of this type. It is a sodium iodide crystal (a scintillator) under the Gran Sasso mountain in Italy, and it belongs to the Italian-Chinese collaboration DAMA (short for DArk MAtter). Interestingly, the loss of discrimination power and the gain in target mass almost compensate each other, and the sensitivity of cryogenic and scintillation detectors is not very different.

Annual modulation Few years ago, the DAMA collaboration reported a possible detection of WIMP dark matter Belli (1997); Bernabei et al.(1998); Bernabei et al.(1999); Bernabei et al.(2000); Bernabei et al.(2003). Their most recent data Bernabei et al.(2003) span 7 years and show a 6.3σ modulation in their total counting rate (signal+background) with a period of 1 year and an amplitude of ~ 0.02 events/day per kg of detector and keV of visible recoil energy (Figure 9). This kind of yearly modulation in a WIMP signal was predicted by Drukier, Freese & Spergel(1986) and Freese, Frieman & Gould (1988) on the basis that the velocity of the Earth around the Sun adds vectorially to the velocity of the Sun in the Galaxy to produce a yearly modulation in the average speed of the WIMPs relative to the Earth (the WIMPs are assumed on average at rest in the Galaxy). For an observer on the Earth, the WIMP ‘wind’ arrives at a higher speed when the Earth and the Sun move in the same

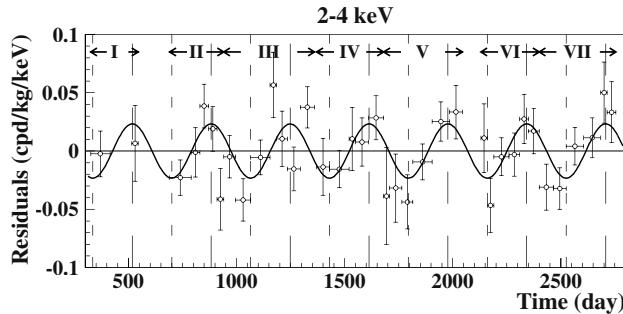
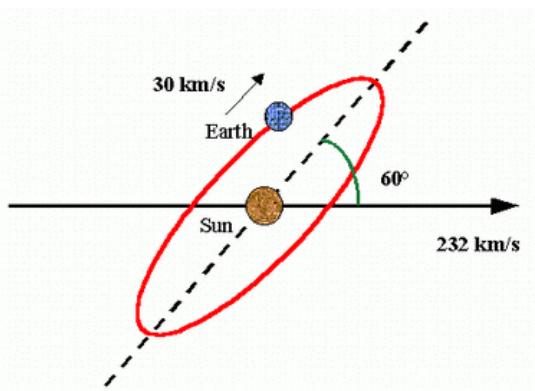


Figure 16.9. Annual modulation of the total counting rate (background plus possible dark matter signal) in seven years of data with the DAMA- NaI detector. A constant counting rate has been subtracted to give the ‘residuals.’ The significance of the modulation is 6σ and its period is 1 year. The interpretation of the yearly modulation as due to a WIMP signal is controversial. (Figure from Bernabei et al.(2003).)

direction and at a lower speed when they move in opposite directions. (The two velocities are actually misaligned by $\sim 60^\circ$; see Figure 10.) The WIMP flux, and the WIMP detection rate, both proportional to the relative speed of the Earth and the WIMPs, are similarly modulated.

While the presence of a yearly modulation in the DAMA data seems now to be established, its interpretation as due to WIMPs is controversial. Firstly, it is not hard to imagine that the background itself can undergo seasonal variations with a period of a year. The DAMA collaboration has examined many possible sources of background variation, including a yearly modulation of the cosmic ray intensity underground due to winter-summer temperature changes in the upper atmosphere. They claim to have found no annual variation in the

Figure 16.10. Sketch illustrating the directions of the Sun’s and the Earth’s motions during a year. As the Sun moves in the Galaxy (here at 232 km/s, 60° out of the plane of the Earth’s orbit), the Earth moves around the Sun (here at 30 km/s). The vectorial sum of their velocities gives the velocity of the Earth with respect to the Galaxy. Assuming the WIMPs to be on average at rest in the Galaxy, it follows that the average speed of the WIMPs relative to the Earth is modulated with a period of 1 year.



background level that would produce an amplitude as big as the observed one. Secondly, the EDELWEISS cryogenic detector has sensitivity comparable to DAMA's but has not recorded any nuclear recoil event Benoît et al.(2002). And the CDMS-I cryogenic detector, also of comparable sensitivity, has detected nuclear recoil events attributed to ambient neutrons in the shallow site where the detector was running Akerib et al.(2003). Comparison of these experimental results is however not as straightforward as it may seem, because the relationship between detection rates in cryogenic and scintillation detectors depends, among other things, on the kind of WIMP-nucleus interaction and on details of the WIMP velocity distribution in the halo, which are both poorly known.

This dependence is apparent in the expression for the expected counting rate per recoil energy bin and unit detector mass dR/dE . We have

$$\frac{dR}{dE} = \int \frac{N_T}{M_T} \times \frac{d\sigma}{dE} \times nvf(\vec{v}, t) d^3v, \quad (16.40)$$

where N_T and M_T are the number of target nuclei and the detector mass, respectively, $d\sigma/dE$ is the WIMP-nucleus differential cross section, and $nvf(\vec{v}, t)$ is the WIMP flux impinging on the detector. Here n denotes the WIMP density, v the WIMP speed, and $f(\vec{v}, t)d^3v$ the WIMP velocity distribution. We write $M_T/N_T = M$, the nuclear mass, $n = \rho/m$, and $d\sigma/dE = \sigma_0|F(q)|^2/E_{\max}$, where σ_0 is the total scattering cross section of a WIMP off a fictitious point-like nucleus, $|F(q)|^2$ is a nuclear form factor that depends on the momentum transfer $q = \sqrt{2ME}$ and is normalized as $F(0) = 1$, and $E_{\max} = 2\mu^2v^2/M$ is the maximum recoil energy imparted by a WIMP of speed v ($\mu = mM/(m + M)$ is the WIMP-nucleus reduced mass). Hence

$$\frac{dR}{dE} = \frac{\rho\sigma_0|F(q)|^2}{2m\mu^2} \int_{v > \sqrt{ME/2\mu^2}} \frac{f(\vec{v}, t)}{v} d^3v. \quad (16.41)$$

Notice that one can only measure the product $\rho\sigma_0$ with this technique. Notice also that the event rate at energy E depends on the WIMP velocity distribution at speeds $v > \sqrt{ME/2\mu^2}$. This integration limit depends on the nuclear mass, and thus detectors with different kinds of nuclei are sensitive to different regions of the WIMP velocity space. Moreover, the cross section σ_0 scales differently for spin-dependent and spin-independent WIMP-nucleus interactions. Finally, while there is a consensus on the spin-independent nuclear form factors, spin-dependent form factors are sensitive to detailed modeling of the proton and neutron wave functions inside the nucleus ?, see and references therein]Jungman:1996.

For spin-independent interactions with a nucleus with Z protons and $A - Z$ neutrons, one has

$$\sigma_0 = \frac{\mu^2}{\pi} |ZG_s^p + (A - Z)G_s^n|^2 \simeq A^2 \frac{\mu^2}{\pi} |G_s^p|^2, \quad (16.42)$$

where G_s^p and G_s^n are the scalar four-fermion couplings of the WIMP with point-like protons and neutrons, respectively ?, see, e.g., lesarcs. The last approximation holds for the case $G_s^p \simeq G_s^n$, which is typical of a neutralino. The spin-independent event rate, proportional to σ_0/μ^2 , scales with the square of the atomic number A (if we neglect the form factor). It is this dependence on A that allows detectors with relatively heavy nuclei to reach down to WIMP-proton cross sections typical of weak interactions.

For spin-dependent interactions, one has instead

$$\sigma_0 = \frac{4\mu^2}{\pi} \frac{J+1}{J} |\langle S_p \rangle G_a^p + \langle S_n \rangle G_a^n|^2, \quad (16.43)$$

where J is the nuclear spin, $\langle S_p \rangle$ and $\langle S_n \rangle$ are the expectation values of the spin of the protons and neutrons in the nucleus, respectively, and G_a^p and G_a^n are the axial four-fermion couplings of the WIMP with point-like protons and neutrons ?, see, lesarcs, tovey. There is no increase of the spin-dependent rate with A^2 , and spin-dependent cross sections of the order of weak cross sections are hard to reach with current detector technology.

Given all these ambiguities in the comparison of cryogenic and scintillator results, it is the author's opinion that the important issue if WIMP dark matter has been detected is not settled yet. A bigger DAMA detector and an upgraded CDMS detector running in the low-background Soudan mine are currently taking data. EDELWEISS is also improving their sensitivity, and new experiments, like CRESST-II and ZEPLIN-IV, should start taking data shortly.

Current bounds and future reach The sensitivity of some future experiments is shown in Figure 11, together with the current best bounds from the cryogenic detectors CDMS-I Akerib et al.(2003) and EDELWEISS Benoît et al.(2002), and the region where DAMA claims evidence for a WIMP signal Bernabei et al.(2003). As it is conventional in comparing results from different experiments, the figure shows the WIMP-proton spin-independent cross section obtained from experimental data using Eq. (16.42) under the assumption of a Maxwellian distribution with conventional parameters for the WIMP velocity. For an historical perspective, the figure also displays the first observational bound on WIMP dark matter obtained by Ahlen et al.(1987).

We also indicate theoretical predictions for a Dirac neutrino with standard model couplings, and for the lightest neutralino in two supersymmetric scenarios: minimal supergravity as in Ellis, Ferstl & Olive (2000) (shaded yellow

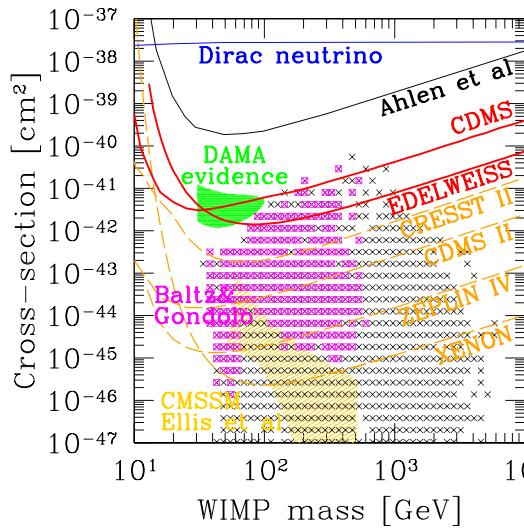


Figure 16.11. Current best bounds (lines labeled ‘CDMS’ and ‘EDELWEISS’) on the dark matter WIMP-proton spin-independent cross section as a function of WIMP mass, together with the expected reach of a few of the upcoming experiments (‘CRESST-II’, ‘CDMS-II’, ‘ZEPLIN-IV’, ‘XENON’), some theoretical expectations (‘Dirac neutrino’, ‘Baltz & Gondolo’, ‘CMSSM’), and the first bound on WIMP dark matter (‘Ahlen et al.’).

See text for details.

region) and weak-scale MSSM as in Baltz & Gondolo (2001); Baltz & Gondolo (2003) (black crosses and magenta squares). Theory models assume that the respective dark matter particles fill up the galactic halo. Dirac neutrinos are excluded as main constituents of galactic dark matter in the mass range 3 GeV – 3 PeV (the bounds continue linearly to the right of the figure). These are the bounds from dark matter searches used in Section 2.2 (and Figure 3) to conclude that a yet-undetected particle species is needed to provide cold dark matter. For the neutralino, the expected scattering cross section varies in a wide range. Even with the most restrictive assumptions of the Constrained Minimal Supersymmetric Standard Model (CMSSM) in Ellis, Ferstl & Olive (2000), the cross section at a given neutralino mass can change by an order of magnitude when the other supersymmetric parameters are changed (shaded yellow region in Figure 11). For neutralino masses $m_\chi \sim 400$ GeV, the CMSSM model parameters may conspire to make the spin-independent cross section arbitrarily small (if such is the case, the total cross section will be dominated by spin-dependent terms, which are however much smaller for the heavy nuclei in current detectors). The highest CMSSM cross sections are within reach of the upcoming experiments, although outside the current best bounds. Other studies of supergravity models ?, e.g.]Feng:2000 are less restrictive and result in somewhat larger cross sections. Enlarging the parameter space beyond supergravity, as for example in Baltz & Gondolo (2001); Baltz & Gondolo (2003), opens up more possibilities for the values of the cross section. This happens basically for two main reasons; (i) The supergravity relation between the masses of squark and sleptons on one side and Higgs bosons on the other side is removed; the Higgs boson can thus be lighter, and the spin-dependent scattering cross sec-

tion, which varies essentially with the fourth inverse power of the Higgs mass, can be larger; (ii) It is no longer required that the electroweak symmetry breaking is achieved radiatively, with the consequence that the gaugino and higgsino content of the lightest neutralino can be arbitrary; the scattering cross section is then enhanced because mixed neutralinos couple to nucleons stronger than pure gauginos or pure higgsinos. This explains the larger extent of the region covered by the Baltz & Gondolo (2001); Baltz & Gondolo (2003) models in Figure 11. Finally, the region in Figure 11 marked by the magenta squares indicates the subset of the supersymmetric models examined by Baltz & Gondolo (2001); Baltz & Gondolo (2003) that could be able to quantitatively explain the $\sim 3\sigma$ deviation between the measured value and the Standard Model value of the magnetic moment of the muon.⁵ These models have relatively light masses for supersymmetric partners, and give neutralino-proton cross sections which are relatively large and within the reach of the most ambitious future experiments.

Directional detection We conclude this section by mentioning the very intriguing possibility of WIMP detectors that are sensitive to the direction of nuclear recoils (directional detectors).

The advantages of measuring the recoil direction are multiple: a more powerful background discrimination; the detection of the new modulation effects, such as a daily modulation in the arrival direction of WIMPs due to the Earth rotation around its axis; and the exciting possibility of reconstructing the WIMP velocity distribution in the solar neighborhood. The latter is possible because of a simple relation between the WIMP velocity distribution $f(\vec{v}, t)$ and the nuclear recoil rate differential in both energy E and recoil direction \hat{q} Gondolo (2002c),

$$\frac{dR}{dEd\Omega} = \frac{n\sigma_0|F(q)|^2}{4\pi\mu^2} \hat{f}(w, \hat{q}) \quad (16.44)$$

where $d\Omega$ is an infinitesimal solid angle around the direction \hat{q} , $w = \sqrt{ME/2\mu^2}$, and

$$\hat{f}(w, \hat{q}) = \int \delta(\vec{u} \cdot \hat{q} - w) f(\vec{v}, t) d^3v \quad (16.45)$$

is the Radon transform of the WIMP velocity distribution.

A promising development in this direction is the DRIFT detector Snowden-Ifft, Martoff & Burwell (2000). This detector consists of a negative ion time projection chamber, the gas in the chamber serving both as WIMP target and as ionization medium for observing the nuclear recoil tracks. The direction of

⁵The size of the deviation has been hard to determine conclusively because of the difficulty of the non-perturbative QCD calculations involved, in particular their dependence on the data used as input. There is also a story on sign errors in some of the theoretical calculations....

the nuclear recoil is obtained from the geometry and timing of the image of the recoil track on the chamber end-plates. A 1 m³ prototype has been successfully tested, and a 10 m³ detector is under consideration.

Directional detection is particularly powerful for detecting structure in the dark matter velocity space, as discussed in the next section on the Sagittarius stream.

Sagittarius stream Recent observations of the stellar component of the Galactic halo show evidence of a merger history that has not yet become well mixed, and corroborate previous indications that halos form hierarchically. In particular, the Sloan Digital Sky Survey Newberg et al.(2003) and the Two Micron All Sky Survey Majewski, Skrutskie, Weinberg & Ostheimer(2003) have traced the tidal stream Ibata et al.(2001) of the Sagittarius (Sgr) dwarf spheroidal galaxy. The Sagittarius dwarf spheroidal galaxy, of roughly $10^9 M_{\odot}$, is a satellite of our own much larger Milky Way Galaxy, located inside the Milky Way, ~ 12 kpc behind the Galactic Center and ~ 12 kpc below the Galactic Plane Ibata et al.(1997). Two streams of matter are being tidally pulled away from the main body of the Sgr galaxy and extend outward from it. These streams, known as the leading and trailing tidal tails, are made of matter tidally pulled away from the Sgr galaxy. It appears that the leading tail is showering matter down upon the solar neighborhood Majewski, Skrutskie, Weinberg & Ostheimer(2003). The flow is in the general direction orthogonal to the Galactic plane and has a speed of roughly 300 km/s. This speed is comparable to that of the relative speed of the Sun and the WIMPs in the general dark halo.

It is natural to expect that dark matter is associated with the detected tidal streams. Hence one can hope to detect the stream in direct detection experiments. The detectability depends on the density of dark matter in the stream. The mass-to-light ratio M/L in the stream is unknown, but is plausibly at least as large as that in the Sgr main body; in fact, the M/L in the stream may be significantly larger because the dark matter on the outskirts of the main body would be tidally stripped before the (more centrally located) stars. Various determinations of the M/L for the Sgr main body give values in the range 25 to 100 (see the discussion in Majewski, Skrutskie, Weinberg & Ostheimer(2003) and references therein). Freese, Gondolo & Newberg (2003) and Freese, Gondolo, Newberg & Lewis(2003) have estimated the density of dark matter in the stream, and find it to be in the range 0.3% to 23% of the local (smoothed) dark halo density. This agrees with a previous theoretical study on the tidal disruption of satellite galaxies falling into the halo of our own Milky Way by Stiff, Widrow & Frieman (2001). These authors found that, with probability of order 1, the Sun should be situated within a stream of density $\sim 4\%$ of the local Galactic halo density.

The additional flux of WIMPs from the stream shows up as a 0.3–23% increase in the rate of nuclear recoils at energies below a characteristic energy E_c , the highest energy that WIMPs in the stream can impart to a target nucleus. Hence, there is a step in the energy recoil spectrum; the count rate in the detector is enhanced at low energies, but then returns to the normal value (due to Galactic halo WIMPs) at all energies above the critical energy E_c . This feature can be observed as a sharp decrease in the count rate above a characteristic energy that depends on the mass of the target nucleus, the mass of the WIMP, and the speed of the stream relative to the detector. Figures 12(a) and (b) show how the recoil spectrum dR/dE is modified by the presence of the stream for a sodium iodide detector (like in DAMA) and a germanium detector (like for CDMS and EDELWEISS). For the sake of illustration, the plots assume a stream density equal to 20% of the local halo density.

Excitingly, the effect of the stream should be detectable in DAMA, CDMS-II, and other upcoming detectors, and may already be present in the current DAMA data. A detail calculation Freese, Gondolo & Newberg (2003) predicts the presence of stream WIMPs in the data with a significance of 100σ for DAMA and 11σ for CDMS if the stream density is 20% of the local halo density, and a significance of 24σ for DAMA and 3σ for CDMS if the stream

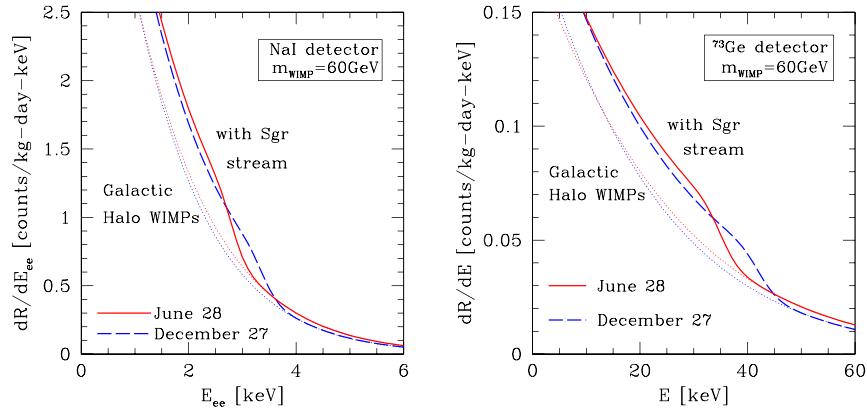


Figure 16.12. Effect of the presence of WIMPs in the Sagittarius leading tidal arm. Count rate of 60 GeV WIMPs in (a) an NaI detector such as DAMA and (b) a Ge detector such as CDMS and EDELWEISS, as a function of recoil energy. The dotted lines (towards the left) indicate the count rate due to Galactic halo WIMPs alone for an isothermal halo. The solid and dashed lines indicate the step in the count rate that arises if we include the WIMPs in the Sgr stream for $v_{str} = 300$ km/s in the direction $(l, b) = (90^\circ, -76^\circ)$ with a stream velocity dispersion of 20 km/sec. The plot assumes that the Sgr stream contributes an additional 20% of the local Galactic halo density. The solid and dashed lines are for June 28 and December 27 respectively, the dates at which the annual modulation of the stream is maximized and minimized. (Figure from Freese, Gondolo & Newberg (2003).)

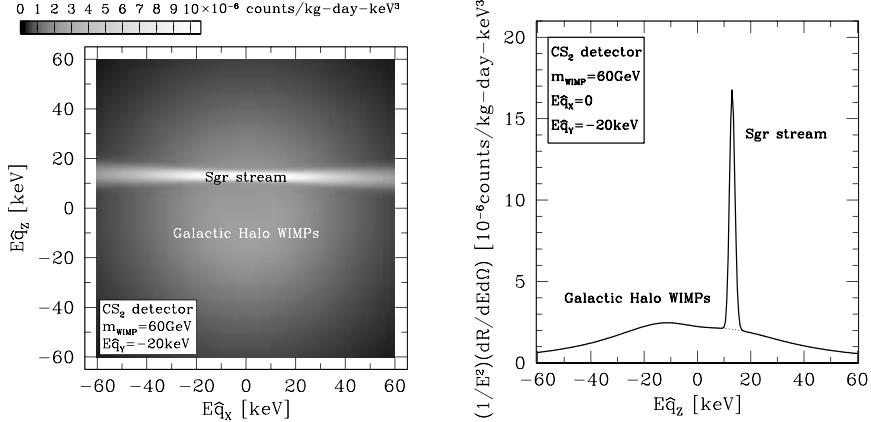


Figure 16.13. Count rate of 60 GeV WIMPs in a CS₂ detector (DRIFT) as a function of recoil energy E and direction of the nuclear recoil ($\hat{q}_x, \hat{q}_y, \hat{q}_z$). Here X points toward the galactic center, Y toward the direction of galactic rotation, and Z toward the North Galactic Pole. On the left is a density plot of the count rate in the 2-dimensional slice $E\hat{q}_y = -20$ keV. On the right is the 1-dimensional section through $E\hat{q}_y = -20$ keV and $\hat{q}_x = 0$. The horizontal axis represents recoils in the direction of the Galactic center (left) and Galactic anticenter (right); the vertical axis represents recoils in the direction of the North Galactic Pole (upward) and South Galactic Pole (downward). The gray scale indicates the count rate per kilogram of detector per day and per unit cell in the 3-dimensional energy space $E\hat{q}$. Lighter regions correspond to higher count rates. The white band in the upper part is the location of nuclear recoils due to WIMPs in the Sgr stream. The fuzzy gray cloud at the center contains recoils due to WIMPs in the local isothermal Galactic halo. The two WIMP populations can in principle be easily separated, given a sufficiently long exposure. (Figure from Freese, Gondolo & Newberg (2003).)

density is 4% of the local halo density. (These significance figures are however very sensitive to the velocity assumed for the stream, cfr. Freese, Gondolo, Newberg & Lewis(2003).)

Directional detection will be a fantastic means of recognizing the presence of a dark stream through the Solar system. The recoil distribution due to WIMPs in a stream is very much different from the recoil distribution due a Maxwellian velocity distribution. The corresponding Radon transforms that appear in Eq. (16.44) are: for a stream of velocity \vec{V} ,

$$\hat{f}(w, \hat{q}) \simeq \delta(w - \vec{V} \cdot \hat{q}) \quad (16.46)$$

which is non-zero only on the surface of a sphere in $(w\hat{q}_x, w\hat{q}_y, w\hat{q}_z)$ space; for a Maxwellian of bulk velocity \vec{V} and velocity dispersion σ ,

$$\hat{f}(w, \hat{q}) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left[-\frac{(w - \vec{V} \cdot \hat{q})^2}{2\sigma^2}\right], \quad (16.47)$$

which is a smooth gaussian distribution. For our Sgr stream, consider a next-generation DRIFT detector of 30 m^3 (DRIFT-2). The difference in the recoil direction distributions of stream and isothermal WIMPs is apparent in Figure 13, where we plot the differential detection rate $E^{-2}dR/dEd\Omega$ for DRIFT-2 under the assumption of a 20% stream density. As seen in the figure, a large DRIFT detector will have the capability of clearly identifying WIMPs in the Sgr stream.

Indirect detection

Besides the direct detection of galactic neutralino dark matter in the laboratory, we can search for dark matter neutralinos by looking for the products of their annihilation. We distinguish three types of searches according to the place where neutralino annihilations occur. The first is the case of neutralino annihilation in the Sun or the Earth, which gives rise to a signal in high-energy neutrinos; the second is the case of neutralino annihilation in the galactic halo, or in the halo of external galaxies, which generates gamma-rays and other cosmic rays such as positrons and antiprotons; the third is the case of neutralino annihilations around black holes, in particular around the black hole at our Galactic Center.

All these annihilation signals share the property of being proportional to the *square* of the neutralino density. This follows from the fact that the neutralino is a Majorana fermion, i.e. is identical to its antiparticle. Two neutralinos can annihilate to produce standard model particles. Simple stoichiometry then tells us that the annihilation rate, being proportional to the product of the densities of the initial particles, is proportional to the square of the neutralino density. In more detail, we have

$$\Gamma_{\text{ann}} = \frac{\sigma_{\text{ann}} v \rho^2}{m^2}, \quad (16.48)$$

where Γ_{ann} is the neutralino annihilation rate per unit volume (i.e. the number of neutralinos that are annihilated per unit volume and unit time), σ_{ann} is the neutralino-neutralino annihilation cross section, v is the relative speed of the two annihilating neutralinos, ρ is the neutralino mass density, and m is the neutralino mass. Recall that the annihilation cross section σ_{ann} goes as $1/v$ at small speeds, as required by kinematical arguments, and thus the product $\sigma_{\text{ann}}v$ does not vanish linearly with v (and is not small at the relatively small speeds of neutralinos in galactic halos). Notice also that the number of annihilations per unit volume and unit time is given by $\frac{1}{2}\Gamma_{\text{ann}}$, where the factor of $\frac{1}{2}$ correctly converts between the number of annihilation events and the number of neutralinos that are annihilated (2 per annihilation). It is easy to get confused with this factor of $\frac{1}{2}$.

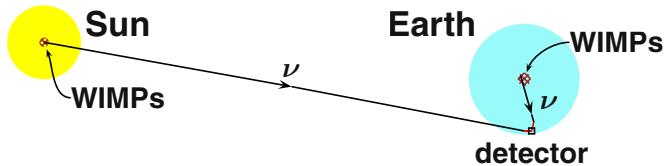
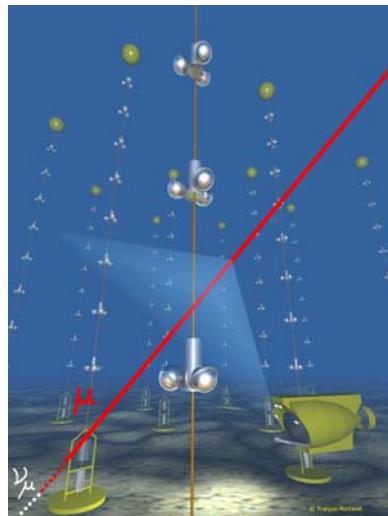


Figure 16.14. Illustration of indirect detection of WIMPs using high-energy neutrinos emitted in WIMP annihilations in the core of the Sun or of the Earth.

If in the annihilation rate Γ_{ann} we insert a typical weak interaction cross section and a typical value for the average dark matter density in the Universe, the annihilation rate we obtain gives undetectably small signals. Indirect detection is possible because dark matter is not distributed uniformly in space. Galaxies and clusters of galaxies are overdensities in the dark matter field, as is any possible substructure in galactic halos. Furthermore, dark matter may be concentrated gravitationally around massive objects, and may even get trapped inside planets and stars. The neutralino annihilation rate, proportional to the square of the neutralino density, increases substantially in these dark matter concentrations, sometimes to the point of giving observable signals.

Figure 16.15. Principle of operation of a water Cherenkov neutrino telescope. An incoming muon neutrino (here a muon neutrino) is converted into a charged lepton (a muon) in the material surrounding the detector (in the rock at the bottom of the sea). The charged lepton moves faster than light in water and thus Cherenkov light (blue cone) is emitted along its trajectory. The Cherenkov light is collected by the array of photomultipliers suspended on strings. Cherenkov neutrino telescopes in ice work on the same principle. (Background figure by François Montanet, ANTARES Collaboration; tracks and Cherenkov light by the present author.)



High energy neutrinos from the core of the Sun or of the Earth. Neutralinos floating around the solar system can occasionally collide with nuclei in the Sun and in the planets (the Earth, in particular). In these collisions, they may lose enough kinetic energy to end up with a speed smaller than the escape speed, thus becoming gravitationally trapped. After some time, the trapped neutralinos will sink to the core of the celestial body in which they are captured, and will possibly reach a condition of thermal equilibrium (Figure 14).

Once concentrated in the center, neutralinos annihilate copiously. The annihilation rate is maximal when all captured neutralinos annihilate (a condition called equilibrium between capture and annihilation). Whether this condition is satisfied depends on the relative strength of the annihilation and scattering cross sections, and ultimately on the parameters of the particle and halo models. (See Jungman, Kamionkowski & Griest (1996) and references therein for complete formulas.)

Of the annihilation products produced in the center of the Earth and the Sun only the neutrinos make it to the surface; all the other products are absorbed or decay within a short distance of production. All three flavors of neutrinos are produced for neutralino masses which are currently allowed. Direct production of a neutrino pair is however strongly suppressed in neutralino annihilation, due to the Majorana nature of the neutralino. Annihilation neutrinos are instead produced as secondaries in the decay chains of the primary particles produced in the neutralino-neutralino annihilation. As a consequence, the neutrino energy spectrum is a continuum, and the typical energy of neutrinos from neutralino annihilations is about a tenth of the neutralino mass. Given the current constraints, this means a neutrino energy between few GeVs and few TeVs.

Neutrinos of this energy can be detected in Cherenkov neutrino telescopes, whose principle of operation is depicted in Figure 15. A charged-current interaction in the material surrounding the detector (rock, ice, water) converts the neutrino into its corresponding charged lepton, which then radiates Cherenkov light in the detector medium (ice or water). Several neutrino telescopes are currently operational (among them the Super-Kamiokande detector in Japan and the AMANDA detector at the South Pole), and others are under construction or development (IceCube at the South Pole, ANTARES and NESTOR in the Mediterranean). Other neutrino telescopes have played a role in dark matter searches in the past, such as the IMB, the Fréjus, the MACRO, and the Baksan experiments.

The current experimental situation for this indirect detection method is summarized in Figure 16(a) for neutrinos from WIMPs in the Sun, and Figure 16(b) for neutrinos from WIMPs in the Earth. The figures show the current best bounds from the MACRO, Baksan, Super-Kamiokande, and AMANDA

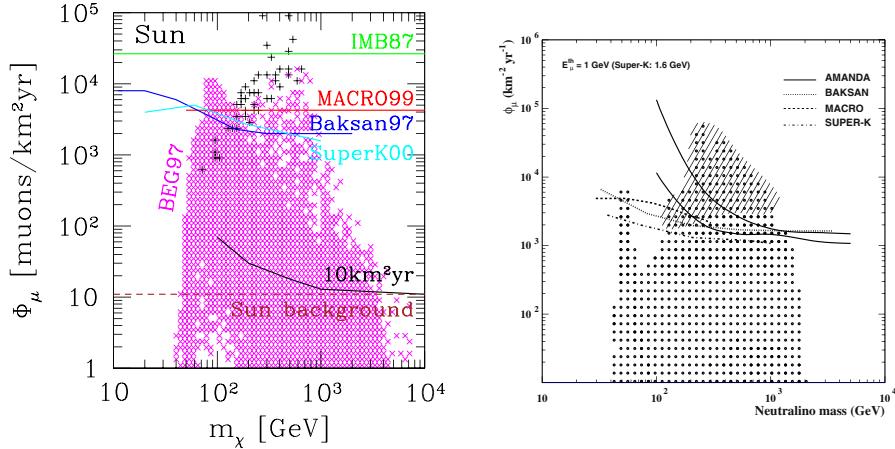


Figure 16.16. Indirect searches for neutralino dark matter using high-energy neutrinos from (left) the Sun and (right) the Earth. On the vertical axis is the flux of neutrino-induced muons that traverse the neutrino telescope, on the horizontal axis is the neutralino mass. ‘IMB87’ is the historically first upper limit, ‘MACRO,’ ‘BAKSAN,’ ‘SuperK,’ and ‘AMANDA’ are the limits from the corresponding experiments. The regions marked by \times on the left and by dots on the right are the predictions of supersymmetric models defined at the weak scale Bergström, Edsjö, & Gondolo (1998); Ahrens et al.(2002). The $+$ signs on the left and the shaded region on the right indicate the regions where there are models that have been excluded by direct dark matter searches. The line labeled ‘ $10\text{km}^2\text{yr}$ ’ shows the maximum reach of such an exposure in IceCube, and the line labeled ‘Sun background’ marks the level of high-energy neutrino emission due to cosmic ray interactions on the surface of the Sun, which is the ultimate applicability limit of this method. (Figure on the left from Gondolo (2000); figure on the right from Ahrens et al.(2002).)

experiments, as well as the first bound obtained using this technique by the IMB collaboration in 1987. Also shown is the reach of the IceCube experiment after an exposure of $10 \text{ km}^2 \text{ yr}$, and the ultimate applicability of this method for the Sun, which is set by the emission of high-energy neutrinos in cosmic ray interactions with nuclei on the surface of the Sun.

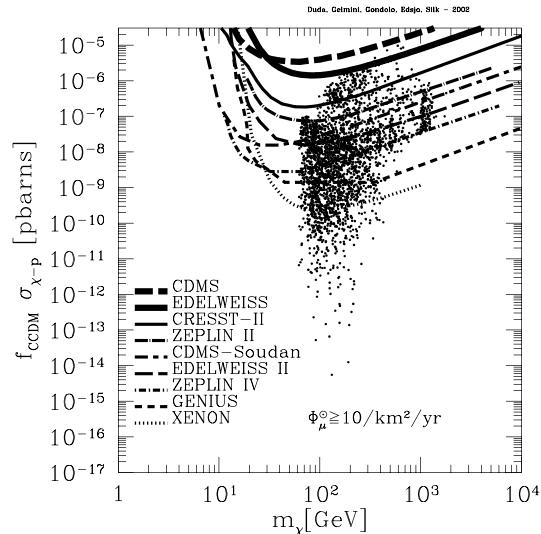
Expectations from theoretical models in Figure 16 (MSSM with seven weak-scale parameters) range from cases which are already excluded by this method to cases which this indirect method will be unable to explore. In comparison, direct searches have a different coverage of theoretical models. The reach of direct searches is indicated in the Figure 16 by $+$ signs in the left panel and by the shading in the right panel (direct searches exclude only some of the theoretical models that are projected onto these regions from the higher-dimensional supersymmetric parameter space). There are models that can be explored by direct searches and not by indirect searches of high-energy neutrinos from the Sun and the Earth. And vice versa, there are theoretical models that cannot be explored by direct searches but can be explored by indirect searches of high-

energy neutrinos from the Sun or the Earth. The latter aspect is illustrated in Figure 17, where models that can be reached by indirect searches for neutrinos from neutralinos in the Sun are marked by dots in the scattering cross section–mass plane, and compared with the sensitivity of several current and future direct search experiments. Several models fall beyond the reach of even the most ambitious direct dark matter searches. This shows the (everlasting) complementarity between direct and indirect neutralino searches.

Gamma-rays and cosmic rays from neutralino annihilation in galactic halos. We shift now our attention to signals originating in neutralino annihilations which occur in the halo of our galaxy or in the halo of external galaxies.

The annihilation products of importance are those that are either rarely produced in astrophysical environments or otherwise have a peculiar characteristic that make them easily recognizable. In the first category are rare cosmic rays such as positrons, antiprotons, and antideuterons. In the second category are gamma-rays, whose spectrum is expected to contain a gamma-ray line at an energy corresponding to the neutralino mass (besides a gamma-ray continuum; see Figure 18). The gamma-ray line is produced directly in the primary neutralino annihilation into $\gamma\gamma$ or $Z\gamma$. Positrons, antiprotons, deuterons, and the gamma continuum are generated in the particle cascades that follow the decay of the primary annihilation products. Their spectra are therefore broad, with a typical energy which is only a fraction of the neutralino mass, and a shape whose details depend on which annihilation channels are dominant. Two neutralinos can in fact annihilate into a variety of primary products, depending on

Figure 16.17. Complementarity of direct and indirect neutralino dark matter searches. The figure shows that several supersymmetric models that are within the (expected) reach of a big neutrino telescope of $10 \text{ km}^2 \text{ yr}$ exposure are beyond the reach of current and future direct detection experiments. The vertical axis is the product of the neutralino-proton spin-independent scattering cross section $\sigma_{\chi-p}$ and the local neutralino density in units of $0.3 \text{ GeV}/\text{cm}^3 f_{CCDM}$. The horizontal axis is the neutralino mass m_χ . (Figure from Duda et al.(2003).)



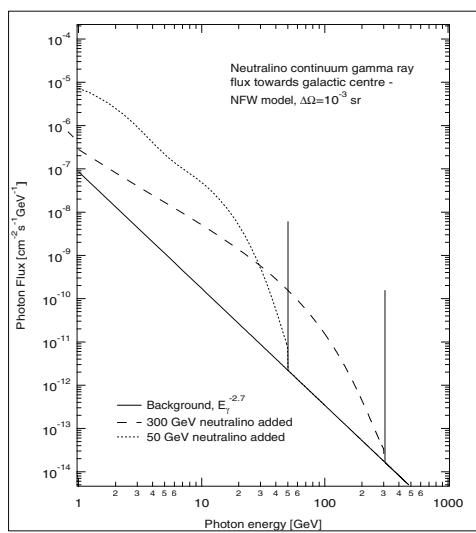


Figure 16.18. Two examples of gamma-ray spectrum from neutralino annihilation in the galactic halo. The numerical values refer to a specific model for the dark halo of our galaxy and for observations in the direction of the galactic center, but the spectral characteristics are general. Neutralino annihilations produce a gamma-ray line at an energy corresponding to the mass of the neutralino, and a gamma-ray continuum generated in the particle cascades following the primary annihilation. The figure illustrates that the shape of the continuum spectrum depends on the neutralino mass, but notice that it also depends on the neutralino composition. (Figure from Bergström, Ullio, & Buckley (1998).)

their masses and compositions: fermion pairs $f\bar{f}$, Higgs boson pairs H_iH_j , gauge boson pairs W^+W^- , ZZ , etc.

Detection of the gamma-ray line would be a smoking-gun for neutralino dark matter, since no other astrophysical process is known to produce gamma-ray lines in the 10 GeV – 10 TeV energy range. Good energy resolution is crucial to detect the neutralino gamma-ray line. The simulation in Figure 19 shows that the upcoming GLAST detector should have an adequate energy resolution.

For these signals, the dependence of the annihilation rate on the square of the density, see Eq. (16.48), has dramatic consequences. The predicted signals may change by several orders of magnitude when the model for the dark matter

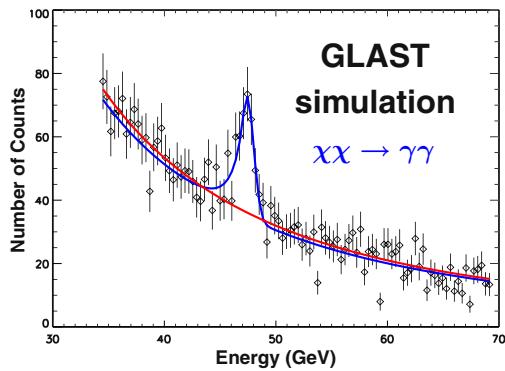


Figure 16.19. Simulation of a gamma-ray annihilation line from the annihilation of ~ 48 GeV neutralinos, superimposed on a gamma-ray background of astrophysical origin. The simulation includes the finite energy resolution of the upcoming GLAST detector. (Figure from the GLAST Science Brochure.)

density is changed, even without violating observational bounds on the latter. Truly, these observational limits are not very stringent, given the understandable difficulty of measuring the dark matter density. Anyhow the problem is currently there, and can be divided roughly into two questions: (1) what is the radial dependence of the average dark matter density in a galaxy, especially in our own? (2) how much substructure, i.e. clumps and streams, is there in galactic dark halos?

Dependence on the density profile Historically, the density profile of galactic dark halos has been given in terms of empirical density profiles whose density is constant in a central region and decreases as r^{-2} at large radii. The latter is the main ingredient in obtaining a flat rotation curve in the outer regions, which is a primary evidence for dark matter in galaxies. Central among these functions is the cored isothermal profile

$$\rho_{\text{BS}}(r) = \frac{\rho_0 a^2}{r^2 + a^2}, \quad (16.49)$$

where a is called the core radius Bahcall & Soneira (1980). This parametrization is so simple and so much used that it is sometimes called the ‘canonical’ density profile. Another interesting empirical parametrization is the density profile of Persic, Salucci & Stel (1996),

$$\rho_{\text{PS}} = \rho_0 \frac{a^2(r^2 + 3a^2)}{3(r^2 + a^2)^2}, \quad (16.50)$$

which provides good fits to rotation curves of hundreds of spiral galaxies (which are not as flat as one might think!).

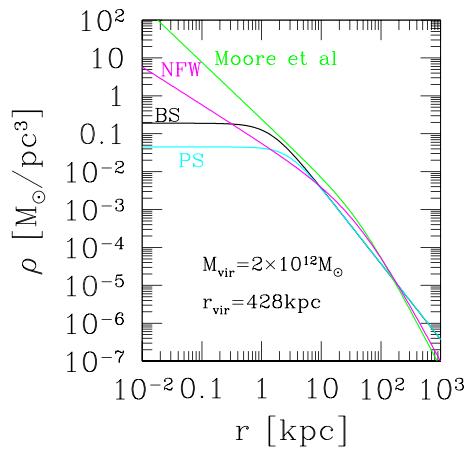
Numerical simulations of structure formation in the Universe have discovered that pure cold dark matter halos do not follow the previous empirical density profiles but instead have a universal shape whose parameters depend on the mass (or age) of the system. Navarro, Frenk & White (1996) have found this universal profile to have the form

$$\rho_{\text{NFW}}(r) = \frac{\rho_s r_s^3}{r(r + r_s)^2}, \quad (16.51)$$

where the parameter r_s is the radius at which the radial dependence of the density changes from r^{-1} to r^{-3} . The empirical dependence r^{-2} is then seen as an approximation in the transition region. Moore et al.(1998) suggest instead that the universal profile may be steeper at the center,

$$\rho_{\text{Moore}}(r) = \frac{\rho_s r_s^3}{r^{3/2}(r + r_s)^{3/2}}. \quad (16.52)$$

Figure 16.20. Dark matter density profiles for a galaxy resembling our own. Models ‘BS’ Bahcall & Soneira (1980) and ‘PS’ Persic, Salucci & Stel (1996) are empirical parametrizations which possess a central region with constant density (core). Models ‘NFW’ Navarro, Frenk & White (1996) and ‘Moore et al.’ Moore et al.(1998) are derived from numerical simulations of structure formation in the Universe, and in them the density in the central region increases as a power law of radius (cusp). All four models are normalized to the same total mass and virial radius.



Which of these two profiles better represents the results of numerical simulations is a question that must be answered by higher resolution simulations (which seem to be pointing to an inner slope γ that depends on the mass of the system).

The four profiles mentioned above are plotted in Figure 20 for a galaxy that could be our own, with total mass $M_{\text{vir}} = 2 \times 10^{12} M_\odot$ and virial radius $r_{\text{vir}} = 428 \text{ kpc}$.

The essential difference between the empirical and the numerical profiles, for what concerns neutralino signals, is their behavior at small radii. The empirical profiles have a central region with constant density, called a core, while the numerical profiles have a density that increases toward the center as a power law, called a cusp. Since the neutralino annihilation signals scale as the square of the density, any power law $r^{-\gamma}$ with $\gamma > 0$ is bound to give a higher signal from the central region, for a given total mass. Moreover, if $\gamma \geq 3/2$, the annihilation rate in the central region formally diverges, because $\int \rho^2 r^2 dr \propto \int r^{-2\gamma+2} dr$.

It is therefore crucial for our purpose to know which class of profiles, with a core or with a cusp, better resembles reality. Unfortunately, constraining a given dark matter profile using the kinematics of the central region is a hard problem, because the dynamics of the central regions is usually dominated by the visible matter, and when it is not, like in low surface brightness (LSB) spiral galaxies, the central parts are so small that the angular resolution of the observations is a major concern. Discrepant data and an apparent lack of universality in LSB profiles has generated endless controversies in the astrophysical community. At any rate, it must not be forgotten that the profiles mentioned above obtained in numerical simulations include cold dark matter only, and astro-

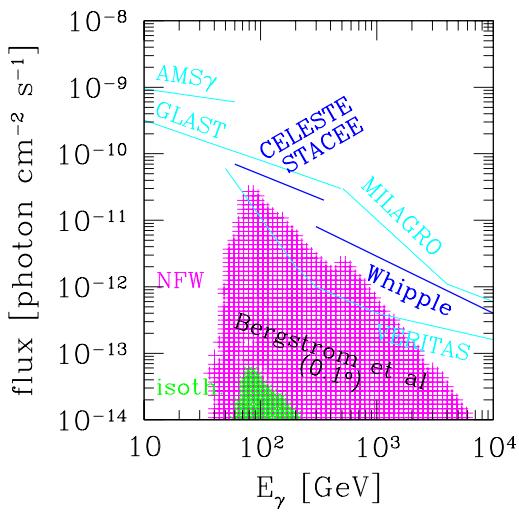


Figure 16.21. Expected gamma-ray flux in the gamma-ray line from neutralino annihilation in our galactic halo, coming from the direction of the Galactic Center (without the spike discussed in next Section). The photon energy E_γ is to good approximation equal to the neutralino mass. The upper set of points is for an NFW model Bergström, Ullio, & Buckley (1998), the lower set for a cored isothermal model. The gamma-ray fluxes differ by a factor of about 500. Also shown for approximate comparison are the sensitivities of current and upcoming gamma-ray telescopes to gamma-ray point sources. (Figure from Gondolo (2000).)

physical processes connected for instance with the gas and the formation of stars may well modify the density profile of dark matter (theoretical work in this direction is not lacking).

In the face of this situation, when making predictions for annihilation signals from neutralino dark matter in the halo, it is prudent for the moment to consider both possibilities, core and cusp.

The considerable differences in indirect neutralino signals between a cored and a cuspy density profile are illustrated in Figure 21 for a gamma-ray signal from neutralino annihilation from the direction of the Galactic Center (without the spike contribution discussed in the next Section). There is a factor of 500 difference between assuming an NFW profile or a cored isothermal profile in the theoretical calculation of the gamma-ray flux in the seven-parameter MSSM model in Bergström, Ullio, & Buckley (1998). Superposed on the plot are the sensitivities to gamma-ray point sources of various gamma-ray telescopes, both current and upcoming (as of 2000). The comparison with the theoretical expectations in the figure is not direct, however, because the neutralino emission may not be point like in some of the telescopes. The sensitivity curves are thus only meant to provide an approximate comparison.

It must be mentioned here that EGRET has detected gamma-ray emission from the direction of the Galactic Center Mayer-Hasselwander et al.(1998). However, in a reanalysis of the EGRET data by Hooper & Dingus(2002), the EGRET signal seems to originate from a source which is displaced with respect to the Galactic Center. Hooper & Dingus(2002) use their reanalysis to place constraints on the gamma-ray flux from neutralino annihilations, coming to an upper limit on the gamma-ray flux above 1 GeV which is about a factor of 2

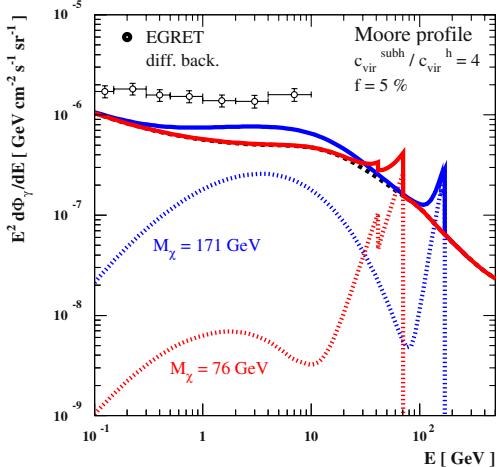


Figure 16.22. Expected isotropic gamma-ray background from neutralino annihilations in the early Universe. Dotted lines are the signal from neutralino annihilations; solid lines are the sum of the neutralino signal and a gamma-ray background of astrophysical origin. Two neutralino models are shown, together with the current EGRET measurements of the isotropic extragalactic background. The assumptions used for the dark matter profile are indicated. (Figure from Ullio, Bergström, Edsjö & Lacey (2002).)

higher than the theoretical predictions for neutralino masses $m_\chi \sim 50$ GeV, assuming an NFW profile. See Hooper & Dingus(2002) for details.

Effect of halo substructure The same numerical simulations that predict a cuspy dark halo profile also predict the existence of numerous dark clumps in galactic halos. These clumps are a natural outcome of the hierarchical formation of structure in the Universe. Small structures form first, and larger structures, like galaxies, grow by attracting and swallowing smaller structures. This process continues to the present day. Clumps that fall into a galaxy are pulled apart by tidal interactions: the material pulled out forms tidal streams that crisscross the galactic halo. The central parts of some of the clumps may survive for a long time, and become a galactic satellite.

The overall picture of hierarchical structure formation has found confirmation in a variety of context, from observations of galaxy clusters and of merging galaxies to the halo substructure detected in our own galaxy (see the discussion on the Sagittarius stream in Section 3.1 above, for example). A numerical discrepancy should however be mentioned. A counting of visible satellites of our own galaxy gives a number of luminous satellites that is much smaller than the expected number of dark satellites. A resolution to this discrepancy may be that only a small fraction of dark satellites becomes luminous. It is not clear why this should happen, but on the other hand we do not fully understand how galaxies become luminous in the first place. Thus for what concerns signals from neutralino annihilation, it makes sense to examine the effect of adding substructure, i.e. clumps and streams, to galactic dark halos.

Substructure in the halo tends to increase the annihilation signals because of the dependence of the annihilation rate on the square of the dark matter den-

sity (see Eq. (16.48)). The enhancement factor is linearly proportional to the density enhancement $\delta \equiv \rho'/\rho$. To understand why the dependence is linear instead of quadratic in δ , consider a box of volume V containing a total mass M . The density in the box is $\rho = M/V$ and the annihilation rate integrated over the whole box is

$$R_{\text{ann}} = \Gamma_{\text{ann}} V = \frac{\sigma_{\text{ann}} v}{m^2} \frac{M^2}{V}. \quad (16.53)$$

Now let all the mass be concentrated equally into N small boxes, each of volume V' , so that each box contains a mass $M' = M/N$ and thus has a density $\rho' = M'/V'$. The density enhancement is then $\delta = V/NV'$. The annihilation rate from the whole box is now given by a sum over the N small boxes as

$$R'_{\text{ann}} = N\Gamma'_{\text{ann}} V' = N \frac{\sigma_{\text{ann}} v}{m^2} \frac{M'^2}{V'} = \frac{\sigma_{\text{ann}} v}{m^2} \frac{M^2}{V} \frac{V}{NV'} = R_{\text{ann}} \delta. \quad (16.54)$$

One power of the density ρ in Γ_{ann} is compensated by a decrease in the volume where annihilations occur. Hence the signal enhancement is linear in the density increase.

An interesting consequence of the annihilation signal enhancement has been explored by Bergström, Edsjö, & Ullio (2001) and Ullio, Bergström, Edsjö &

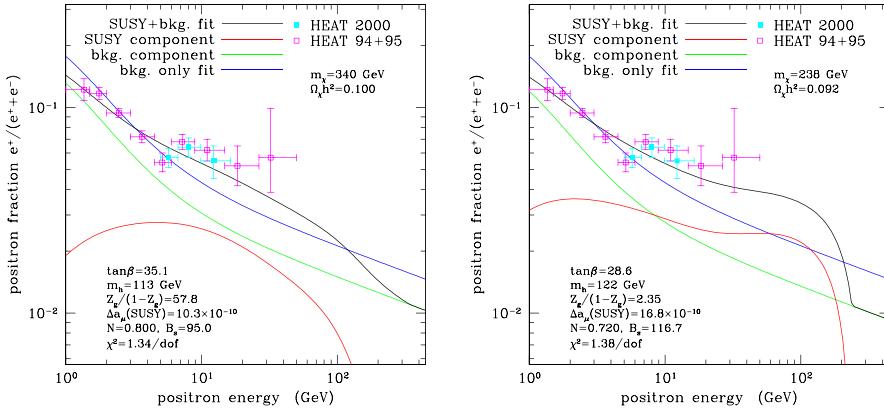


Figure 16.23. Two examples of neutralino models that provide a good fit to the excess of cosmic ray positrons observed by the HEAT collaboration. The two sets of data points (open and filled squares) are derived from two different instruments flown in 1994-95 and 2000. The lines represent: (i) the best expectation we have from models of cosmic ray propagation in the galaxy ('bkg. only fit'), which underestimate the data points above ~ 7 GeV; (ii) the effect of adding positrons from neutralino annihilations (lines 'SUSY component', 'SUSY+bkg. fit', and 'bkg. component', the latter being the resulting background component when the data are fitted to the sum of background and neutralino contributions). (Figures from Baltz, Edsjö, Freese, & Gondolo(2002)).

Lacey (2002). These authors have found that the density enhancements produced during the formation of the large scale structure in the Universe may lead to a substantial increase in the isotropic gamma-ray background from neutralino annihilations in the early Universe. Moreover, they found that this increase in the gamma-ray signal is not very sensitive to details of the galactic density profile. Expected gamma-ray spectra may in some models be close to the measured gamma-ray background, as illustrated in Figure 22. The gamma-ray spectra include both a continuum part and gamma-ray lines (two for each neutralino case: one for $\chi\chi \rightarrow \gamma\gamma$, the other for $\chi\chi \rightarrow Z\gamma$). The gamma-ray lines are asymmetrically broadened because photons emitted at earlier times have a larger redshift. The gamma-ray background due to neutralino annihilations should be searched for at high galactic latitudes, where the galactic emission is expected to be minimal. Detection of the line features depicted in Figure 22 would not require an energy resolution much better than the present one.

Another exciting application of a signal enhancement due to clumps in the galactic halo is the boost of the positron signal from neutralino annihilation up to the level of the excess of cosmic ray positrons observed by the HEAT collaboration. The HEAT collaboration flew two different detectors on balloons, and claims to have detected a ratio of positron to electron fluxes above ~ 7 GeV that is higher than the flux expected in current state-of-the-art models of cosmic ray production and propagation in the galaxy. These models aim at explaining all correlated signals in gamma-rays, radio waves, protons, electrons, positrons, heavy nuclei, etc. that are produced by cosmic rays in our galaxy.

One possible explanation is that the positron excess is due to the positrons generated in neutralino annihilation in the galactic halo. If the neutralinos are produced thermally in the early universe, which is the most common assumption, the annihilation cross section σ_{ann} is forced to be small by the requirement that the neutralino relic density is large enough for neutralinos to be the dark matter. Using the average value of the local dark matter density, of the order of 0.3 GeV/cm^3 , then leads to a positron signal which is more than an order of magnitude smaller than the excess measured by HEAT. Increasing the annihilation cross section σ_{ann} does not make the signal higher, because the density ρ decreases inversely with σ_{ann} , and hence the annihilation signal, being proportional to $\sigma_{\text{ann}}\rho^2$, decreases. Kane, Wang & Wells (2002) suggested that the neutralinos may not have been produced thermally in the early Universe, and were thus able to decouple the annihilation rate in the halo from the constraint coming out of the relic density requirement. Alternatively, Baltz, Edsjö, Freese, & Gondolo(2002) have suggested that substructure in the galactic halo may provide the necessary boost factor to the positron signal.

Enhancing the positron signal through clumps also enhances other annihilation signals, such as antiprotons and gamma-rays. Baltz, Edsjö, Freese,

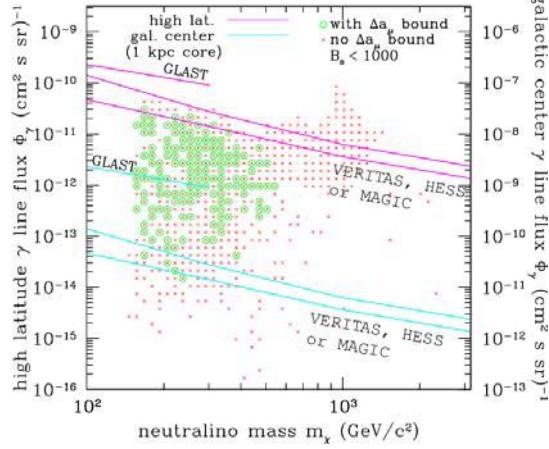


Figure 16.24. Sensitivity of upcoming gamma-ray telescopes to neutralino models that can explain the HEAT positron excess with neutralino clumps in the galactic halo. Model points are indicated by crosses; circles denote those models that in addition can also account for the measured deviation in the muon magnetic moment. The upper set of sensitivity curves corresponds to the high latitude gamma-ray line flux (scale on the left); the lower set of curves to the direction toward the galactic center (scale on the right; no steep spike around the central black hole is assumed). (Figure from Baltz, Edsjö, Freese, & Gondolo(2002).)

& Gondolo(2002) have performed a detailed analysis of these enhancements, and have concluded that it is possible to explain the HEAT positron excess with boost factors as small as 30, but typically higher, without obtaining too many antiprotons or gamma-rays. Figure 23 shows two examples of neutralino models that provide a good fit to the positron excess. On the left, the neutralino has mass $m_\chi = 340$ GeV and is an almost pure gaugino (gaugino fraction $Z_g = 0.98$); the boost factor is 95 and other parameters are listed in the figure. On the right, the neutralino is mixed (gaugino fraction $Z_g = 0.70$) with a mass of 238 GeV; the boost factor is 116.7. The χ^2 per degree of freedom is quite good for both fits, 1.34 and 1.38 respectively.

The ultimate test of the explanation of the positron excess by means of neutralino clumps will be the detection of a signal in gamma-rays. Gamma-ray production would in fact be enhanced by the same mechanism that would enhance positron production. Baltz, Edsjö, Freese, & Gondolo(2002) have found that almost all neutralino models that can explain the positron excess are within the sensitivity reach of upcoming gamma-ray telescopes (see Figure 24). The realistic possibility of confirming (or disproving) the neutralino origin of the positron excess is fascinating.

Signals from neutralino annihilation at the Galactic Center. The last indirect neutralino signals we consider are neutrinos, gamma-rays and radio

waves from a possible dark matter concentration around the black hole at the Galactic Center.

Evidence for the presence of a black hole at the center of our galaxy comes from studies of the motion of stars in orbit around the center. The speeds of these stars decrease from the center as the inverse square root of the radius, which is the primary indication for the existence of a point mass at the center. The mass of the central object is measured to be $\sim 4 \times 10^6$ solar masses, which are contained within a sphere of less than ~ 0.05 pc Eckart & Genzel (1997); Ghez, Klein, Morris & Becklin (1998); Ghez et al.(2003). No stellar or gas system can be so dense, indicating that the central object is most probably a black hole. The position of the black hole happens to coincide with the position of a strong radio source called Sagittarius A*, which is thus identified with the central black hole.

The radio emission from Sgr A* is easily explained by thermal emission of hot matter falling into the black hole. However, contrary to many of the similar black holes observed at the center of external galaxies, our galactic black hole does not emit intensely in the X-ray band, and it is controversial if it emits gamma-rays. Models for such ‘quiet’ black holes do exist, however, such as those involving advection-dominated accretion flows (ADAFs).

Further evidence for a black hole at the center of the Milky Way comes from the 2001 observation of a X-ray and infrared flares from the Galactic Center Baganoff et al.(2001); Porquet et al.(2003); Genzel et al.(2003b). The flare time scale and intensity can nicely be explained if the flare is produced near a black hole ?, see, e.g.,]Aschenbach:2004.

Dark matter may be driven near the black hole gravitationally, and may form a dense concentration around it Gondolo & Silk (1999). We will call this concentration a spike, so as to distinguish it from the dark matter cusps of Section 3.2.2. The formation of a spike is gravitational phenomenon similar to but less efficient than accretion of matter, in that the latter involves dissipation of energy and angular momentum and can thus produce concentrations which are smaller and denser.

How strong is the dark matter concentration around Sgr A*, or around a generic black hole? This is still a matter of investigation. The simplest case is that of adiabatic compression, and was analyzed by Gondolo & Silk (1999).

Adiabatic compression of an initial dark matter distribution produces two kinds of spikes. If the initial distribution before the black hole forms is thermal, the spike is shallow, with density profile $\rho \propto r^{-3/2}$. If the initial phase-space distribution is a power-law in energy, the spike is steep, i.e. $\rho \propto r^{-\gamma}$ with $\gamma > 2$. The physical reason for the higher concentration in the second case is the presence of many dark matter orbits with low speed, which are more easily driven into bound orbits when the black hole forms.

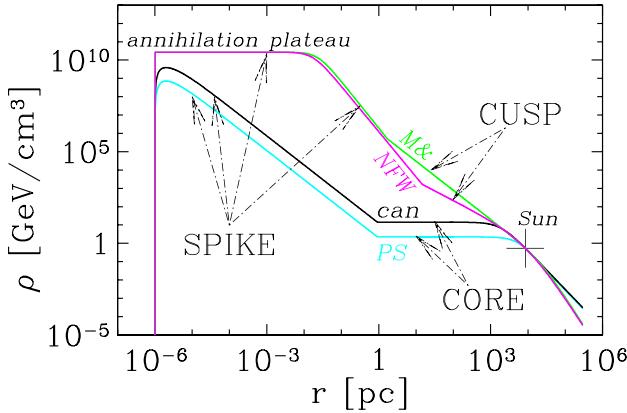


Figure 16.25. Density profiles of spikes that form adiabatically around the black hole at the center of our Galaxy. The position of the Sun is indicated by a cross. Four models for the halo profile are shown: two with cores ('PS' by Persic, Salucci & Stel (1996) and 'can' by Bahcall & Soneira (1980)) and two with cusps ('NFW' by Navarro, Frenk & White (1996) and 'M&' by Moore et al.(1998)). The spikes form within the radius of influence of the black hole, $r_{\text{infl}} \sim 1$ pc. In the 'annihilation plateau' neutralino annihilations have been so rapid as to deplete the number of neutralinos. (Figure from Buckley et al.(2001).)

These two kinds of spike are illustrated in Figure 25 for the same four halo models discussed in Section 3.2.2. Models with a core, like those by Bahcall & Soneira (1980) and Persic, Salucci & Stel (1996), give rise to a shallow spike around the central black hole, while models with a cusp, like those of Navarro, Frenk & White (1996) and Moore et al.(1998), produce steep spikes. In the very inner regions, the density may become so high that neutralino-neutralino annihilations may have had the time to deplete the number of neutralinos and an 'annihilation plateau' is formed. The typical radius of a spike around a black hole is determined by the radius of influence of the black hole, $r_{\text{infl}} \sim GM/\sigma_v^2$, which is the radius at which the gravitational potential energy becomes comparable to the typical kinetic energy in the dark matter gas (M is the black hole mass and σ_v is the gas velocity dispersion). For the black hole at the Galactic Center, $r_{\text{infl}} \sim 1$ pc.

Neutralino annihilation is enhanced in the spike, because of the dependence of the annihilation rate on square of the density. The Galactic Center then becomes a source of neutrinos, gamma-rays, radio waves, etc. from neutralino annihilation (Figure 26). The intensity of the emission depends on the steepness of the spike. If the spike is shallow, neutralino annihilation is generally undetectable. On the contrary, a steep spike at the Galactic Center produces interesting signals.

For example, if, disregarding Hooper & Dingus's reanalysis mentioned above, we attribute the EGRET gamma-ray emission from the Galactic Center to neutralino annihilation in a spike born out of an NFW profile, we obtain a high-energy neutrino flux that is either excluded or mostly detectable in a km^3 neutrino telescope (Figure 27). The flux of neutrino-induced muons above 25 GeV would be detectable over the atmospheric neutrino background for neutralino masses between ~ 100 GeV and ~ 2 TeV, while heavier neutralinos would already be excluded from the current limit on the neutrino emission from the Galactic Center Habig et al.(2001).

As a second and more dramatic example Gondolo (2000b), electrons and positrons from neutralino annihilation would emit synchrotron radiation as they spiral in the magnetic field that plausibly exists around the central black hole. While this synchrotron radiation is innocuous for a shallow adiabatic spike, it may exceed the observed radio emission by several orders of magnitude if the spike is steep. The radio synchrotron emission at 408 MHz is shown in Figure 28 for an adiabatic spike born out of an NFW profile, under two assumptions for the radial dependence of the magnetic field (a constant field of 1 mG and a field in equipartition with the infalling gas). In both cases, all dark matter neutralinos in the seven-parameter MSSM models considered are strongly excluded.

The examples above have assumed that the spike formed adiabatically and maintained its shape till the present time. This may not be the case.

For example, if the central black hole formed through the merging of two black holes of similar mass, Merritt, Milosavljević, Verde & Jimenez(2002) have shown that the spike would become shallow at the end of the merging, because dark matter particles would be kicked out of the black hole region via a gravitational sling-shot effect. The final shallow spike would not give dramatic signals from neutralino annihilation at the Galactic Center. In a realistic

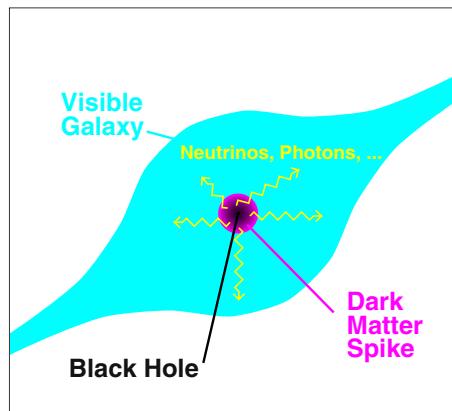


Figure 16.26. Artistic conception of emission from a dark matter spike around a black hole at the galactic center. Neutralino annihilation in the spike produces intense fluxes of neutrinos, gamma-rays, radio waves, etc. Some of these signals may be detectable.

scenario of this type, the two merging black holes of similar mass would be accompanied by their host galaxies whose mass would also be similar. The merging would then constitute what is known as a major merging. A major merging is capable of destroying the Galactic disk, and so must not have occurred after the disk formed about 10^{10} years ago. Thus, for this scenario to work, black holes of millions of solar masses should already have been in place at very early times. Is this possible? While hard to explain theoretically, supermassive black holes have indeed been observed in very distant quasars, at redshift $z > 6$, so the scenario may be plausible.

In these considerations, it must be kept in mind that there is a stellar spike around the black hole at the Galactic Center. The steepness of this stellar spike is however not very well known. With large uncertainties, Genzel et al.(2003) estimate the slope of the stellar spike to be $\gamma_{\text{stars}} \sim 1.3\text{--}1.4$. This means that the current stellar spike is probably shallow. We may think that the stellar spike is our best proxy for the dark matter spike. If so, also the dark matter spike would also be shallow, and thus inconsequential for neutralino signals. However, the dark matter and stellar spikes follow very different evolution histories, because contrary to the dark matter, binary collisions of stars and coalescence of two stars into one at collisions effectively relax the stellar system to a shallower spike.

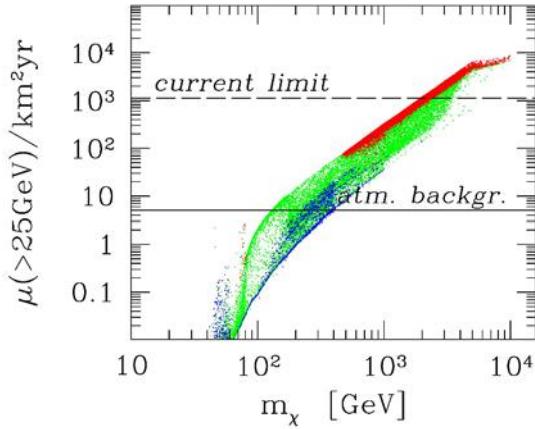
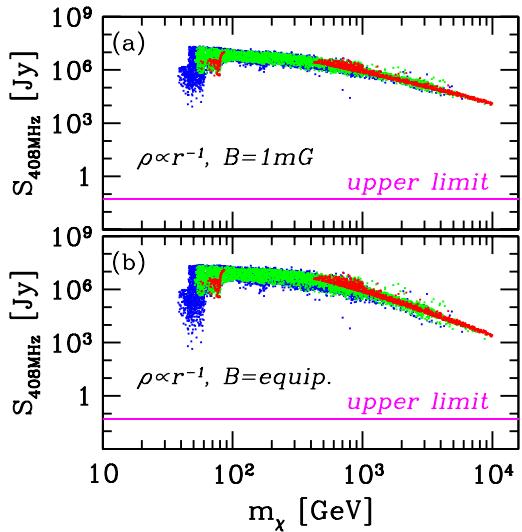


Figure 16.27. Predicted neutrino signal from neutralino dark matter annihilation in a steep adiabatic spike at the Galactic Center expressed as number of neutrino-induced muons per $\text{km}^2\text{-yr}$ in a neutrino telescope. The spike corresponds to an NFW profile. Each dot in the figure corresponds to a point in a seven-parameter weak-scale MSSM, and is normalized so that the gamma-ray flux from the spike coincides with the gamma-ray signal from the Galactic Center observed by EGRET. Models with heavy neutralinos are excluded by the current limit on the neutrino emission from the Galactic Center; models with neutralinos as light as ~ 100 GeV could be detected above the atmospheric neutrino background. (Figure from Gondolo(2002d).)

Figure 16.28. Synchrotron emission at 408 MHz expected from neutralino dark matter annihilation in a steep adiabatic spike at the Galactic Center. The spike corresponds to an NFW profile, and the synchrotron radiation is emitted by electrons and positrons produced in neutralino annihilation. Upper panel: constant magnetic field; lower panel: equipartition magnetic field. All dark matter neutralinos in the seven-parameter weak-scale MSSM considered are excluded by several orders of magnitude. (Figure from Gondolo (2000b).)



The final word has not yet been said regarding the distribution of dark matter at the Galactic Center, or around black holes and other compact objects in general. This is one of the exciting points of contact between the study of dark matter and the study of the formation and evolution of galactic nuclei.

4. Conclusions

Current cosmological data imply the existence of non-baryonic dark matter. We have discussed some of the most popular candidates and shown that none of the candidates known to exist, i.e. the active neutrinos, can be non-baryonic cold dark matter. Hence to explain the nature of cold dark matter we need to invoke hypothetical particles that have not been detected yet. Some of these hypothetical particles have been suggested for reasons different than the dark matter problem (such as sterile neutrinos, neutralinos, and axions), some others have been proposed mainly as a solution to the cold dark matter problem (e.g., self-interacting dark matter, WIMPZILLAs, etc.). Although most studies focus on the first category of candidates, especially neutralinos and axions, we should keep an open mind.

To illustrate how we can find out if dark matter is made of elementary particles, we have used neutralino dark matter as our guinea pig to survey several methods to search for non-baryonic dark matter. These methods range from a direct detection of dark matter particles in the laboratory to indirect observation of their annihilation products produced in the core of the Sun or of the Earth and in galactic halos, including our own. Direct searches may have found a signal from WIMPs (the annual modulation), but this claim is highly controversial

at the moment. Future direct searches have great promise, and might even be able to explore the local velocity distribution of WIMPs. These searches are complemented by indirect searches for high-energy neutrinos from the core of the Sun or of the Earth. Indirect searches using gamma-rays and cosmic rays from annihilations in the galactic halo are subject to uncertainties related to the detailed structure of the dark matter halo. Even more so are predictions for dark matter signals from the Galactic Center. Despite this, some anomalies in cosmic ray fluxes, namely a positron excess, may be explained by neutralino annihilation, and future gamma-ray observations may discover a gamma-ray line from neutralino annihilation in our galactic halo.

All of the examples we have presented are without doubt simple, elegant, and compelling explanations for the nature of non-baryonic dark matter. As we ponder on which one of them is realized in Nature, we must remember the words of astrophysicist Thomas Gold (as quoted by Rocky Kolb): “For every complex natural phenomenon there is a simple, elegant, compelling, wrong explanation.”

References

- Abazajian, K., Fuller, G. M., & Patel, M. 2001. “Sterile neutrino hot, warm, and cold dark matter,” Phys. Rev. D64, 023501
- Abrams, D., et al. [CDMS Collaboration] 2002. “Exclusion limits on the WIMP nucleon cross-section from the cryogenic dark matter search,” Phys. Rev. D66, 122003
- Ahlen, S. P., Avignone, F. T., Brodzinski, R. L., Drukier, A. K., Gelmini, G. & Spergel, D. N. 1987. “Limits On Cold Dark Matter Candidates From An Ultralow Background Germanium Spectrometer,” Phys. Lett. B195, 603
- Ahmad, Q. R., et al. [SNO Collaboration] 2002. “Direct evidence for neutrino flavor transformation from neutral-current interactions in the Sudbury Neutrino Observatory,” Phys. Rev. Lett. 89, 011301
- Ahrens, J., et al. [AMANDA Collaboration] 2002. “Limits to the muon flux from WIMP annihilation in the center of the earth with the AMANDA detector,” Phys. Rev. D66, 032006
- Akerib, D. S., et al. [CDMS Collaboration] 2003. “New results from the Cryogenic Dark Matter Search experiment,” Phys. Rev. D68, 082002
- Allanach, B. C., Kraml, S. & Porod, W. 2003. “Theoretical uncertainties in sparticle mass predictions from computational tools,” JHEP 0303, 016
- Aschenbach, B., Grosso, N., Porquet, D., & Predehl, P. 2004. “X-ray flares reveal mass and angular momentum of the Galactic Center black hole.” Preprint astro-ph/0401589
- Asztalos, S. J., et al. 2004. “An improved RF cavity search for halo axions,” Phys. Rev. D69, 011101

- Baganoff, F. K. et al. 2001. "Rapid X-ray flaring from the direction of the supermassive black hole at the Galactic Centre," *Nature* 413, 45
- Bahcall, J. N. & Soneira, R. M. 1980. "The universe at faint magnitudes. I - Models for the galaxy and the predicted star counts," *ApJS* 44, 73
- Baltz, E. A., Edsjö, J., Freese, K., & Gondolo, P. 2002. "Cosmic ray positron excess and neutralino dark matter," *Phys. Rev.* D65, 063511
- Baltz, E. A. & Gondolo, P. 2001. "Implications of Muon Anomalous Magnetic Moment for Supersymmetric Dark Matter," *Phys. Rev. Lett.* 86, 5004
- Baltz, E. A. & Gondolo, P. 2003. "Improved constraints on supersymmetric dark matter from muon g-2," *Phys. Rev.* D67, 063503
- Belli, P. 1997. Talk at *TAUP* 97, LNGS, Italy [published in R. Bernabei et al., *Nucl. Phys. B* (Proc. Suppl.) 70 (1999) 79].
- Benoît, A., et al. [EDELWEISS Collaboration] 2002. "Improved exclusion limits from the EDELWEISS WIMP search," *Phys. Lett.* B545, 43
- Bergström, L. 2000. "Non-baryonic dark matter: observational evidence and detection methods," *Rept. Prog. Phys.* 63, 793
- Bergström, L., Edsjö, J., & Gondolo, P. 1998. "Indirect detection of dark matter in km-size neutrino telescopes," *Phys. Rev.* D58, 103519
- Bergström, L., Edsjö, J., & Ullio, P. 2001. "Spectral gamma-ray signatures of cosmological dark matter annihilations," *Phys. Rev. Lett.* 87, 251301
- Bergström, L. & Gondolo, P. 1996. "Limits on direct detection of neutralino dark matter from $b \rightarrow s\gamma$ decays," *Astropart. Phys.* 5, 263
- Bergström, L., Ullio, P., & Buckley, J. H. 1998. "Observability of gamma rays from dark matter neutralino annihilations in the Milky Way halo," *Astropart. Phys.* 9, 137
- Bernabei, R., et al. 1998. "Searching For Wimps By The Annual Modulation Signature," *Phys. Lett.* B424, 195
- Bernabei, R., et al. 1999. "On a further search for a yearly modulation of the rate in particle darkmatter direct search," *Phys. Lett.* B450, 448
- Bernabei, R., et al. 2000. "Search for WIMP annual modulation signature: Results from DAMA/NaI-3 and DAMA/NaI-4 and the global combined analysis," *Phys. Lett.* B480, 23
- Bernabei, R., et al. 2003. "Dark matter search," *Riv. Nuovo Cim.* 26, 1
- Binetruy, P., Girardi, G., & Salati, P. 1984. "Constraints On A System Of Two Neutral Fermions From Cosmology," *Nucl. Phys.* B237, 285
- Birrell, N. D. & Davies, P. C. W., 1982. "Quantum Fields in Curved Space" (Cambridge: Cambridge University Press)
- Buckley, J., et al. 2001. "Gamma-ray summary report," in *Proc. of the APS/DPF/DPB Summer Study on the Future of Particle Physics (Snowmass 2001)*, ed N. Graf, eConf C010630, P407 [astro-ph/0201160]
- Cembranos, J. A. R., Dobado, A., & Maroto, A. L. 2003. "Brane-world dark matter," *Phys. Rev. Lett.* 90, 241301

- Cheng, H. C., Feng, J. L., & Matchev, K. T. 2002. "Kaluza-Klein dark matter," Phys. Rev. Lett. 89, 211301
- Chung, D. J. H., Kolb, E. W., & Riotto, A. 1988. "Nonthermal supermassive dark matter," Phys. Rev. Lett. 81, 4048
- Chung, D. J. H., Kolb, E. W., & Riotto, A. 1999. "Superheavy dark matter," Phys. Rev. D59, 023501
- Chung, D. J. H., Notari, A., & Riotto, A. 2003. "Minimal theoretical uncertainties in inflationary predictions," JCAP 10, 012
- Dine, D., Fischler, W., & Srednicki, M. 1981. "A Simple Solution To The Strong CP Problem With A Harmless Axion," Phys. Lett. B104, 199
- Dolgov, A. D. 2002. "Neutrinos in Cosmology," Phys. Rep. 370, 333
- Drukier, A. K., Freese, K., & Spergel, D. N. 1986. "Detecting Cold Dark Matter Candidates," Phys. Rev. D33, 3495
- Duda, G., Gelmini, G., Gondolo, P., Edsjö, J., & Silk, J. 2003. "Indirect detection of a subdominant density component of cold dark matter," Phys. Rev. D67, 023505
- Edsjö, J. & Gondolo, P. 1997. "Neutralino relic density including coannihilations," Phys. Rev. D56, 1879
- Edsjö, J., Schelke, M., Ullio, P., & Gondolo, P. 2003. "Accurate relic densities with neutralino, chargino and sfermion coannihilations in mSUGRA," JCAP 0304, 001
- Eckart, A., & Genzel, R. 1997. "Stellar proper motions in the central 0.1 pc of the Galaxy," MNRAS284, 576
- Ellis, J. R., Ferstl, A., & Olive, K. A. 2000. "Re-evaluation of the elastic scattering of supersymmetric dark matter," Phys. Lett. B481, 304
- Ellis, J. R., Hagelin, J. S., Nanopoulos, D. V., Olive, K. A., & Srednicki, M. 1984. "Supersymmetric Relics From The Big Bang," Nucl. Phys. B238, 453
- Feng, J. L., Matchev, K. T., & Wilczek, F. 2000. "Neutralino dark matter in focus point supersymmetry," Phys. Lett. B482, 388
- Freese, K., Frieman, J. A., & Gould, A. 1988. "Signal Modulation In Cold Dark Matter Detection," Phys. Rev. D37, 3388
- Freese, K., Gondolo, P., & Newberg, H. 2003. "Detectability of weakly interacting massive particles in the Sagittarius dwarf tidal stream." Preprint astro-ph/0309279
- Freese, K., Gondolo, P., Newberg, H., & Lewis, M. 2003. "The Effects of the Sagittarius Dwarf Tidal Stream on Dark Matter Detectors," Phys. Rev. Lett., to appear [Preprint astro-ph/0310334]
- Fukuda, Y., et al. [Super-Kamiokande Collaboration] 1998. "Evidence for oscillation of atmospheric neutrinos," Phys. Rev. Lett. 81, 1562
- Fukugita, M., Hogan, C. J., & Peebles, P. J. E. 1998. "The Cosmic Baryon Budget," ApJ503, 518

- Fulling, S. A. 1979. "Remarks On Positive Frequency And Hamiltonians In Expanding Universes," *Gen. Rel. Grav.* 10, 807
- Fulling, S. A., 1989. "Aspects of Quantum Field Theory in Curved Spacetime" (Cambridge: Cambridge University Press)
- Genzel, R. et al. 2003. "The Stellar Cusp Around the Supermassive Black Hole in the Galactic Center," *ApJ*594, 812
- Genzel, R., Schödel, R., Ott, T., Eckart, A., Alexander, T., Lacombe, F., Rouan, D., & Aschenbach, B. 2003b. "Near-infrared flares from accreting gas around the supermassive black hole at the Galactic Centre," *Nature*425, 934
- Ghez, A. M., Klein, B. C., Morris, M., & Becklin, E. E. 1998. "High Proper-Motion Stars in the Vicinity of Sagittarius A*: Evidence for a Supermassive Black Hole at the Center of Our Galaxy," *ApJ*509, 678
- Ghez, A. M., Salim, S., Hornstein, S. D., Tanner, A., Morris, M., Becklin, E. E., & Duchêne, G. 2003. "Stellar Orbits Around the Galactic Center Black Hole." Preprint astro-ph/0306130
- Gnedin, O. Y. & Ostriker, J. P. 2001. "Limits on Collisional Dark Matter from Elliptical Galaxies in Clusters," *ApJ*561, 61
- Goldberg, H. 1983. "Constraint On The Photino Mass From Cosmology," *Phys. Rev. Lett.*50, 1419
- Gondolo, P. 1996. "Phenomenological introduction to direct dark matter detection," in *XXXI Rencontres de Moriond: Dark Matter in Cosmology, Quantum Measurements, Experimental Gravitation*, Les Arcs, France [hep-ph/96-05290].
- Gondolo, P. 2000a. "Indirect detection of dark matter." Plenary talk at *Neutrino 2000*, Sudbury, Canada.
- Gondolo, P. 2000b. "Either neutralino dark matter or cuspy dark halos," *Phys. Lett.* B494, 181
- Gondolo, P. 2002c. "Recoil momentum spectrum in directional dark matter detectors," *Phys. Rev.* D66, 103513
- Gondolo, P. 2002d. "Neutrinos from dark matter." Talk at the NSF Meeting on *Neutrinos and Subterranean Science*, Washington, D.C.
- Gondolo, P., Edsjö, J., Bergström, L., Ullio, P., & Baltz, E. A. 2000. "DarkSUSY: A numerical package for dark matter calculations in the MSSM." Preprint astro-ph/0012234
- Gondolo, P., Edsjö, J., Ullio, P., Bergström, L., Schelke, M., & Baltz, E. A. 2002. "DarkSUSY: A numerical package for supersymmetric dark matter calculations." Preprint astro-ph/0211238. See <http://www.physto.se/~edsjo/darksusy>
- Gondolo, P. & Gelmini, G. B. 1991. "Cosmic Abundances Of Stable Particles: Improved Analysis," *Nucl. Phys.* B360, 145
- Gondolo, P. & Silk, J. 1999. "Dark matter annihilation at the galactic center," *Phys. Rev. Lett.*83, 1719
- Griest, K. & Seckel, D. 1991. "Three Exceptions In The Calculation Of Relic Abundances," *Phys. Rev.* D43, 3191

- Habig, A. [Super-Kamiokande Collaboration] 2001. “An indirect search for WIMPs with Super-Kamiokande.” Preprint hep-ex/0106024
- Hagiwara, K., et al. [Particle Data Group Collaboration] 2002. “Review Of Particle Physics,” Phys. Rev. D66, 010001
- Hall, L. J., Moroi, T., & Murayama, H. 1998. “Sneutrino cold dark matter with lepton-number violation,” Phys. Lett. B424, 305
- Hooper, D., & Dingus, B. L. 2002. “Limits on supersymmetric dark matter from EGRET observations of the galactic center region.” Preprint astro-ph/0210617
- Ibata, R., Lewis, G. F., Irwin, M., Totten, E., & Quinn, T. 2001. “Great Circle Tidal Streams: Evidence for a Nearly Spherical Massive Dark Halo around the Milky Way,” ApJ551, 294
- Ibata, R. A., Wyse, R. F. G., Gilmore, G., Irwin, M. J., & Suntzeff, N. B. 1997. “The Kinematics, Orbit, and Survival of the Sagittarius Dwarf Spheroidal Galaxy,” AJ113, 634
- Jungman, G., Kamionkowski, M., & Griest, K. 1996. “Supersymmetric dark matter,” Phys. Rep. 267, 195
- Kane, G. L., Wang, L. T., & Wells, J. D. 2002. “Supersymmetry and the positron excess in cosmic rays,” Phys. Rev. D65, 057701
- Kim, J. E. 1979. “Weak Interaction Singlet And Strong CP Invariance,” Phys. Rev. Lett.43, 103
- Kuzmin, V., & Tkachev, I. 1998. “Ultra-high energy cosmic rays, superheavy long-living particles, and matter creation after inflation,” JETP Lett. 68, 271
- Lauer, T. R., Statler, T. S., Ryden, B. S., & Weinberg, D. H. 1986 (approx). “A New and Definitive Meta-Cosmology Theory.” Scanned images available on Steve McGaugh’s website at www.astro.umd.edu/~ssm/mond/flowchart.html.
- Ma, C.-P. 1999. “Neutrinos and Dark Matter.” Preprint astro-ph/9904001
- Majewski, S. R., Skrutskie, M. F., Weinberg, M. D., & Ostheimer, J. C. 2003. “A Two Micron All Sky Survey View of the Sagittarius Dwarf Galaxy. I. Morphology of the Sagittarius Core and Tidal Arms,” ApJ599, 1082
- Markevitch, M., et al. 2003. “Direct constraints on the dark matter self-interaction cross-section from the merging galaxy cluster 1E0657-56.” Preprint astro-ph/0309303
- Mayer-Hasselwander, H. A. et al. 1998. “High-energy gamma-ray emission from the Galactic Center,” A&A335, 161
- Merritt, D., Milosavljević, M., Verde, L., & Jimenez, R. 2002. “Dark Matter Spikes and Annihilation Radiation from the Galactic Center,” Phys. Rev. Lett.88, 191301
- Miralda-Escudé, J. 2002. “A Test of the Collisional Dark Matter Hypothesis from Cluster Lensing,” ApJ564, 60
- Moore, B., Governato, F., Quinn, T., Stadel, J., & Lake, G. 1998. “ApJL,” Resolving the Structure of Cold Dark Matter Halos499, L5

- Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996. "The Structure of Cold Dark Matter Halos," ApJ462, 563
- Newberg, H. J., Yanny, B., Grebel, E. K., Hennessy, G., Ivezić, Z., Martinez-Delgado, D., Odenkirchen, M., Rix, H.-W., Brinkmann, J., Lamb, D. Q., Schneider, D. P., & York, D. 2003. "Sagittarius Tidal Debris 90 Kiloparsecs from the Galactic Center," ApJL596, 191
- Peccei, R. D., & Quinn, H. R. 1977. "CP Conservation In The Presence Of Instantons," Phys. Rev. Lett.38, 1440
- Persic, M., Salucci, P., & Stel, F. 1996. "The universal rotation curve of spiral galaxies - I. The dark matter connection," MNRAS281, 27; Erratum *ibid.* 283, 1102 (1996)
- Porquet, D., Predehl, P., Aschenbach, B., Gross, N., Goldwurm, A., Goldoni, P., Warwick, R. S., & Decourchelle, A. 2003. "XMM-Newton observation of the brightest X-ray flare detected so far from Sgr A*," A&A407, L17
- Primack, J. R. 2001. "Whatever Happened to Hot Dark Matter?," SLAC Beam Line31N3, 50 [Preprint astro-ph/0112336]
- Sarkar, S. 2003. "Neutrinos from the Big Bang." Preprint hep-ph/0302175
- Shifman, M. A., Vainshtein, A. I., & Zakharov, V. I. 1980. "Can Confinement Ensure Natural CP Invariance Of Strong Interactions?," Nucl. Phys. B166, 493
- Silk, J. 2003. Lectures at this School.
- Snowden-Ifft, D. P., Martoff, C. J., & Burwell, J. M. 2000. "Low pressure negative ion drift chamber for dark matter search," Phys. Rev. D61, 101301
- Spergel, D. N. & Steinhardt, P. J. 2000. "Observational Evidence for Self-Interacting Cold Dark Matter," Phys. Rev. Lett.84, 3760
- Spergel, D. N., et al. 2003. "First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters," ApJS 148, 175
- Stiff, D., Widrow, L.M., & Frieman, J. 2001. "Signatures of hierarchical clustering in dark matter detection experiments," Phys. Rev. D64, 083516
- Tovey, D. R., Gaitskill, R. J., Gondolo, P., Ramachers, Y., & Roszkowski, L. 2000. "A new model-independent method for extracting spin-dependent cross section limits from dark matter searches," Phys. Lett. B488, 17
- Ullio, P., Bergström, L., Edsjö, J., & Lacey, C. 2002. "Cosmological dark matter annihilations into γ rays: A closer look," Phys. Rev. D66, 123502
- Verde, L., et al. 2002. "The 2dF Galaxy Redshift Survey: the bias of galaxies and the density of the Universe," MNRAS335, 432
- Wald, R. M., 1994. "Quantum Field Theory In Curved Space-Time And Black Hole Thermodynamics"(Chicago, IL: University of Chicago Press)
- Yoshida, N., Springel, V., White, S. D. M., & Tormen, G. 2000. "Weakly Self-interacting Dark Matter and the Structure of Dark Halos," ApJL544, L87

Zhitnitsky, A. R. 1980. "On possible suppression of the axion hadron interactions," Sov. J. Nucl. Phys. 31, 260