2002/2003

Proceedings of the students' workshop held at
Chalmers University of Technology and Göteborg University
14 and 15 May 2003

—

Following the course 'Hot Topics in Astrophysics' held at
Chalmers University of Technology and Göteborg University
October 2002 – May 2003

—

Edited by:

Alessandro B. Romeo,

Christoffer Petersson, Daniel Persson & Oscar Agertz

Contents

Solar Activity and Terrestrial Climate:
Is the Sun Responsible for Global Warming?

Tommy Lindfors

Göteborg University
SE-41296 Göteborg, Sweden
(tommy.lindfors@gbgsd.se)

*

Abstract

During the last century the Earth's surface temperature has risen some 0.8 degrees. According to most scientists this raise is supposed to be consequence of our extensive use of coal and oil. As we shall see in this paper different forms of solar activity has risen dramatically since year 1900. The question is then whether these changes might be affecting the Earth climate and how this mechanism then is working. In the search for such a mechanism we study the connection between the solar wind, the heliosphere and the cosmic ray flux. We also investigate if variations in these kinds of indirect solar activity might be responsible for the changing Earth's climate.

1 Introduction

For five billion years has our Earth been situated next to a big dangerous star, the Sun. This star has been the source of all the energy and light necessary for the development of the Earth. The last hundred years the temperature on the Earth has been raising drastically. Now our scientists are worried that human activity this time might bee the main course behind the warming of the Earth. Scientists have for more than a hundred years argued that increasing human use of fossil fuels would lead to a bigger amount of greenhouse gases in the atmosphere and a warmer climate on Earth. Today most observers are convinced that this is the mechanism behind the global warming we have experienced lately.

A more detailed study shows that the global temperature has been raising some 0.8 degrees from 1900 until 2000. The first 40 years of this period until 1940 the temperature increased 0.4 degrees. From 1940 until 1975 there were now warming at all the temperature rather fell 0.1 degree. But after 1975 there have been another raise of another 0.4 degrees. The global temperature raise the last hundred years have in a bigger perspective been rather dramatic. The temperature 1900 was some 0.3 degrees lower than the mean

*Hot Topics in Astrophysics 2002/2003, Alessandro B. Romeo, Christoffer Petersson, Daniel Persson & Oscar Agertz (Eds.), Chalmers University of Technology and Göteborg University, 2003.

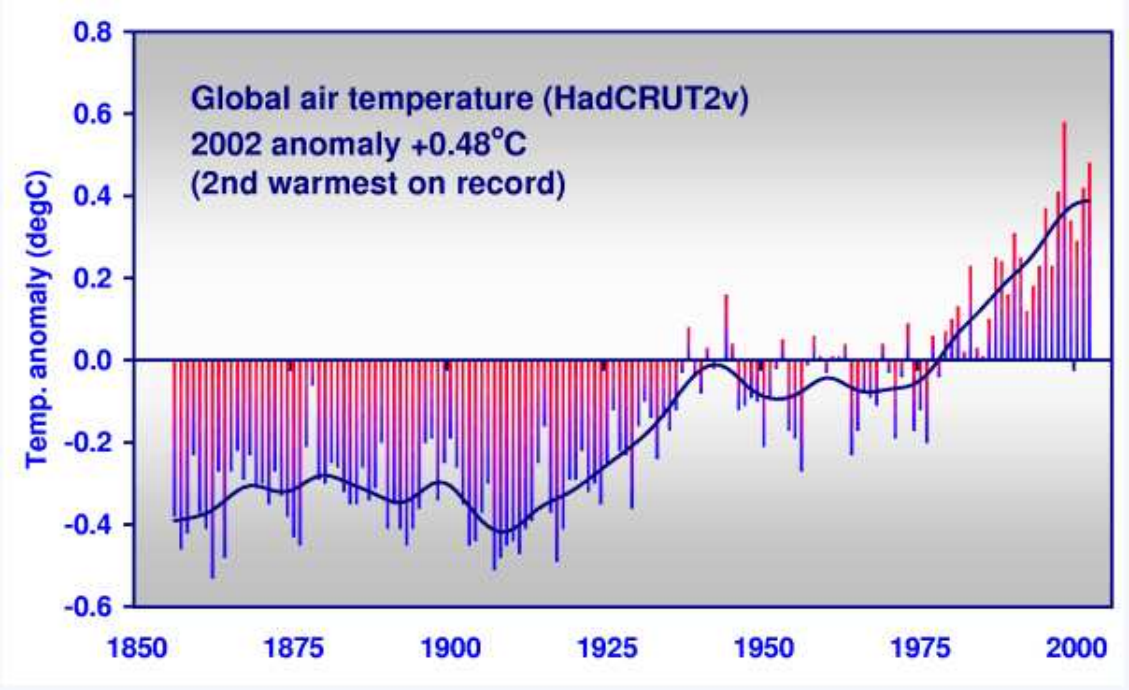

Figure 1: Variations of the Earth's surface temperature for the past 140 years. Departure in temperature from the 1961 to 1990 average (from IPCC homepage).

temperature for the entire millennium. This means that only the 0.4 degree temperature raise from 1975 until today is a raise above what has been normal on the Earth. This then is the only raise that is supposed to be man maid and can be a result of our productions of additional greenhouse gases.

1.1 The Greenhouse Gases

The main greenhouse gases in Earth's atmosphere are water vapor (H_2O) and carbon dioxide (CO_2). The amount of water vapor in the atmosphere is frequently changing and not easy to manipulate with human activity. The rate of Water in the atmosphere is usually close to 1 %.

Our extensive use of fossil products as oil and coal is raising the concentration of (CO_2) in the Earth's atmosphere. Higher amounts of greenhouse gases in the atmosphere are raising the temperature on Earth. Consequently the raising temperature in Earth climate is seen as a result of our growing use of oil and coal. The rate of (CO_2) in the atmosphere has during the last 40 years risen from 0.0320 % to 0.0365 %. Could this then bee the main course behind the global warming?

The Intergovernmental Panel of Climate Change (IPCC), a panel of scientists under the United Nations, at least thinks so. In a report from 2001 this temperature raise caused IPCC to write: "...here is new and strong evidence that most of the warming observed over the last 50 years is attributable to human activities." (see IPCC 2002).

A few months later the IPCC head KE Trenberth explained: "Further, known causes such as changes in the Sun and volcanic activity in the past 50 years have, if anything, led to cooling in this interval leaving only the human-caused increase in greenhouse gases as the culpit. This reasoning has also been quantitatively confirmed with climate models." (see Science 2001).

This is a strange analysis coming from a qualified scientific board knowing that almost

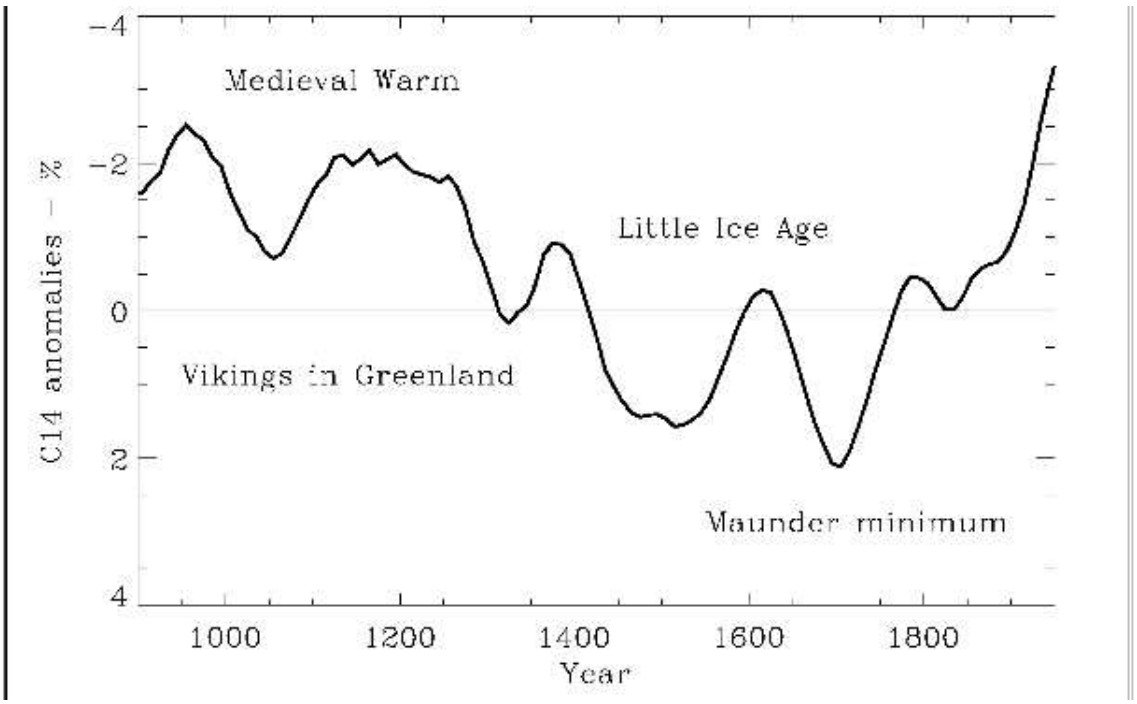

Figure 2: Temperature and ^{14}C changes since year 1000 (from Svensmark 1998).

all former changes in Earth temperature has been coursed by changes in the radiation of the Sun. The raise in global temperature from 1900 to 1940 can hardly be depending on raising abundance of carbon dioxide in the atmosphere since this raise were very small. From 1940 until 1975 the abundance of carbon dioxide in Earth atmosphere was continuously growing at increasing speed but the global temperature didn't raise but rather fell 0.1 degree in these 35 years. This is a strong argument against the theory of global warming mainly as a result of growing amounts of carbon dioxide. Probably this is the reason why greenhouse gases are seen as responsible for the global warming since 1970.

Lack of solid proofs in the case against the raising rate of greenhouse gases as been responsible for the global warming is why skeptical scientists like the Danish astronomer H. Svensmark comment on the IPCC report as only being "based on an analysis of Solar Irradiance variations" (see Svensmark 2001).

1.2 Solar Activity and Temperature

Looking further back in the temperature history of the Earth we can see that the temperature has changed a lot. Between 1000 and 1300 solar activity was very high and we had the Medieval Warm period when for example the Vikings settled in Greenland and wine was made from grapes grown in England. After 1300 solar activity decreased considerably and a long cold period followed. The settlers disappeared from Greenland and in the winter-times the River Thames regularly froze in London. This cold period, Little Ice Age, lasted with variations until the end of the nineteenth century when solar activity started to raise again and with it raised the global temperature. A particular cold period in the Little Ice Age was the Maunder Minimum between 1650 and 1720. This was e period with extreme low solar activity and extreme cold weather.

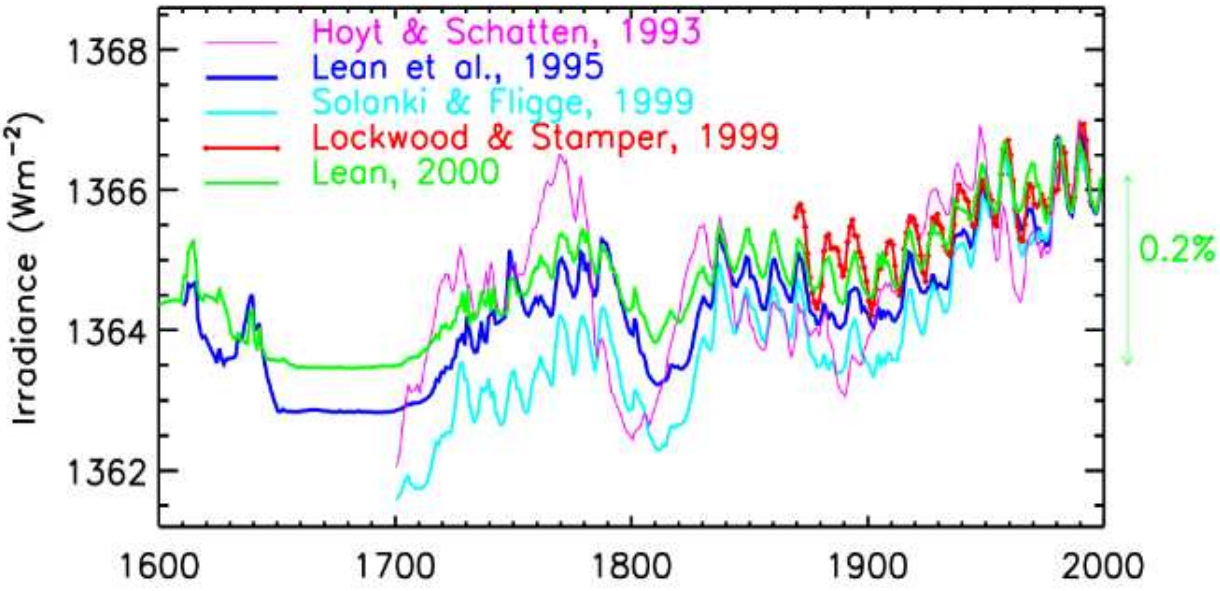

Figure 3: Solar activity for four centuries. Five reconstructed curves (from Svensmark 1999).

1.3 Reconstructed Solar Radiation

Today we can reconstruct curves showing the varying solar activity for the latest four hundred years. Even if there is some differences between these curves they all show the same long term trends for the solar activity. The registered changes in solar activity quite obviously are varying in the same mode as the global temperature. For example is different phases in the Little Ice Age clearly registered in the solar activity and in temperature and the coldest period the Maunder Minimum 1650 until 1720 is visible in solar activity. These curves over solar activity all verify the raise in solar activity since 1900 and they all also register a 30 years brake in the raise before the solar activity begin to raise again as we have seen in the temperature curve. Even if the raise in solar activity during the last century is quite noticeable the total raise are not more than 0.2 % of the radiation. This raise is not big enough to in it self explain the raise in temperature we have registered since 1900.

During the last millennium there has been both warm and could periods but the most dramatic changes have occurred since 1900 and the hottest years the last millennium seems to have been 1998 and then 2002. The warming of the Earth during the last hundred years is maybe not so extraordinary seen in a longer perspective, but the increase in temperature during this period is surely worth a study.

2 With a Violent Star as a Neighbour

2.1 Sunspots and Radiation

The Sun is a variable star that has had an almost constant energy output for billions of years. Energy from the Sun has been the basis for development of life on earth from the very beginning. The solar radiation emerging from the Sun's atmosphere contains wavelengths spanning the entire spectrum from X rays to radio waves. The Sun has its

maximum flux radiation in the visible wavelengths where the human eye got it's maximum sensitive and where the green plants have developed their highest sensitivity. The Sun's mean radiation at the Earth distance, known as "the solar constant" is $S = 1366 \pm W/m^2$.

With modern technology we know this solar constant not to be constant. Patterns in the changing features on the surface of the Sun have been studied for centuries. Dark sunspots pass over the surface of the Sun as it rotates once every 27 days. Besides that there is a regular 11-12 year pattern observed in the length of the period between sunspots maximums.

It is not only the sunspot frequency that varies with this 11-12 year periodicity. The same periodicity can be found in all kinds of radiation from the Sun. In shorter wavelengths as X-ray an UV rays the periodicity is obvious. Even if the intensity of the flux in these wavelengths is low the difference between the highest and the lowest intensity is bigger here than in the visible light. The UV-radiation for example is responsible for the formation of ozone molecules in the higher atmosphere of the Earth. Consequently we can find this 11 years periodicity both in the abundance of ozone in the atmosphere and in the curve for the Sun's total radiation flux as well. But the amplitude in the periodically changing total flux from the Sun is not more then some 0.15 %. It is hard to see how so small changes in the solar radiation could have any significant influence on the climate on Earth.

Apart from the electromagnetic radiation emanating from the Sun there also is a continuous flux of charged particles streaming out in the Space from the surface of the Sun. These charged particles, this solar wind, mainly consist of protons and electrons emitted with a speed of about 400 km/s. Several types of solar events can cause particles with much higher energy and velocity to be superimposed on this background solar wind and form enormous mass ejections from the surface of the Sun. Even in the solar wind and in these mass ejections we find the same 11 year periodicity.

2.2 The Sun Dynamo Produces a Gigantic Magnetic Field

Today scientists are working with theories about the magnetic field of the Sun as being the engine giving electromagnetic radiation and particle emission from the Sun this 11 year periodicity. To explain the solar cycle scientists visualize an enormous dynamo inside the Sun. Here the magnetic field is generated when enormous layers of electrified gas in the interior of the Sun are rotating next to other layers of gas with different relative speed. The relative motion of neighbouring layers of gas containing charged particles generate enormous magnetic fields in and around the Sun. This dynamo is situated some 200 million meters under the surface of the Sun. Here, a third of the way down to the core, the radiate zone meets the convective zone and the gigantic magnetic field is produced.

The solar wind and the magnetic flux streaming out from the Sun are supporting each other in producing a giant heliosphere around the Sun and our solar system. This heliosphere is working as a magnetic shield protecting the solar system against cosmic ray flux from the outer space making our solar system a friendlier place to live in. The particles in the cosmic ray that mainly is coming from other parts of our galaxy is some hundred times more energetic than the particles in the solar wind.

The development and use of different types of space technology under the past few decades have resulted in a much better understanding of a lot of solar variability and its mechanisms. Comparison of solar and stellar radiation indicates that the Sun is potentially capable of wider range of variability than what we have witnessed lately. About the source and mechanism of this variability we still know very little.

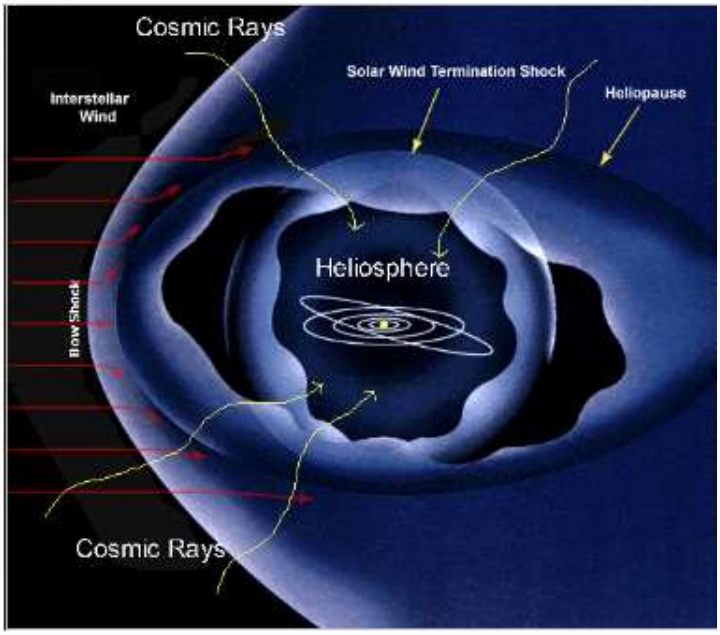

Figure 4: The Heliosphere (from Svensmark 1999).

3 Solar Activity Affecting Earth Climate

3.1 Sunspots and Climate

1881 William Herschel, a famous astronomer, suggested that the number of sunspots directly controlled the price of wheat at the London market. Herschel had observed that less rain fell when there were few sunspots and according to Herschel this could be the mechanism that disturbed the wheat production.

If we cannot explain how solar activity affects Earth climate then our explanation is no science. Several attempts have been done for more than a century to explain how correlation between solar variability and climate disturbances can be explained. A lot of other phenomena in our daily lives has been related to this regular changes in the sunspot activity of the Sun as well. Obvious correlations between the number of sunspots and drought periods in different regions of the Earth has been reported as well as correlations between the changes in the water temperature in shallow lakes and the sunspot numbers. But lacking scientific argumentation has until now been disqualifying for this kind of theories. Not even W. Herschel could explain how dark spots on the face of the Sun could produce rain in England.

3.2 Sunspot Variation, the Heliosphere and the Origin of ^{14}C

In the same way as the heliosphere is protecting our solar system against the cosmic ray flux the magnetic field of the Earth, the magnetosphere is protecting us from the solar wind and its energetic particles. The solar wind consists mostly of protons and electrons, coming from the Sun. The much more energetic cosmic rays, about 100 times more energetic, that have succeeded in penetrating the heliosphere will not be stopped until they crush against molecules in the upper layers of the Earth's atmosphere. Here molecules and atoms are broken and new are formed. One type of atoms being formed under these circumstances are the ^{14}C isotope that is being formed out of a ^{14}N atom and

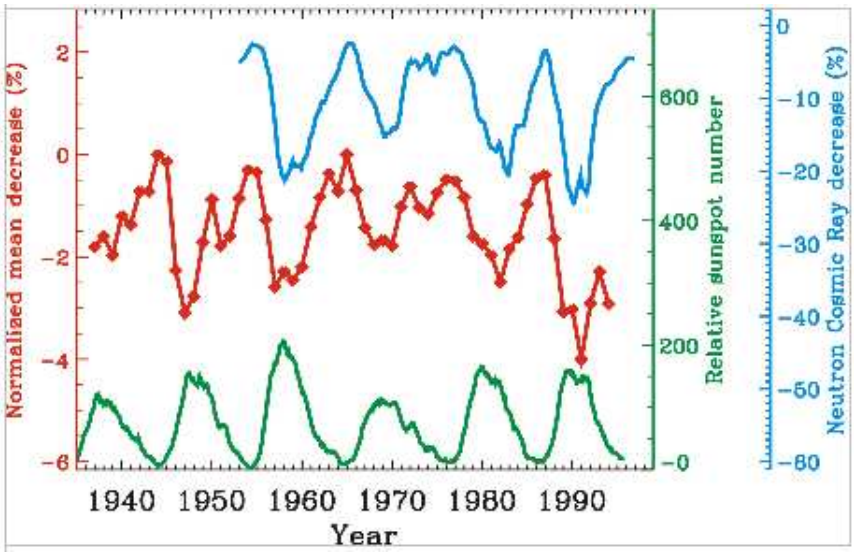

Figure 5: The top blue curve represents the cosmic ray flux from the neutron monitor Climax, Colorado. Middle red curve is annual mean variations in Cosmic Ray flux as measured by ionization chambers. The bottom green curve is the relative sunspot number. Note that the amplitudes are not well correlated (from Svensmark 1999).

a free neutron. As a result of this we can find the same 11 year variation in the abundance of ^{14}C in the higher layers in the atmosphere as in the solar cycle.

The ^{14}C isotopes are radioactive with a half-live that is 5730 years. Today the ratio $^{14}C/^{12}C$ is a commonly used tool to date old organic matter. Scientists know how this ratio has been changing for more than 50 000 years back in time. When a living organism like a tree dies it stop breading new carbon dioxide and no new ^{14}C is being picked up. After 5730 years the ratio of $^{14}C/^{12}C$ in a peace of wood has then dropped to 50 % of the original and after another 5730 years it has dropped to 25 % and so on. By checking this ratio in dead organisms it is then possible to find out how old this peace of wood or whatever organism is.

As a consequence information about how this $^{14}C/^{12}C$ ratio has been changing far back in history is important and has been studied intensely. This means that solar activity now is possible to trace far back in time by studying how the $^{14}C/^{12}C$ ratio or the abundance of other similar radioactive isotopes has varied in the atmosphere. These radioactive isotopes are all produced in the upper layers of the atmosphere by high energetic cosmic rays just like ^{14}C. A striking result from this kind of records is that there is a strong correlation between hot and cold climate periods and periods of high and low solar activity.

Variation in Time of Cosmic Ray Flux and Sunspot Number

For half a century now we have been able to detect high energetic cosmic rays penetrating the Earth atmosphere. In Fig. 5 we can at the bottom row see how the number of sunspots vary. At the curves above we can see that the cosmic ray flux got exactly the same 11 year periodicity, both measured with neutron counter and ion chamber. The only difference is that when the sunspot numbers got a maximum the cosmic ray flux got a minimum. This is a result of the strengthening or weakening of the heliosphere around the Sun and the Earth. With high solar activity we got a lot of sunspots and a strong heliosphere around our solar system. Few highly energetic particles from the cosmic ray can then reach the Earth atmosphere and few ^{14}C isotopes are produced. With low solar activity

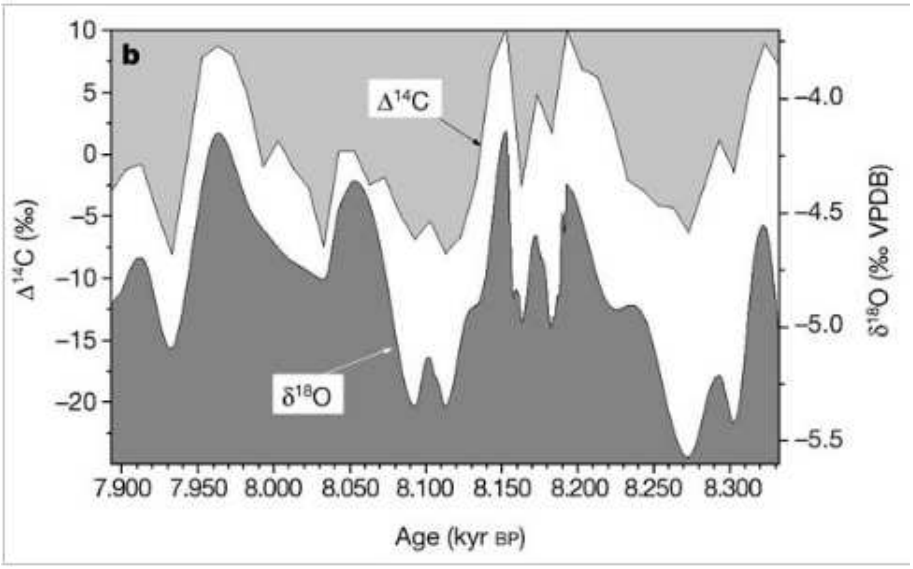

Figure 6: The coherence between solar variability and the monsoon in Oman 8 000 years ago. The changing monsoon has been registered by the formation of stalagmites in caves in Oman. The upper grey curve shows the variation in solar activity, ^{14}C, while the bottom curve in a similar way represents the monsoon and the growing stalagmites, ^{18}O (from Neff et al. 2001).

on the other hand we got few solar spots and a weak heliosphere around the solar system. The Earth atmosphere will then be hit by a lot of high energetic cosmic ray particles that has been able to penetrate the weak heliosphere. Then a lot of ^{14}C isotopes will be produced too. This means that the production of ^{14}C in Earth atmosphere is inversely proportional to the strength of the solar activity and the magnetic field from the Sun. A strong magnetic field around the solar system gives lower ^{14}C ratio in our atmosphere while a weaker magnetic field gives high ^{14}C ratio. Information about how this $^{14}C/^{12}C$ ratio has been changing far back in history can then be used to calculate variations in solar activity thousands of years back in time.

Solar Variability and the Monsoon in Oman, 8000 Years Ago

This amazing technique has made it possible for us to study how variations in solar activity has been influencing the climate on Earth thousands of years back. In Fig. 6 the upper curve for ^{14}C demonstrate how solar activity have changed some 8 000 years ago. The lower curve for ^{18}O demonstrates in a similar way that the monsoon in Oman has changed during the same period. It is obvious that there existed a strong link between the solar activity and the climate in this part of the Arabia. What kind of mechanism that keeps this relation working we still know very little about.

3.3 Ion-Pair Production

In a study published 1959 E.P. Ney can show convincing curves over how new ion-pairs are produced in the atmosphere over Thule. Ney here finds big differences between years with high solar activity and years with low solar activity. At altitudes of 10 km and higher Ney finds a 20 - 30 % higher production of ion-pairs when the solar activity is low than when it is high. This indicates that the ion-pairs are produced with energy from the

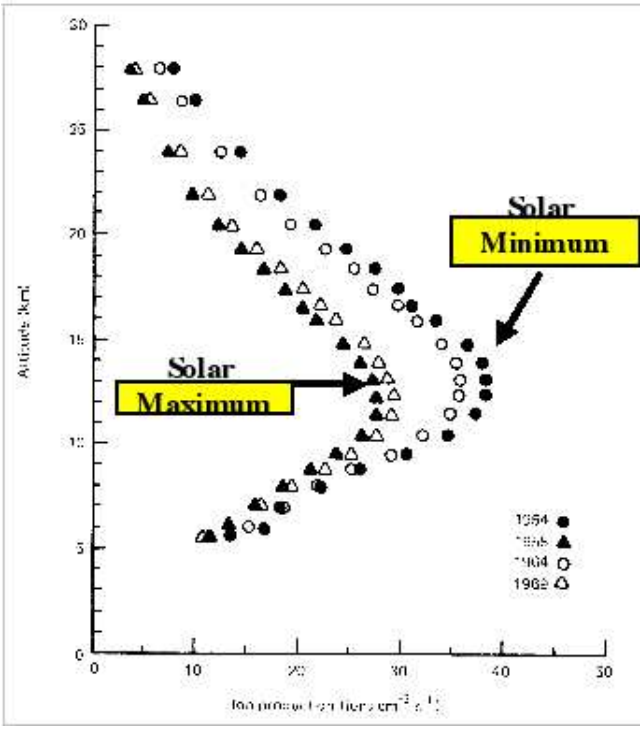

Figure 7: Ion-pair production over Thule. The two rows with triangles on the left-hand side represents years with high solar activity (1958, 1969). The rows with circles on the right represents years with low solar activity (1954, 1964) (from Ney 1959).

cosmic rays much the same way as we have seen concerning the production of ^{14}C and other radioactive isotopes in the atmosphere. Then the production of ion-pairs in higher layers of our atmosphere varies in abundance in the same way as ^{14}C and the number of sunspots.

"The meteorological variable subject to the largest solar-cycle modulation in the dense layers of the atmosphere is the atmospheric ionization produced by cosmic rays." writes EP Ney (see Ney 1959).

3.4 Cosmic Rays and Cloud Frequency

The Earth's climate is a result of the radiation from the Sun. The amount of this radiation that is absorbed, redistributed or reradiated is fundamental for the development of the climate on the Earth. Any changes in the energy that is received at the surface of the Earth or radiated from the Earth will immediately have an effect on our climate. The mixture of the gases in the atmosphere is of great importance in this process. A change in the amount of greenhouse gases in the atmosphere will then have a big impact. Even more so will a change in cloudiness around the Earth affect the climate. H Svensmark and N Marsh have studied correlations between the curve for galactic cosmic rays flux and the global cloudiness. In this curves they have found a strong correlation between lower liquid clouds and the cosmic ray curve (see Fig. 8). In higher regions they could find no correlations between the formation of clouds and the cosmic ray flux. With low solar activity and a weak heliosphere around the solar system more galactic cosmic rays penetrate the Earth atmosphere and there would be more cloud formation on lower altitudes. The cloud formation is then supposed to start with the atmospheric ionization E.P. Ney reported as a consequence of the galactic rays. This ionization would then help

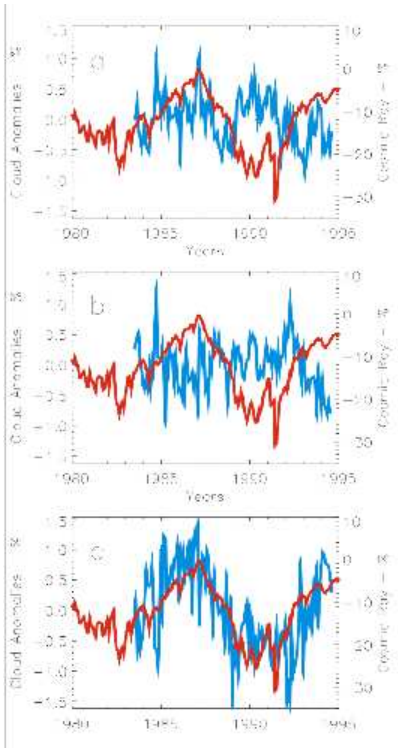

Figure 8: Global average of monthly cloud anomalies for a) high (> 6.5 km) b) middle (3.2-6.5 km) and c) low (< 3.2 km) clouds. The blue line stands for the cloudiness and the red line represents galactic cosmic ray flux. ISCCP D2 data for 1983-1994 used (from Marsh & Svensmark 2000).

in building new aerosols activated as cloud condensation nuclei in the building of new and bigger raindrops.

Hence it is argued that a mechanism to explain the cosmic ray - cloud link might be found in the role of atmospheric ionization in aerosol production and/or growth. A small influence from atmospheric aerosol production from cosmic ray ionization could then have a significant influence on the Earth's climate. Records from the last 1000 years indicate that changes in Earth climate have occurred in accord with variability in cosmic ray intensities. Svensmark et al claims that this possibly could be the missing link between changing solar activity and Earth climate.

In Fig. 9 the cloudiness clearly follows the curve for the cosmic ray flux. The lag of almost two years between the cosmic ray flux and the solar flux before 1987 demonstrates that Earth's cloud cover follows the changes in cosmic ray curve flux more than it follows solar flux. This indicates that the mechanism linking cloud cover with solar activity will involve changes in ionization properties of the atmosphere as a fundamental property.

3.5 Changing Magnetic Flux From the Sun

Today we got a reconstructed history of solar source magnetic flux for the past century. This history agrees well with the available satellite observations for the period since 1964. The solar magnetic field is observed to have more than doubled over the past century. At the same time the galactic cosmic radiation measured at Huancayo consequently has decreased by 3.5 %. As a result of this there would follow a decrees in the amount of lower liquid clouds and according to Marsh & Svensmark 2000, one can infer a $1.4\,Wm^{-2}$

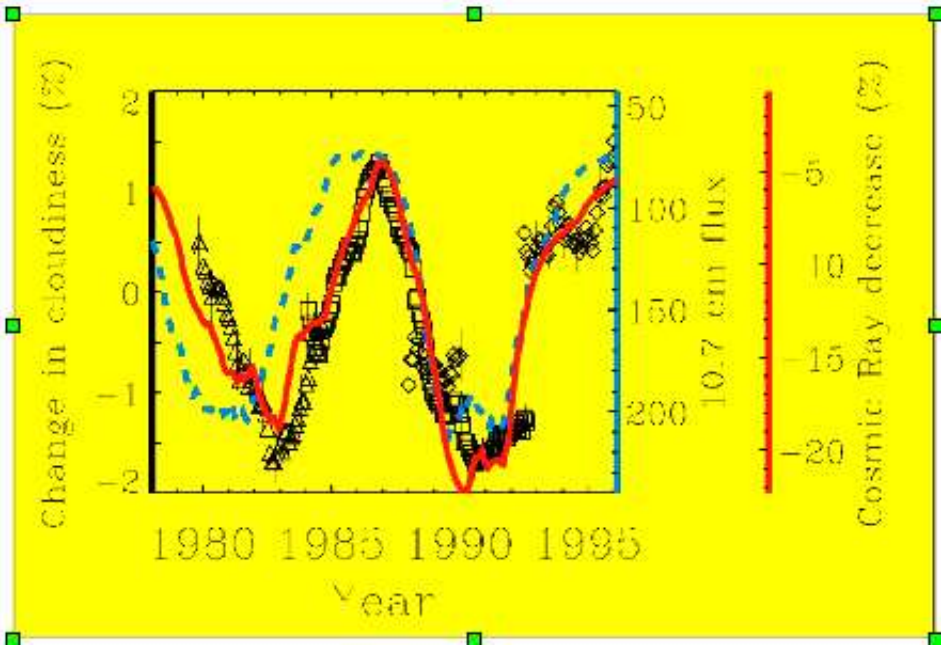

Figure 9: Composite figure showing changes in Earth's total cloud cover, obtained from satellite observations. The solid red curve is showing the cosmic ray flux and the broken blue curve shows the solar flux (from Svensmark 1998).

warming attributed to low cloud cover changes over the past century according to their model. During the same period the estimated heating from increased CO_2 emissions is 1.5 Wm^{-2} and the changes in solar irradiance received at Earth are 0.4 Wm^{-2} (see Lockwood & Stamper 1999).

According to these calculations high solar activity and a consequent decrease in cosmic radiation could be as important for the global warming as the growth in greenhouse gases.

4 Discussion

Parts of the theories I have been discussing in this paper are quite controversial. For example, some of the theories of Svensmark et al. are clearly opposing the Kyoto Protocol and the common opinion among scientists about the greenhouse gases and the global warming. Consequently they have been criticized too.

Are there any significant correlation between the flux of cosmic particles and the observed cloud cover?

We got no complete coverage of the global cloud changes. This would be necessary for reliable estimates in the total change of cloudiness. What Svensmark et al. got is rather a mixture of several different sets of data from different satellites. As such the observations are questionable and other observers are reporting different results.

Besides that there are also other aspects of the cloudiness that varies - for example the optical thickness and water content! These aspects might be of importance too.

Is there any physical process connecting the flux of cosmic particles and the development of clouds?

Today statements about the climatologic effects of changes in ionization are pure speculation, according to EP Ney. We got no observed evidence of ions affecting condensation

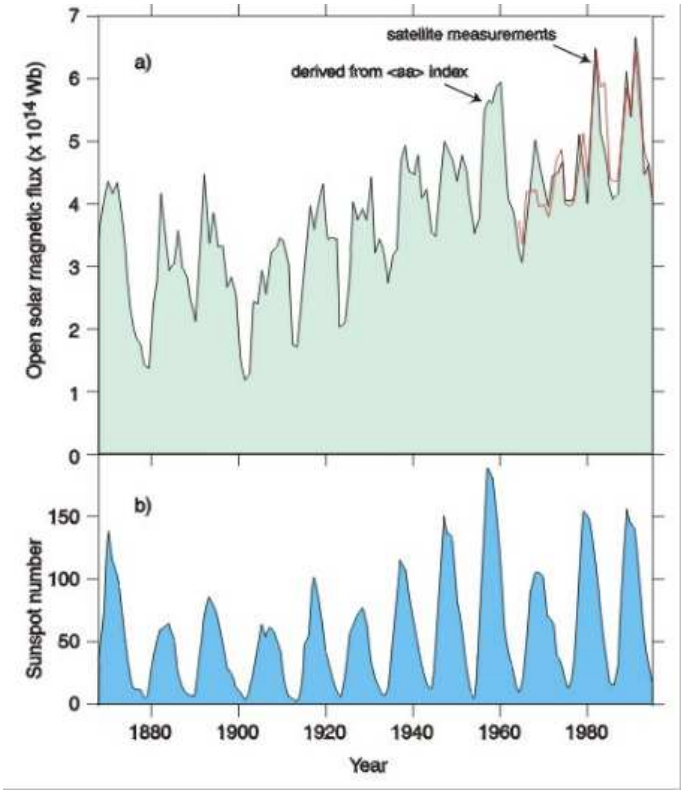

Figure 10: Annual mean of the coronal source magnetic flux. The dark blue curve at the bottom shows the variations of the smoothed sunspot number (from Lockwood & Stamper 1999).

of clouds in the atmosphere.

Noticeable is that the ionization EP Ney reported were obvious at an altitude of 10 km and more, but the influence on cloudiness that Svensmark et al reported occurred at a lower altitude then 3 km. Svensmark et al. are arguing that high-energetic secondary particles from the cosmic ray ionization penetrate the atmosphere to depths of low cloud formation.

Besides that the observed changes in cloudiness Svensmark et al. are building their theory on could for example be a result of El Niño or of volcanic eruptions. There are quite a few studies of such events (from Jorgensen & Hansen 1999).

5 Conclusions

The recorded variation in direct solar flux some 0.2 % under de last century is well known and its effects have been thoroughly studied. Effects of the UV-radiation from the Sun are also well known. The new ozone molecules that are formed in higher layers of the atmosphere have been studied as well as effects of the ozone as a greenhouse gas.

But not much is known about indirect effects of solar wind and cosmic radiation on the Earth's climate. Effects on cloudiness, temperature and rain are not known and neither do we know the mechanisms coursing these effects. Neither do we know much about the effects of the increasing magnetic flux from the Sun.

Consequently it is important to support the conclusions from the 2001 workshop at CERN. This workshop found the scientific indications of a cosmic ray-cloud link both interesting and important. They also agreed that plausible microphysical mechanisms

might exist but that their significance not yet is known. The workgroup then agreed on the urgent need to perform controlled laboratory measurements to test the cosmic ray-cloud link in a particle beam at CERN (from CERN 2001).

Acknowledgments

I would like to thank Alessandro Romeo for a very inspiring course. I also want to thank my colleague and editor Daniel Persson for his support during the preparation of this paper. Without his help there had been no paper.

References

CERN 2001, Conclusions Workshop on ion-aerosol-cloud interactions.

Jorgensen T. S., Hansen W. A., 1999, Journal of Atmospheric, Solar and Terrestrial Physics, 62, 79-80

Lean J., 1997, The Sun's Variable Radiation and its Relevance for Earth

Lockwood, Stamper, Wild, 1999, Nature, 432

Marsh N., Svensmark H., 2000, Phys. Rev. Lett., 85, 5004-5007

Marsh N., Svensmark H., 2001, Space Science Review, 85

Marsh N., Svensmark H., 2002, Journal Geo. Res., 102

Neff et al., 2001, Nature, 411, 290 - 293

Ney E. P., 1959, Nature, 378-380

Svensmark H., Friis-Christensen E., 1997, Journal of Atmospheric, Solar and Terrestrial Physics, 59, 1225-1232

Svensmark H 1998, Ph. Rev. Lett., 81, 5027-5030

Svensmark H 2000, Space Science Review, 93, 155-166

IPCC 2002, International Symposium on Climate Change 2001, `http://www.ipcc.org`

Dynamics of the Crab Nebula

Oscar Agertz

Göteborg University
SE-41296 Göteborg, Sweden
(gu99osag@dd.chalmers.se)

*

Abstract

In this article I present the neutron star and some simple physics explaining the features of pulsars. The magnetic dipole model as well as the aligned rotator model is also presented to give a feeling for the energy release mechanisms of pulsars. The remarkable Crab nebula is presented, with special attention to the recent observations of the extraordinary structures in the synchrotron part of if. A simple structural explanation is attempted as well as a discussion of what we can and cannot model with todays physics.

1 Introduction

Compact objects like white dwarfs, neutron stars and black holes are formed when normal stars die, i.e. when most of their nuclear fuel have been consumed. At this point, the radiation pressure from nuclear fusion is too small to withstand a gravitational collapse. The end-stage of a star is completely determined by its initial mass and the amount of mass ejected into the *InterStellar Medium* (ISM) in the later stages of a stars life.

These new formed objects differ very much from normal stars. White dwarfs do not posses the ability to produce thermal pressure due to fusion, instead they are supported by the pressure of degenerate electrons. Neutron stars, which are more massive than white dwarfs, are also supported by this degenerate fermi pressure, but here it is supplied by neutrons since both species are fermions (spin-$\frac{1}{2}$ particles). Black holes are completely collapsed objects from stars too massive to render sufficient fermi pressure to withstand absolute gravitational collapse. Table 1 shows some characteristics of these objects.

This article will focus on the special features of a rapidly rotating neutron stars (often referred to as pulsars) and their incredible dynamical effects on their environment. A full theory of these objects and their interaction with the ISM is far from fully developed and the topic includes a lot of complicated astrophysics. Just the way of observing them spans over a wavelength spectrum of well over ten orders of magnitude.

The theory of extremely compact matter is a very active field of research in which there is a lot of physics that is in need to be developed.

*Hot Topics in Astrophysics 2002/2003, Alessandro B. Romeo, Christoffer Petersson, Daniel Persson & Oscar Agertz (Eds.), Chalmers University of Technology and Göteborg University, 2003.

The rest of this paper is organized as follows. In Sect. 2 the neutron star is described in simple terms followed by a thorough treatment of the pulsar in Sect. 3. Sect. 4 treats the Crab nebula and its pulsar and presents the state of the art observations followed by the conclusions in Sect. 5.

Table 1: Distinguishing Parameters of Compact Objects

Object	Mass (M)	Radius (R)	Mean Density (g cm^{-3})	Surface Potential (GM/Rc^2)
Sun	$M_\odot$	$R_\odot$	$\sim$1	$\sim 10^{-6}$
White dwarf	$\leq M_\odot$	$\sim 10^{-2} R_\odot$	$\leq 10^7$	$\sim 10^{-4}$
Neutron star	$\sim$1-3$M_\odot$	$\sim 10^{-5} R_\odot$	$\leq 10^{15}$	$\sim 10^{-1}$
Black hole	Arbitrary	$2GM/c^2$	$\sim M/R^3$	~ 1

2 Neutron Stars

The properties of *Neutron Stars* (NS) are derived from the sophisticated *Equation Of States* (EOS) for of degenerate matter above the *neutron drip* (densities above $\rho \sim 4.3 \cdot 10^{11}$ g cm^{-3}). The point of an EOS is to describe how the pressure varies with density, i.e. $P(\rho)$, and the different models that exist differs in terms of density range, composition, way of interaction and in treatment of many-body theory. As an example, the *Baym-Bethe-Pethick* EOS treats the degenerate matter as e^- and n together with an equilibrium nuclide, while the *Bethe-Johnson* EOS composes a neutron star of n, p and a collection of more exotic particles (Λ, $\Sigma^{\pm,0}$, $\Delta^{\pm,0}$ and Δ^{++}) appearing in the ultra-high density range.

With an appropriate choice of EOS together with the general relativistic equation of hydrostatic equilibrium (the *Oppenheimer-Volkoff* equations) and the demand of stability renders neutron stars with the parameters as seen in Table 1. The interior might differ in some ways but the important thing is that we find a large population of charged particles. Fig. 1 displays two possible configurations calculated from completely different EOS (*Reid* and *TNI*).

3 Pulsars

3.1 History

The theory of more compact objects than white dwarfs was predicted by a number of theorists like Baade and Zwicky in 1934 and Oppenheimer and Volkoff in 1939. The very proof for the existence of these stellar configurations came by the observations of pulsars.

In 1967, a group of Cambridge astronomers lead by Anthony Hewish detected objects in deep space emitting periodic pulses of radio waves. The connection between neutron stars and pulsars was not at all clear at the time. The first argument that pulsars where fast rotating neutron stars with surface magnetic fields of around 10^{12} G was put forward by Gold in 1968. He presented the physics behind the remarkable stability of the pulse period as well predicted the small increase of period as the pulsar loses rotational energy. This phenomenon was discovered in the Crab pulsar the next year. This energy loss was later to be connected to the energy required to power the impressive nebula.

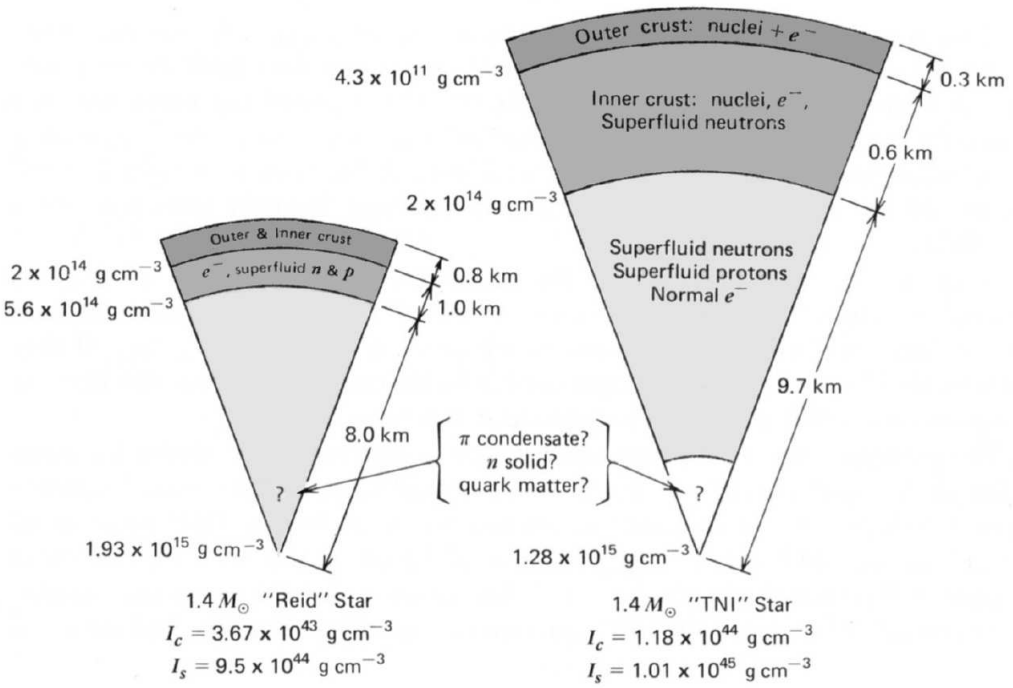

Figure 1: Cross sections of $1.4\,M_\odot$ Reid and TNI stars displaying different regions of particles made possible by different EOS. The moments of inertia for the entire crust as well as the super-fluid interior are also shown. Notice that the Reid-star is much more contracted due to allowed pion condensation. (From Shapiro 1983)

There are three observational facts that make rotating neutron stars the only candidate for pulsars:

1. Observed pulsar periods lie in the range 1.6 ms to 4.3 s.

2. Pulsar periods increase very slowly, but never displays a decrease except for the case of stellar glitches due to star-quakes.

3. Pulsars are extremely good clocks, with some measured periods with 13 significant digits.

The clock mechanism can be explained by studying three possible origins: rotation, pulsation or binary system. A bound system satisfies a relation of the form

$$\Omega^2 R \sim \frac{GM}{R^2}, \tag{1}$$

where R is the radius. Considering a white dwarf rotating at break-up velocity and inserting the relevant numbers into Eq. (1) it is easy to see that the period $P = \frac{2\pi}{\Omega} \geq 1\,\mathrm{s}$ is too large. This does not only rule out a rotating white dwarf but also a white dwarf binary system because R here becomes the orbital radius which is even larger in that case.

By making very simple arguments based upon the size of the systems, pulsations are ruled out as well for both white dwarfs and neutron stars. Using Eq. (1) we see that a binary neutron star system very well can be adjusted to fit into the observed period range of argument 1 above. This configuration would however lead to an enormous amount of

radiation in form of gravitational waves which would lead to a spin-up of the system, making argument 2 invalid. A black hole does not have a structure which to attach a precise periodic emitter as is observed.

These arguments of simple physics only leaves the rotating neutron star as a reasonable explanation for pulsars.

3.2 Observations

Before taking a look of the physical models of pulsars and the spectacular Crab nebula with its pulsar, there are a some observed properties that is essential to point out. The physical background will not be presented here, just the bare observations. It is important to know that the pulse mechanism is not clearly understood as of today, although the standard "lighthouse" model seems to be the way the radio pulses gets their periodicity.

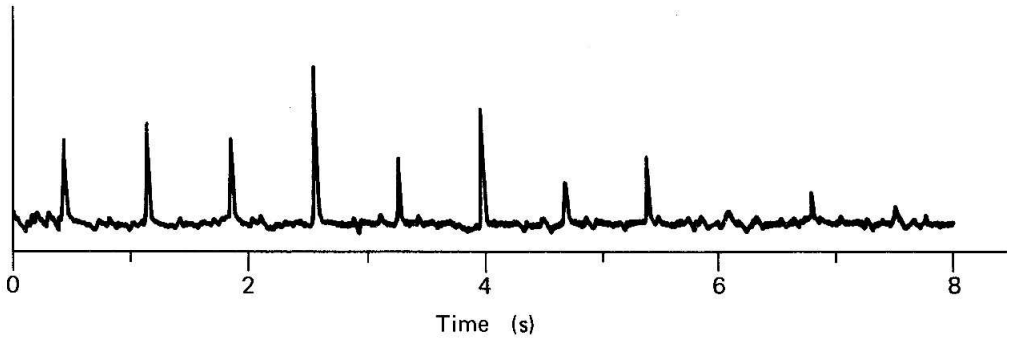

Figure 2: Spectrum showing the individual pulses from one of the first pulsars discovered, PSR 0329+54. The pulses show an interval of 0.714 s at a frequency of 410 MHz (from Shapiro 1983).

- **Pulse Shapes and Spectra:** Today we have observations of around 1000 pulsars. All of these exhibit broadband radio emission in the form of periodic pulses. An example of this can be seen in Fig. 2. On a time-scale of $\leq$ 1 ms the pulse shape becomes quite complex. Most of the observed pulsar spectra consists of not only the characteristic pulses, but also of two or more sub-pulses. The sub-pulses display complicated structures on time-scales as short as 10 μs.

  The specific radio intensity is a power law, $I_\nu \propto \nu^\alpha$. A typical value for α is around -1.5 for $\nu < 1$ GHz and seems to decrease for higher frequencies. A typical intensity for at 400 MHZ is $\sim$ 0.1 Jy.

  It is also an interesting fact that the pulsars show a high degree of linear polarization. Some pulsars manage to deliver 100 %, perhaps suggesting a large portion of synchrotron radiation (this will be treated below) . The amount and position angle is if the linear polarization frequently vary with time across the pulse. Circular polarization is much less important.

- **Periods:** As mentioned before, pulsar periods lie between 1.558 ms (PSR 1937+214) and 4.308 s (PSR 1845-19). Today there are also a few more observed pulsars at possibly even lower periods, suggesting different composition in the pulsar core (very possibly quark matter).

The increase of period-time is also a well observed fact. Typically, we expect to see $\dot{P} \sim 10^{-15}$.

Also observed, especially in the well documented Crab and Vela pulsars are the *spin-ups*. This is sudden shortenings of the pulsar periods. For the Crab pulsar these are typically $|\Delta P|/P \sim 10^{-8}$ accompanied by a period derivative increase, $|\Delta \dot{P}|/\dot{P} \sim 2 \cdot 10^{-3}$, which decays away in around 10 days. The explanation for this is thought to be either instabilities in the superfluid interior or more likely *star-quakes*. These quakes crackings of the stars crust, causing a decrease of angular momentum. This is rather complicated and is best treated in modern research journals.

3.3 The High Rotation and the Strong Magnetic Field

The reason for the impressive rotational speeds and strong magnetic field from the pulsar is, to first order, simply a matter of conservation of physical quantities. A typical strength of a pulsar magnetic field is $\sim 10^{12} - 10^{14}\,\mathrm{G}$. Since the flux is conserved during the contraction of a star to the neutron star stage we know that:

$$R_{\mathrm{i}}^2 B_{\mathrm{i}} = R_{\mathrm{ns}}^2 B_{\mathrm{ns}}. \tag{2}$$

This is a rather good first-order approximation to get an order of magnitude estimation without taking the complicated inner configurations into account.

A main sequence star (the red giant phase will be scaled equally) typically has $R \approx R_\odot$ and $B \sim 100\,\mathrm{G}$. A neutron star has $10^{-5} R_\odot$. Using Eq. (2) we find that the magnetic field should be $\sim 10^{12}\,\mathrm{G}$ which nicely explains the enormous surface magnetic fields of the pulsars.

The rotation is explained in similar fashion, invoking the conservation of angular momentum (and not taking deformation into account):

$$R_{\mathrm{i}} V_{\mathrm{i}} = R_{\mathrm{ns}} V_{\mathrm{ns}} \Leftrightarrow P_{\mathrm{ns}} = P_{\mathrm{i}} \left(\frac{R_{\mathrm{ns}}}{R_{\mathrm{i}}} \right)^2. \tag{3}$$

From Eq. (3) we see that an almost stationary star will reach very high rotation from the dependence of radius-ratio squared. A red giant star typically has a period of the order of several days which suggest a NS period of the order of ms, just as expected.

It is important to know that milli-second pulsars usually only can form as binaries. Pulsars without companions have larger periods. The Crab has an impressive period of 0.0331 s and it does not seem to have had any companion to have spun it up before or after its creation.

3.4 The Magnetic Dipole Model

To be able to discuss the observed properties of pulsars it is appropriate to discuss a very simple pulsar model: the *magnetic dipole* model. This model explains how pulsar emission is derived from the kinetic energy of the rotating neutron star.

There are two situations to study. First the *oblique rotator* where the star rotates uniformly, at a frequency Ω, in vacuum, and possesses a magnetic dipole moment, $\mathbf{m}$, oriented at an angle α to the rotation axis. Secondly we will look at the non-vacuum model, also called the *aligned rotator*.

Independent of the internal field geometry, the magnetic dipole field, $B_{\rm p}$, at the magnetic pole of the star is

$$|\mathbf{m}| = \frac{B_{\rm p} R^3}{2}, \tag{4}$$

where R is the radius of the star. A configuration of this kind radiates energy at a rate (see Rybicki 1979):

$$\dot{E} = -\frac{2}{3c^3}|\ddot{\mathbf{m}}|^2. \tag{5}$$

Considering the configuration of the system we know that

$$\mathbf{m} = \frac{1}{2} B_{\rm p} R^3 (\mathbf{e}_{\|} \cos\alpha + \mathbf{e}_{\perp} \sin\alpha \cos\Omega t + \mathbf{e}'_{\perp} \sin\alpha \cos\Omega t). \tag{6}$$

Here $\mathbf{e}_{\|}$ is the unit vector parallel to the axis of rotation and $\mathbf{e}_{\perp}$ and $\mathbf{e}'_{\perp}$ are the mutually orthogonal unit vectors perpendicular to $\mathbf{e}_{\|}$. From this we find the important relation

$$\dot{E} = -\frac{B_{\rm p}^2 R^6 \Omega^4 \sin^2\alpha}{6c^3}. \tag{7}$$

This is a fundamental equation for pulsars leading to several important consequences. We can directly see that the energy carried away by radiation originates from the rotational energy. Because

$$E = \frac{1}{2} I \Omega^2, \tag{8}$$

where I is the moment of inertia, we see that

$$\dot{E} = I \Omega \dot{\Omega}. \tag{9}$$

It is also appropriate to define a characteristic age T at the present time by

$$T \equiv -\left(\frac{\Omega}{\dot{\Omega}}\right)_0, \tag{10}$$

which can be related to the present age of the pulsar by

$$t \approx \frac{T}{2}. \tag{11}$$

For a full derivation, see Shapiro 1983. This set of equations gives us a good set of predicted numbers to compare with observations. For the Crab pulsar, the number for T is 2486 years, rendering an age of 1243 years. This is in remarkably good agreement with the actual age of 949 years. In astronomy it is the usual case to measure and calculate the age of objects. Concerning the Crab it is historically confirmed that the year of the supernova is 1054 which allows us to test the theory of pulsars.

The model described in this section quantifies some important physical parameters concerning the Crab pulsar. Given a $M = 1.4\,M_{\odot}$, $R = 12$ km and $I = 1.4 \cdot 10^{45}$ g cm^2 with Eqs. (8) and (9) we see that

$$E = 2.5 \cdot 10^{49}\,\text{erg}, \dot{E} = 6.4 \cdot 10^{38}\,\text{s}^{-1}. \tag{12}$$

It is interesting that only I is required to obtain this, and this by using the EOS for neutron stars. Also very interesting is that the value for $\dot{E}$ is very close to the observed

kinetic and radiation energy of about $5 \cdot 10^{38}\,\mathrm{erg\,s^{-1}}$. This is how a pulsar maintains its high energy output for over 900 years. It is important to notice that the radiation in the radio pulse only performs about $10^{31}\,\mathrm{erg\,s^{-1}}$.

From Eq. (7), and thus assuming the dipole model, we find that the Crab has

$$B_\mathrm{p} = 5.2 \cdot 10^{12}\,\mathrm{G} \tag{13}$$

which corresponds nicely to the approximation made in Sect. 3.3.

Electromagnetic radiation is not the only source of energy loss. A very important factor is gravitational radiation (in form of gravitational waves, see Persson 2003). Gravitational waves occurs due the non-spherical geometry of the NS which is the case in reality. To get a more correct age and acceleration rate for the Crab it is more appropriate to use a combined model in the form

$$I\Omega\dot{\Omega} = -\beta\Omega^4 - \gamma\Omega^6 \tag{14}$$

where β and γ are constants for the different types of radiation (see Shapiro 1983). An easy calculation shows that at present day

$$\dot{E}_\mathrm{gw} = 1.4 \cdot 10^{38}\,\mathrm{ergs^{-1}}, \dot{E}_\mathrm{em} = 5.1 \cdot 10^{38}\,\mathrm{ergs^{-1}}. \tag{15}$$

Today the energy release from the Crab pulsar consists of mainly electromagnetic radiation but also an important contribution from gravitational waves. Initially however the GW-contribution dominates entirely, displaying numbers like

$$\dot{E}_\mathrm{gw} = 2.9 \cdot 10^{48}\,\mathrm{ergs^{-1}}, \dot{E}_\mathrm{em} = 3.9 \cdot 10^{45}\,\mathrm{ergs^{-1}}. \tag{16}$$

Fortunately for the simplicity of describing pulsar-nebula dynamics, the high gravitational radiation only dominated the total energy loss for the first 130 years.

3.5 Nonvacuum Pulsar Models: The Aligned Rotator

To get an understanding for the physics in the new observations presented in Sect. 4 the aligned rotator-model will be described here. It is essential that Eq. (7) holds for models other than the simple one presented in Sect. 3.4 (oblique, vacuum and magnetic dipole).

It is appropriate to use the theory developed by Goldreich and Julian in 1969, that predicts that strong electric fields parallel to the pulsar's surface will rip off charged particles from the star and thus create a dense magnetosphere. In reality, most of the e^+e^--plasma comes from the high-energy photons of the electric filled due to pair-production. This region will probably not be filled with particles coming from the material ejected by supernova, because all of this has been violently dragged along the blast-wave.

The main idea of the theory is that the particles inside the *light cylinder* will co-rotate with the pulsar. This cylinder is an imaginary region defined to be within the the radius extending to the point where the rotation reaches the speed of light, see Fig. 3. It is thus defined by

$$R_\mathrm{c} \equiv \frac{c}{\Omega} = 5 \cdot 10^9 P\,\mathrm{cm} \tag{17}$$

rendering $R_\mathrm{c} \approx 1.7 \cdot 10^8\,\mathrm{cm}$ ($\approx 6 \cdot 10^{-11}\,\mathrm{pc}$) for the Crab Pulsar ($P = 0.0331\,\mathrm{s}$). When particles approach this limit they become highly relativistic and it is widely believed that much of the high frequency radiation is emitted near the light cylinder. It is easy to see that whether or not the dipole field is aligned with the rotation axis, plasma near the pulsar will carry away sufficient angular momentum and energy to be responsible for the braking of rotation observed.

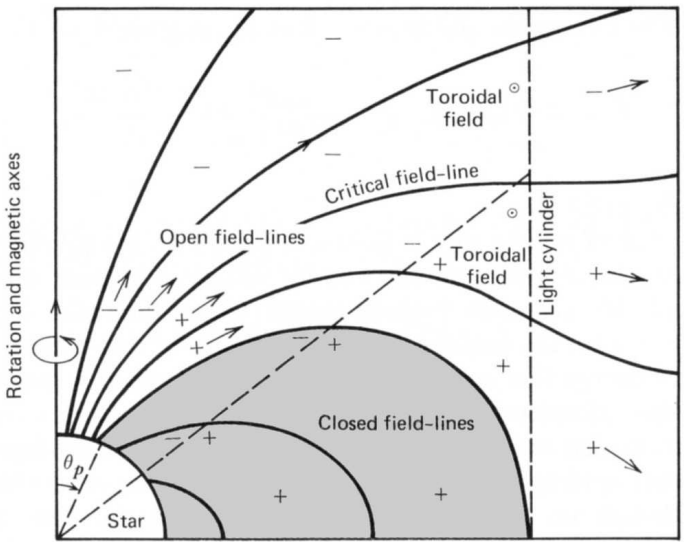

Figure 3: Model of a magnetosphere of a pulsar with parallel magnetic and rotation axes. The particles that are attached to the closed magnetic-field lines co-rotate with the star and form the co-rotating magnetosphere. Open-field lines (the lines that cross through the light cylinder) are deflected back to form a toroidal field component. These are the lines that the particles stream out along. The figure also shows the critical field line where we find the same electrical potential as the exterior ISM. This is where the regions of positive and negative charges are met (from Shapiro 1983).

As mentioned above, the reason for the relativistic wind responsible for the powering of the nebula and the dynamical structures is due to the fact that the enormous electrical field tears of particles from the pulsar surface and creates a plasma around the star - a magnetosphere. This was beautifully shown by Goldreich and Julian in 1969. By assuming that a spinning neutron star has an aligned dipole external magnetic field

$$\mathbf{B}^{(\mathrm{out})} = B_{\mathrm{p}} R^3 \left(\frac{\cos\theta}{r^3} \mathbf{e}_{\hat{r}} + \frac{\sin\theta}{2r^3} \mathbf{e}_{\hat{\theta}} \right) \tag{18}$$

and the interior of the star satisfies

$$\mathbf{E}^{(\mathrm{in})} + \frac{\boldsymbol{\Omega} \times \mathbf{r}}{c} \times \mathbf{B}^{(\mathrm{in})} = 0. \tag{19}$$

The exterior electrical field can be calculated by demanding a continuous crossing through the crust (for a full derivation, see Shapiro 1983).

The important end result is that the magnitude of the electrical field parallel to $\mathbf{B}$ at the surface is approximately

$$E_{\|} \sim \frac{R\Omega}{c} B_p \sim 2 \cdot 10^8 P^{-1} B_{12} \,\mathrm{volt\,cm^{-1}}, \tag{20}$$

where B_{12} is the magnetic field strength in units of 10^{12} G. This enormous field will impart a force to both electrons and ions at the surface that greatly exceeds the gravitational force. As an example, the ration between electric force and gravitational force for a proton is of the order of 10^9!

An important remark for the coming chapters is that the region where the field lines close beyond the light cylinder, called the "open magnetosphere", receives particles permanently lost to the star. This process is fundamental for the pulsar wind and all of the observations (jets, pulses, shocks etc.).

4 The Crab Nebula

Table 2 presents some important facts regarding the Crab nebula.

Table 2: Astronomical facts about the Crab Nebula (from NASA/HST/ASU/J. Hester et al.).

Object Name:	Crab Nebula - NGC 1952 - M1
Object Description:	Supernova Remnant
Time of Creation:	1054
Position (J2000):	R.A. 05h 34m 32s, Dec. +22ą 00ń 52ńń
Constellation:	Taurus
Distance to Earth:	About 6000 light-years (1850 pc)
Dimensions:	The central part is 0.9 pc while the whole nebula is about 4 pc in diameter (depending on observed wavelengths)

4.1 Historical Summary

In July or August of the year 1054, Chinese astronomers witnessed a new star being born in the sky above the southern horn of the constellation Taurus. Its light was bright enough to be seen during the day for over a month, and it remained visible in the evening sky for over a year. This "new star" was a supernova explosion of a star of over $10\,M_{\odot}$. It is not often a event like this can be so directly observable. In the nine centuries since, astronomers have witnessed only two comparable cataclysms in our Galaxy: the supernova explosions of 1572 and 1604.

This specific supernova was forgotten for more than 600 years until the invention of telescopes, which revealed fainter celestial details than the human eye could detect. In 1731, English physicist and amateur astronomer John Bevis observed the strings of gas and dust that form the nebula. While hunting for comets in 1758, Charles Messier spotted the nebula, noting that it had no apparent motion. The nebula became the first entry in his famous *Catalogue of Nebulae and Star Clusters*, first published in 1774 (hence becoming M1). Lord Rosse named the nebula "the Crab" in 1844 because of its tentacle-like structure resembled the legs of the crustacean. In 1939 astronomers concluded that the nebula was expanding and determined that it probably originated from a point source.

In 1948 scientists discovered that the Crab was emitting among the strongest radio waves of any celestial object. Baade noticed in 1954 that the Crab possessed powerful magnetic fields, and in 1963, a high-altitude rocket detected X-ray energy from the nebula. New discoveries kept coming. Finally in 1968 the ends where tied together thanks to the discovery of the emitted bursts of radio waves 30 times per second. The "point source" had to be a pulsar!

The Crab nebula with its pulsar is one of the most studied objects in the sky and has in many ways become the prototype of pulsar physics.

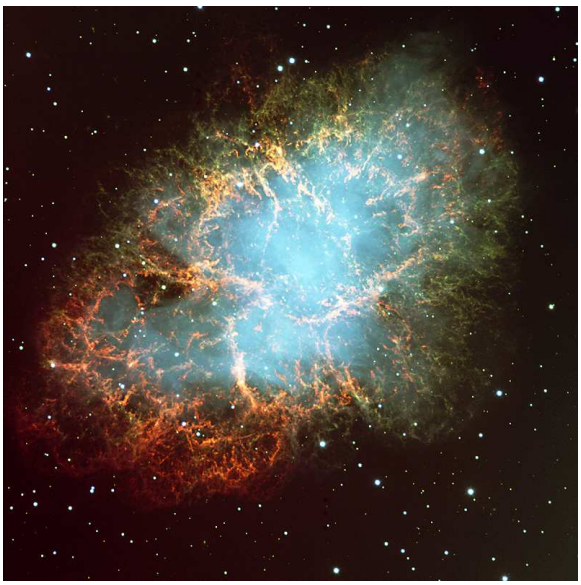

Figure 4: The VLT (UT2 + FORS) view of the Crab Nebula. This is a composite of images in B(blue), R and S II(red), taken in November 1999 (from the ESO homepage).

4.2 Observations

Fig. 4 is one of the most famous ones of the Crab nebula, and a frequently used PR picture for Very Large Telescope (VLT). The red indicates that electrons recombine with protons to form neutral hydrogen. This occurs in the original supernova ejecta from 900 years ago. The blue indicate the synchrotron radiation from the inner nebular part. Although we see important structures such as large filaments and giant O III-regions, there is not much information about the central part, other than a large amount of synchrotron radiation.

As the observational technology has improved, more details of the Crab has been discovered. The last decade has been particularly interesting, in particular the observations shown in Fig. 5 done by Hester et al. (1995) using Hubble Space Telescope (HST). These are a sort of milestone for the physics behind the complicated mechanisms in the Crab. In a series of pictures (see Fig. 6) Hester et al. could present dramatic changes in the appearance of the central regions of the nebula. These include wisp-like structures that move outward away from the pulsar at about half the speed of light, as well as a mysterious "halo" which remains stationary, but grows brighter then fainter over time. Also seen are the effects of two polar jets that move out along the rotation axis of the pulsar. The most dynamic feature seen is a small knot that "dances around". This seems to be a shock front in one of these polar jets. These observations have been presented in the form of a movie, aptly named "The Crab Movie" (1996).

Even though the discoveries made in 1995 showed amazing structures, this was not enough. The most important observations to date was made a couple of years ago. The inner region of the Crab Nebula around the pulsar was observed with Hubble on 24 occasions between August 2000 and April 2001 at 11-day intervals, and with Chandra on eight occasions between November 2000 and April 2001. The Crab was observed with Chandra's Advanced CCD Imaging Spectrometer and Hubble's Wide-Field Planetary Camera. The results were also presented in form of a spectacular movie (2002) which really displays the dynamics of the Crab Nebula and also introduces a strange X-ray ring close to the pulsar (see Fig. 7). The next section will describe and explain some of these

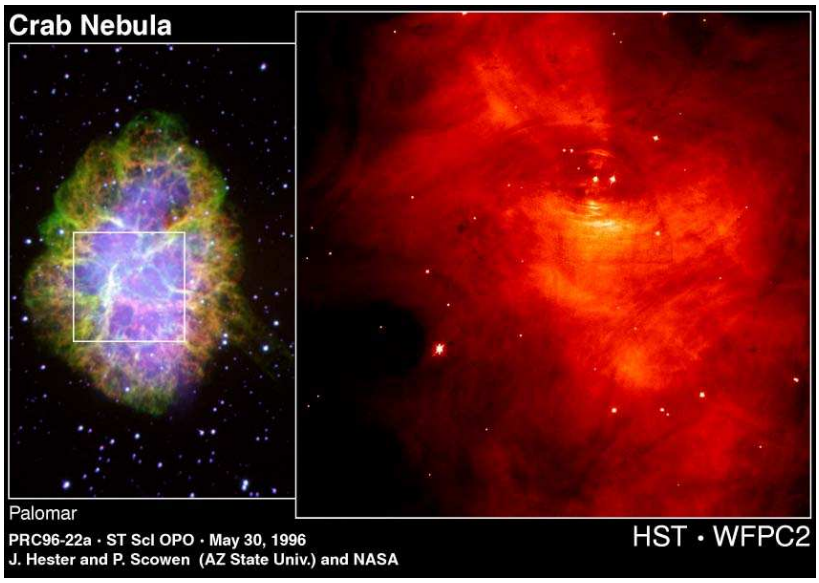

Figure 5: The HST (WFPC2) view of the Crab Nebula and its center taken in May 1996. More complicated structures are observed here than ever before (from the SEDS homepage).

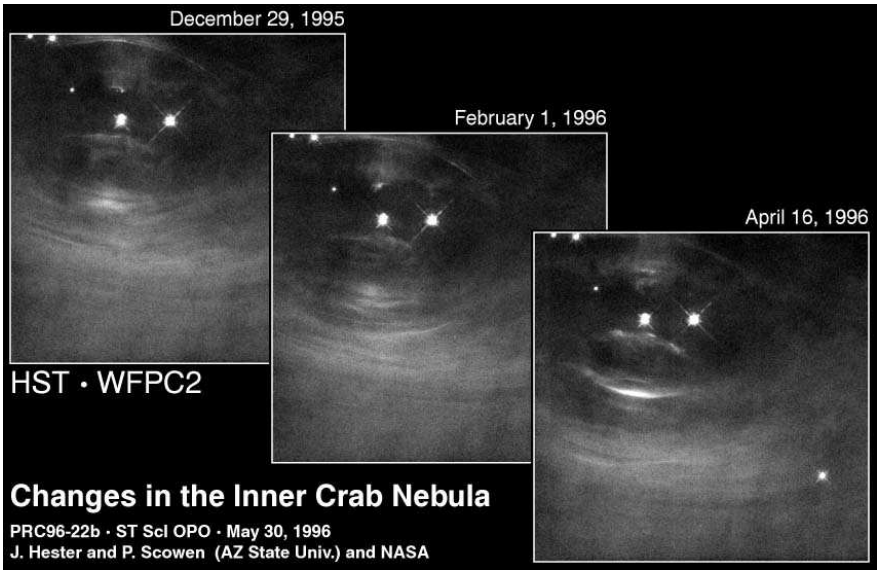

Figure 6: The sequence of images taken by HST in 1995/96 revealing new information about the dynamics of the system (from the SEDS homepage).

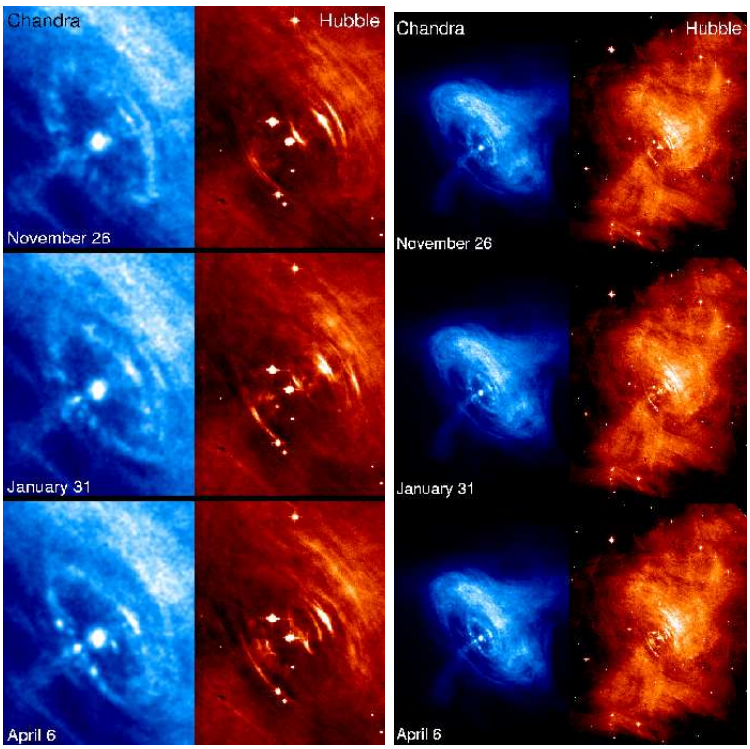

Figure 7: The HST and Chandra images taken in 2000 and 2001. The images to the left shows a close up of the inner part of the nebula, clearly showing the pulsar with the characteristic X-ray ring and jets. To the right we see the same scene a bit further away, displaying how much less extended the X-ray emitting part is (from the HST homepage).

features.

4.3 The Structure and Dynamics

The observations presented in the last section are really amazing, and poses a large set of problems in astrophysics. Many of the things observed are hard or even not possible to explain, as of today.

The Crab Nebula is a prototype of something called a *pulsar wind nebula* (PWN), also referred to as a *plerion*. A sort of definition for this epithet is when a relativistic pulsar wind is confined by external pressure, and thus becomes observable. In a large context it is appropriate to separate between three types of PWN (see Gaensler 2003).

- The young ($\sim 1000\,\mathrm{yr}$) pulsar, inside its equally young supernova remnant (SNR). The pulsar's wind is significantly over-pressured with respect to the cold ejecta inside the SNR shell. It is important to know that the pulsar still has a very high rotational speed. Because of all this, the PWN will expand in all directions and one typically observes a spherically looking radio and X-ray nebula with a pulsar in the center.

- At later stages ($\geq 10000\,\mathrm{yr}$) the story gets more complicated. Reverse shocks are created that fills the SNR with shocked ejecta. The collision between this reverse shock and the expanding PWN is dramatic, and the pulsar nebula gets compressed, resulting in a completely new appearance in the radio and X-ray bands. There is a large possibility for the pulsar to be offset from the center of the PWN.

- A much later times, the SNR has dissipated and the pulsar has acquired a speed of typically about 500 km s^{-1} moving through the ISM. Because of the pulsar wind ram pressure, a bow shock is created, which can be observed in many parts of the spectrum. A fascinating example of this can be seen in Fig. 8.

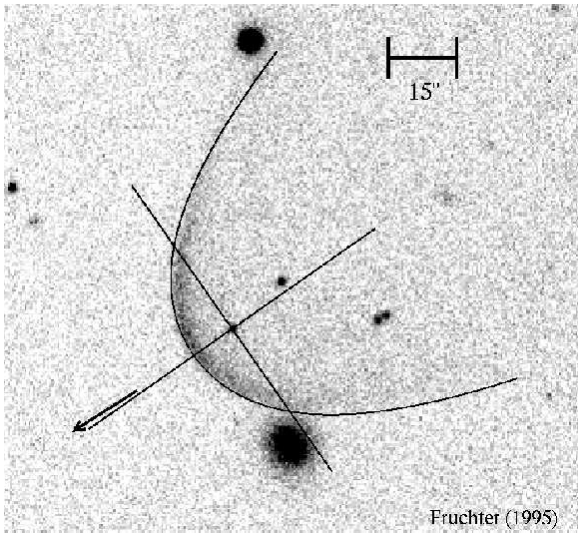

Figure 8: An H-α image of the nebula about PSR 0437-4715. The distance between the pulsar and the leading edge of the H-α emission is about 1400 AU. On this scale, the position of the pulsar is indistinguishable from that of its white dwarf companion (picture from Andrew S. Fruchter).

Here the Crab clearly falls into the first category (naturally, by being the prototype). The morphology of the Crab in the most simple model be approximated as consisting of three zones, as seen in Fig. 9. In the first zone, the pulsar wind zone, a relativistic wind is launched from the pulsar light cylinder (as phenomenologically described in Sect. 3.4). At a certain distance r_w from the pulsar the pressure of the out-flowing wind gets balanced by the outside pressure (gas and magnetic). This is the *termination shock*, and this is where the particles are thermalized up to X-ray emitting energies. Downstream of the shock front (in a shock wave co-moving frame) the flow decelerates and generates the observable synchrotron emission together with the important *inverse Compton radiation* (ICS). ICS is when a photon is scattered on a relativistic e^+ or e^-. This process boosts a low-energy photon's energy by up to a factor of γ^2 (lorentz factor squared). The effectiveness is lessened at high energies due to quantum effects (the cross-section $\frac{d\sigma}{d\Omega}$ is reduced). One should notice that the particle speed, as mentioned before, is measured to be in the order of $0.5\,c$, which impart that the pulsar wind must be of ultra-relativistic speeds (models show a Lorentz factor of $\gamma \sim 3 \cdot 10^6$).

This rather simple picture has gained much validity thanks to the Chandra images, where it was shown that the termination shock does not have an overall spherical geometry but a toroidal one. This equatorial outflow is of course expected, especially when considering Eq. (7).

The termination shock seen in the Crab is a very complicated region (see Fig. 7). The observed inner X-ray ring that remains stationary consist of a large amounts of "variable knots" that probably originates from unstable synchrotron cooling. There is no doubt however (see Hester 2002) that this is structure is associated with the shock that turns the "cold" ultra-relativistic pulsar wind into a synchrotron-emitting plasma.

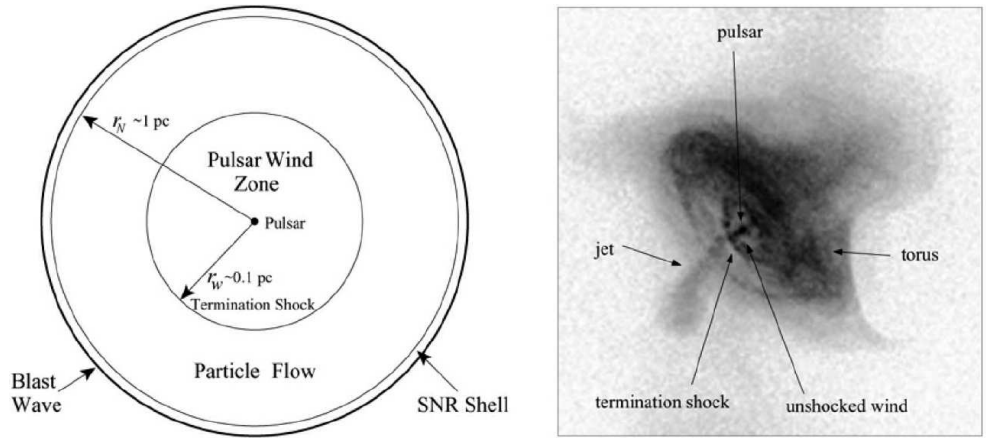

Figure 9: Illustration of PWN (left), compared to a *Chandra* X-ray image of the Crab nebula (right) (from Gaensler 2003).

The moving wisps seems to originate from this X-ray ring, but this is not conclusive. These are likely to be moving magnetic flux tubes undergoing unstable synchrotron cooling, thus explaining their variability. In a way, these are a kind of internal shock structure of gyrating ions. There are currently computer simulations being carried out to try to simulate this behavior conducted by Hester et al.

The next observed feature is the jets. From the Chandra observations it has been possible to see that these jets are perpendicular to the equatorial plane of the system, and hence the jets do not originate from the magnetic poles, but from the pulsar rotation axis. The jets are also seen in the visible light. Along the line of ejection, there are features seen moving with speeds up to $\sim 0.4\,c$. Many diffuse structures can be seen as well as discrete bow-shock features. Hester at al. has derived that the jet must be pushing into the surrounding synchrotron nebula at speeds of $\sim 0.03\,c$. As seen from the aligned rotator model, particles are carried away from the pulsar but the way they get accelerated remains unsolved.

The outer regions of the nebula is more easy to model. Here we see (Fig. 4) the gas of the synchrotron nebula being pushed into the freely expanding gas of the SNR. Radiative shocks are created as well as the well observed *Rayleigh-Taylor* (R-T) instabilities creating the beautiful filaments. Inside of the shocks, we observe, among other elements, the O III skin regions.

In simple terms, the things presented in this section is the current state of the research. There are many things to clarify.

5 Conclusions

The subject of late stellar evolution spans over many parts of physics and introduces many unsolved problems. From stellar and statistical physics we can put mass limits for the end-stage of a dying star. This article has concentrated on the neutron star which normally has a mass of $\sim$1.4-3 $M_\odot$. The interior of the neutron star varies a bit depending of usage of equation of state, but the density and size generated is fairly equal.

Due to conservation of angular momentum and magnetic flux, the neutron star will gain a very high rotation and enormous magnetic fields, introducing interesting observables. These objects are pulsars and one of the most interesting ones is the Crab pulsar

with surrounding nebula. The amazing observations made by modern telescopes, especially HST and Chandra, show remarkable structures and dynamics that poses new ways of probing he physics of pulsar and high-energy astrophysics.

As presented in the last chapters, the observations are as of today more complicated than the state of the art models can handle. This has revived the field greatly, and there are a large number of articles published continuously on the subject. Many of them are trying to model the Crab using advanced computer simulations in *MagnetoHydroDynamics* (MHD), managing to account for some of the features (the X-ray torus) but there is much to be done.

Early work done by Kennel and Coroniti (1984), are sufficient to explain the expansion velocity of the PWN, luminosity, and synchrotron radiation spectrum (from optical wavelengths to X-rays) if the wind consists of electrons and positrons. These are just integral observables and can in no way explain the complicated inner structures.

We have clearly entered a new era of pulsars and their winds. These discoveries are hard problems to solve, but this time there is enough data available to address them in a proper manner.

Acknowledgments

I acknowledge the support by my supervisor Alessandro Romeo of the department of Astronomy and Astrophysics at Chalmers University of Technology and Göteborg University. I also want to thank Ulf Torkelsson (same department) for the discussions concerning pulsar winds.

References

Bogovalov S.V., Khangoulian D.V., 2002, Astronomy Letters, 28, 373

European Space Organisation Homepage, `http://www.eso.org`

Gaensler B. M., 2003, in XXI Symposium on Relativistic Astrophysics: Texas in Tuscany, in press (astro-ph/0303427)

Hester J.J, Scowen P.A., Sankrit, R., et al., 1995, ApJ, 448, 240

Hester et al., 2002, ApJ, 577, L49-L52

Kennel C.F., Coroniti F. V., 1984, ApJ, 283, 710

Mori et al., 2002, ASP Conf. Ser. 271: Neutron Stars in Supernova Remnants, 157

Persson D., 2003, this volume, p. 42

Rybicki G. B, Lightman A. P., 1979, *Radiative Processes in Astrophysics*, John Wiley & Sons, New York

Shapiro S. L., Teukolsky S. A., 1983, *Black Holes, White Dwarfs and Neutron Stars - The Physics of Compact Objects*, John Wiley & Sons, New York

Shibata et al. 2002, in The Universe Viewed in Gamma-Rays, in press (astro-ph/0211621)

Sollerman J., Flyckt V., 2002, The Messenger, 107, 32

Students for the Exploration and Development of Space, `http://www.seds.org`

Swaluw E., 2003, to appear in A&A, in press (astro-ph/0303661)

The Crab Movie (1996), `http://hubblesite.org/newscenter/archive/1996/22/`

The Crab Movie (2002), `http://hubblesite.org/newscenter/archive/2002/24/`

The Hubble Space Telescope Homepage, `http://www.hubblesite.org/`
Weisskopf et al., 2000, ApJ, 536, L81

Binary Pulsars

Gautam Narayan

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(gautam@etek.chalmers.se)

*

Abstract

This paper gives an insight into the wonderful world of binary pulsars. Apart from being objects of general interest what makes them so important are their remarkable properties, which give a perfect validation of the general theory of relativity.

1 Introduction

Binary systems are important not only because they enable us to determine stellar masses but because they are very common and may help us in understanding stellar evolution. A large proportion of millisecond pulsars have been observed in binary systems. Almost all neutron stars associated with X-ray sources are members of binary systems. Also the high percentage of single radio pulsars may find an explanation in the disruption of the binary systems during supernova explosions. It has been observed that millisecond pulsars are quite common in globular clusters and so are low mass X-ray binaries. This supports the theory of millisecond pulsars evolving from these binaries. The X-ray binary systems, the radio binary pulsars and the millisecond pulsars which may be the end product of the evolution of a binary system together give a new understanding of the phenomenon of mass transfer in a binary system which appears to be quite common to the life of many stars in this universe.

In this paper we start with a brief review of the end points of stellar evolution in Sect. 2. In Sect. 3 we review the basics of pulsars and classify them into various types. Further in Sect. 4 we explain some important concepts like the Roche lobe and mass transfer and in Sect. 5 we study the different evolutionary scenarios of the different classes of binary pulsars. In Sect. 6 we take a look at the evolution of binary pulsars in globular clusters and finally in Sect. 7 we present an important application of the binary pulsars.

*Hot Topics in Astrophysics 2002/2003, Alessandro B. Romeo, Christoffer Petersson, Daniel Persson & Oscar Agertz (Eds.), Chalmers University of Technology and Göteborg University, 2003.

2 End Points of Stellar Evolution

Depending on the mass of the remnant core which is left behind after the complete evolution, a star can end its life in one of the following three states:

- White Dwarf
- Neutron Star
- Black Holes

A white dwarf is formed if the mass of the core is less than about $1.4\,M_\odot$ (solar masses). This limiting mass of $1.4\,M_\odot$ is known as the Chandrasekhar Limit. Chandrasekhar studied that in the white dwarf matter would be compressed to very high densities, something like a million times that of water. Thus 1 liter of white dwarf matter will weigh a thousand tonnes (see Narlikar 1999). If the mass of the remnant core is more than $1.4\,M_\odot$ then one can imagine the even higher densities that would exist. At such high densities electrons are forced to penetrate into the nucleus and they combine with protons to form neutrons. Normally neutrons would decay into protons and electrons. However, because of the extremely high densities and the uncertainty principle, all the low-energy electron states are full. Hence formation of new electrons is prevented due to the Pauli's exclusion principle and so the neutrons do not decay.

The number of neutrons in the nucleus keeps on increasing and their binding energies keeps on decreasing. This process continues for some time after which neutrons begin to leak out from the nucleus and some time later the nuclei disappear completely. Matter now consists mainly of neutrons along with minute amounts of electrons and protons. Thus a neutron star is formed. When a massive star reaches the end of its evolution and explodes as a supernova, the simultaneous collapse of its core will not necessarily stop at the density of a white dwarf. If the mass of the collapsing core is larger than Chandrasekhar mass, the collapse continues to form a neutron star. The high magnetic fields are a result of the conservation of magnetic flux. The upper mass limit for neutron stars has been fixed at $2\,M_\odot$. Only stars with masses below this limit can maintain their equilibrium as neutron stars.

If the mass of the core is more than $2\,M_\odot$ then the object collapses indefinitely and a black hole is formed.

3 Pulsars

Pulsars are extremely fast rotating neutron stars having a very high magnetic field. A neutron star has two polar axes: a rotational axis and a magnetic axis just like the Earth; the difference being that in case of the Earth both the rotational and magnetic axes are almost aligned but the axes of a neutron star may be pointing in totally different directions. The rotating star has a swarm of charged particles like electrons in its atmosphere. As the star rotates so does its atmosphere. The charged particles in the outer part of the atmosphere rotate very fast, almost at the speed of light. Such fast rotating particles are known to generate electromagnetic radiation in the presence of a magnetic field (see Narlikar 1999). The radiation is highly beamed and is known as synchrotron radiation. So if the earth happens to be in the sweep through area of the pulsar beam we will get pulses of radiation each time the beam sweeps past us. See Fig. 1 for a schematic picture of this model known as the lighthouse model of a pulsar. It is this pulsating radiation,

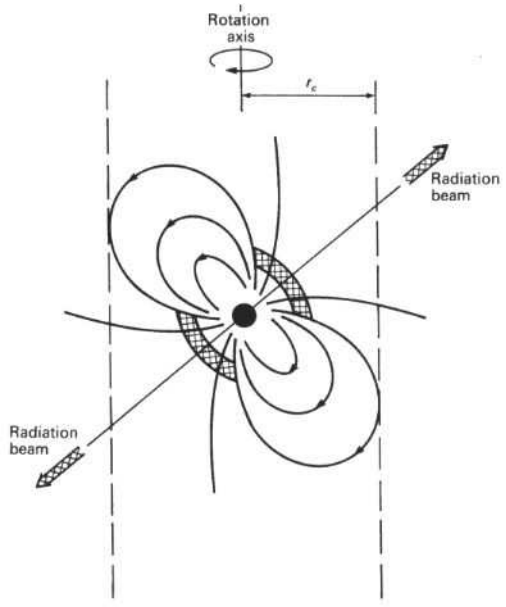

Figure 1: The lighthouse model of a radio pulsar: A rapidly rotating central neutron star with a strong magnetic field, inclined to the rotation axis with radio emission emanating from the magnetic poles (from Lyne & Graham-Smith 1998).

which is observed using radio telescopes and hence the name pulsar, which is the acronym for pulsating radio source.

The first pulsars were discovered using radio telescopes. With the advent of X-ray astronomy in the 1970s many X-Ray pulsars where discovered. In 1974 the first binary radio pulsar was discovered and in 1982 the first millisecond pulsar was discovered. It was the discovery of the millisecond pulsar and the X-ray pulsars that generated interest in the study of binary pulsars. Let us see why.

What happens to the spinning neutron star as it keeps radiating for a long time? The obvious answer is that as time goes on the spinning pulsar slows down and its pulse period increases. Thus the pulse period of the pulsar is an indication of its age. If we divide the observed pulsar period by double the rate at which the period is observed to decrease we can get the age of the pulsar (see Narlikar 1999). A useful diagram using this concept is the $P - \dot{P}$ diagram. This is shown in Fig. 2. The $P - \dot{P}$ diagram can be thought of as the equivalent of the Hertzsprung-Russell diagram for pulsars. $\dot{P}$ is the derivative of the period with respect to time. According to this diagram the young pulsars which have small spin periods occupy the upper right hand part of the diagram. The faster the period the further on top of the diagram they lie. However many millisecond pulsars were observed which seemed to defy this concept. The millisecond pulsars have very low spin periods in the order of a few milliseconds (anywhere between 1.5 and 30 ms). They have even lower slow down rates. Their ages are in the order of a few billion years and they occupy the lower left region of the diagram. Many such pulsars were observed to be in binary systems and they are shown as encircled in the $P - \dot{P}$ diagram. The only explanation for the existence of millisecond pulsars was recycling of a pulsar and this concept made use of the existence of pulsars in binary systems. We shall describe later in detail the process of recycling of pulsars. Also studies in X-ray astronomy proved that X-ray pulses and bursts occurred due to the accretion of matter on compact objects and this fact also pointed to the existence of neutron stars in binary systems. Pulsars can be

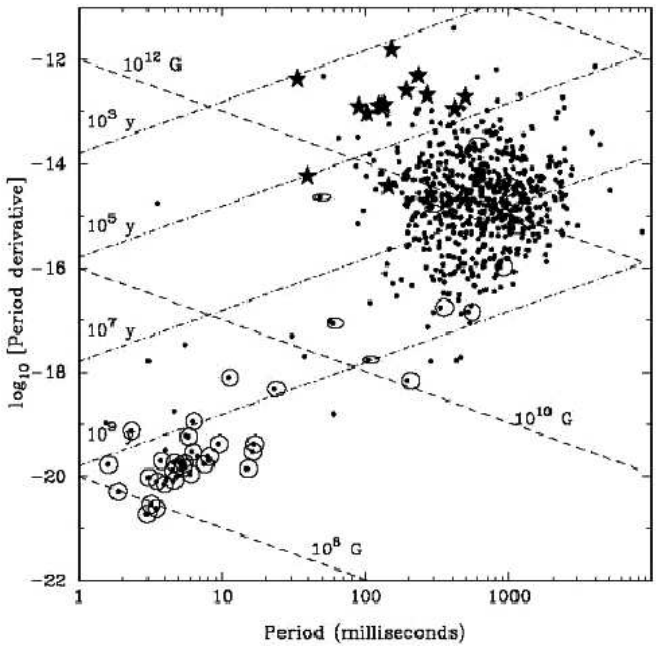

Figure 2: The $P - \dot{P}$ diagram (from Lorimer 2001).

classified as shown in Fig. 3.

4 Some Important Concepts in a Binary System

4.1 The Roche Lobe

In binary systems there is a tidal interaction between the two objects constituting the binary. The surfaces of constant gravitational potential (or equipotential surfaces) which are spherical for single stars become distorted into a tear drop like shape in the presence of a companion star. There is a critical point on the line joining the centers of the two stars where the net force felt by a test particle is zero. This point is known as the inner Lagrangian point. Fig. 4 illustrates this concept. The potential surface which passes through the inner Lagrangian point L_1 is common to both the stars and is called the Roche lobe which has the shape of the inverted numeral 8. The material within each stars Roche lobe is confined to the gravitational force of that star only. However material on or beyond the Roche lobe is attracted more strongly to the companion star. If a star expands and extends beyond its own Roche lobe, then its outer layer is attracted more strongly to the companion star and thus it starts to transfer mass to its companion.

4.2 Mass Transfer

The most important issue in a binary system is that of mass transfer and the crux of it can be understood as follows. When a star evolves in a binary system and expands and fills its Roche lobe it starts to transfer mass to the companion star. Now the mass transfer changes the angular momentum of the donor star and according to the law of conservation, angular momentum has to be conserved. So there is a change in the orbit size of the binary to compensate for the loss of angular momentum. When the mass transfer takes place from a more massive to a less massive star then there is a shrinking effect on the orbit of the binary system and when there is mass transfer from a lower mass

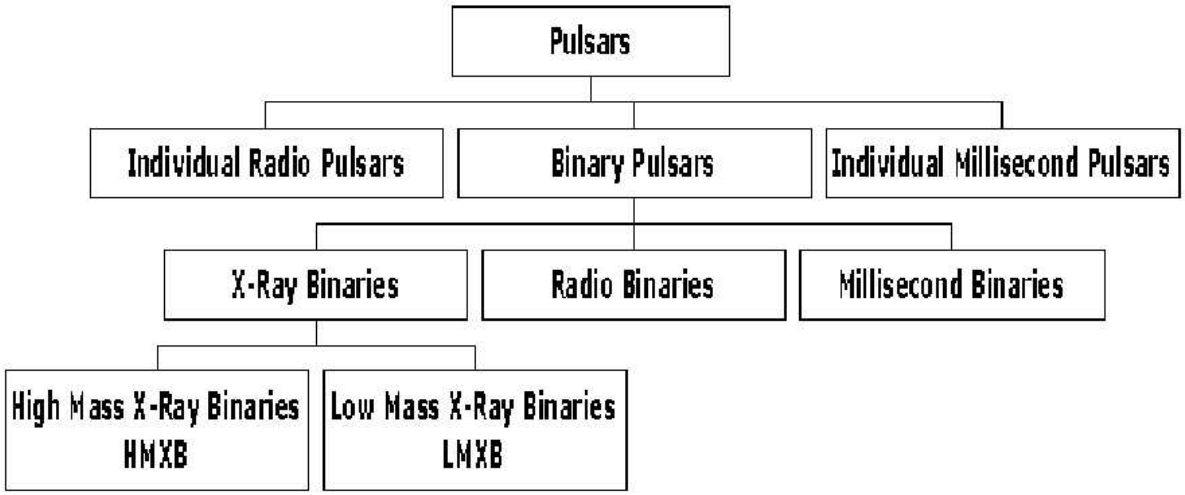

Figure 3: The classification of pulsars

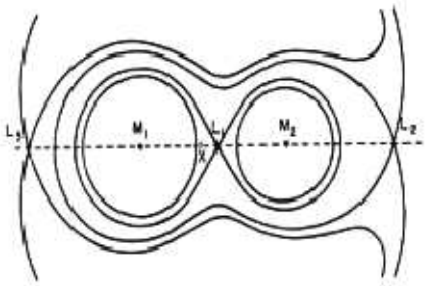

Figure 4: The point L_1 is the inner Lagrangian point and L_2 and L_3 are the outer Lagrangian points (from Bhattacharya & Van den Heuvel 1992).

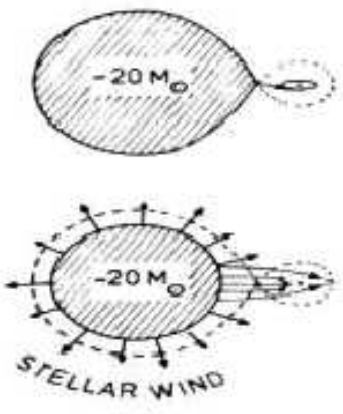

Figure 5: The top figure shows mass transfer via Roche lobe overflow one can observe the accretion disk which is formed in this case. The bottom figure shows mass transfer via strong high velocity stellar wind from a companion still inside its Roche lobe (from Bhattacharya and Van den Heuvel Binary 1992).

star to a higher mass star then there is a increase in the orbit size. An obvious point is that in the case of mass transfer from a more massive to a less massive star, as the mass transfer takes place the orbit of the binary system shrinks causing a further expansion of the star beyond its Roche lobe. This in turn causes more mass loss and a further shrinking of the orbit. Clearly there is a runaway effect and mass transfer continues on a dynamical scale until the more massive star has lost so much mass that it becomes the less massive star compared to its companion.

Mass transfer can be of two types. The first type, which is common in most low mass binary systems is mass transfer via Roche lobe overflow which generally results in the formation of an accretion disk. The second type is mass transfer via stellar winds which are more common in high mass binary systems where the system contains a massive star usually of O, B or Be type. Fig. 5 shows the two different types of mass transfer. In high mass systems mass transfer can take place both through Roche lobe overflow and stellar winds but in low mass systems its only via Roche lobe mass transfer.

5 Evolutionary Scenarios

Fig. 6 shows the various possible evolutionary scenarios of binary pulsars. In a binary system the more massive of the two stars will evolve first. It will expand and overflow its Roche lobe, transfer mass to its companion, undergo a supernova explosion and form a neutron star which is detected as a pulsar. Now there are two possibilities. If the supernova explosion has not disrupted the binary system then we have a normal binary radio pulsar. If the supernova explosion destroys the binary system then we have a young individual pulsar and a runaway star. Here we should mention that from the virial theorem it follows that the binary system gets disrupted if more than half the pre-supernova mass is ejected from the system during the explosion (see Lorimer 2001). The binary radio pulsar exists for a long time until the neutron star is almost dead.

However, by this time the companion star has evolved and it expands and overflows its Roche lobe and starts transferring mass to its companion-the neutron star. An accretion disk is formed around the neutron star and along with the mass transfer there is also a

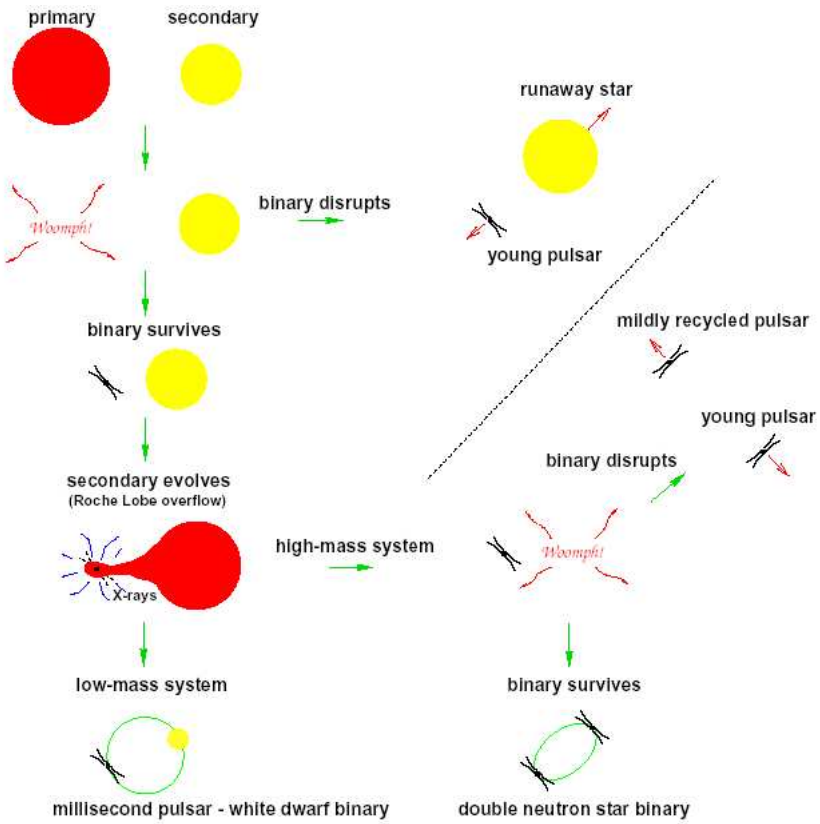

Figure 6: A cartoon depicting the various possible evolutionary scenarios of pulsars (from Lorimer 2001).

transfer of momentum which spins up the almost dead pulsar to very high velocities thus recycling it or giving it a new lease of life. Also the accreting mass falls on the surface of the neutron star and X-rays are generated. This system is now an X-ray Binary. The further evolution depends on the mass of the system. If the companion is massive then it will be called as a *high mass X-ray binary* (HMXB). The companion star may evolve and undergo a supernova explosion. If the explosion does not disrupt the binary system then we have a double neutron star binary. If the system is disrupted by the supernova explosion then we have a young pulsar and a mildly recycled pulsar. In case the companion was of low mass then the system is called a *low mass X-ray binary* (LMXB) and this evolves into a binary system consisting of a millisecond pulsar and a white dwarf.

5.1 Evolution Of High Mass X-Ray Binaries (HMXB)

HMXBs can be of two types. The first type is one where the massive star is a O or B type star. Such HMXBs are called as standard HMXBs. The other type of HMXB is the B-emission type HMXB. In this the massive star is of Be type. The other important difference between the two types is in the type of mass transfer. In the standard HMXB the mass transfer can be both due to stellar winds and Roche lobe transfer. But in case of Be-star X-ray Binary the mass transfer is by episodic ejection of matter in the orbital plane or by a steady stellar wind or by both methods but never by Roche lobe. It is believed that Be-star X-ray binary may be the reason behind the observed X-ray transients. The episodic ejection of mass from the Be star may explain the large, erratic transient outbursts that have been observed. Fig. 7 shows the various evolutionary

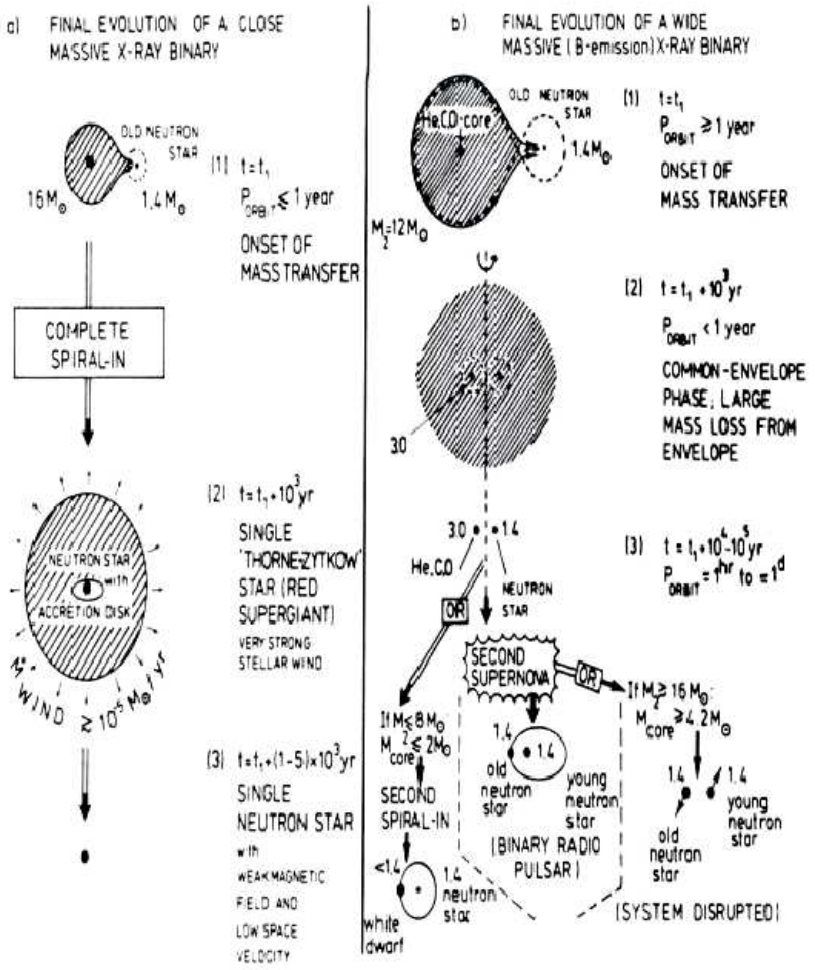

Figure 7: The various possible evolutionary scenarios of HMXBs (from Bhattacharya & Van den Heuvel 1992).

scenarios of HMXBs.

In case of the close HMXB system due to the runaway effect explained earlier the Roche lobe may overflow and engulf the companion star forming a common envelope. The neutron star then spirals into the core of the companion. The envelope is then ejected due to the energy liberated from the accretion of mass from the common envelope to the neutron star. Finally a single mildly recycled radio pulsar is left. In the case of a wide Be type X-ray Binary the envelope is ejected during the spiral-in and a close binary system is left behind. If the companion of the neutron star was initially more massive than 8-12 $M_\odot$, the remnant core undergoes a supernova explosion and if the system survives the explosion then a binary radio pulsar is formed else a runaway old neutron star and a runaway young neutron star are formed.

5.2 Evolution of Low Mass X-Ray Binaries (LMXBs)

We shall now study the evolution of a LMXB, which is depicted in Fig. 8. In this LMXB the mass transfer takes place from the low mass sub-giant with a degenerate helium core of 0.24 $M_\odot$ to a neutron star of around 1 $M_\odot$.

As we have seen earlier the mass transfer in a LMXB takes place through the Roche lobe and there is an increase in the orbit size. So finally we are left with a recycled pulsar

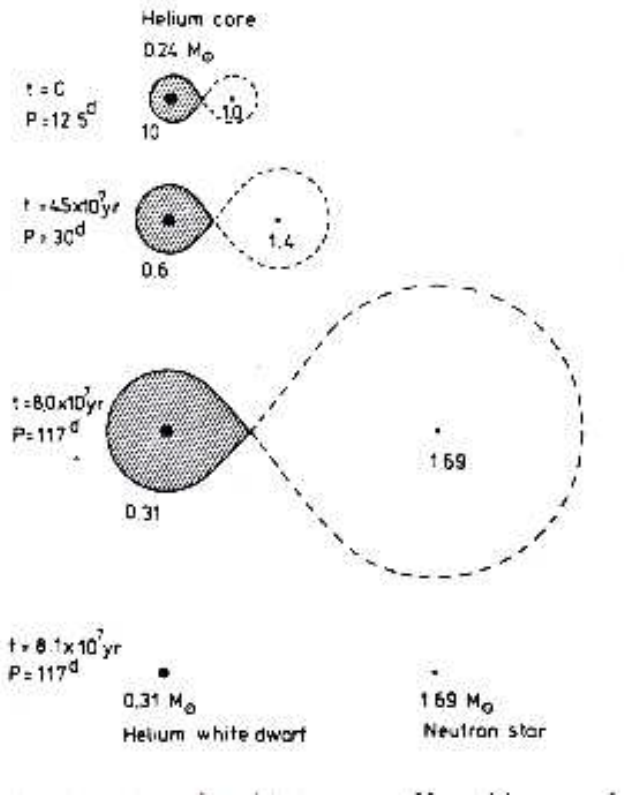

Figure 8: Evolution of a wide LMXB such as the Cygnus X-2 into a wide radio-binary pulsar with a circular orbit and a low mass helium white dwarf companion (from Bhattacharya & Van den Heuvel Binary 1992).

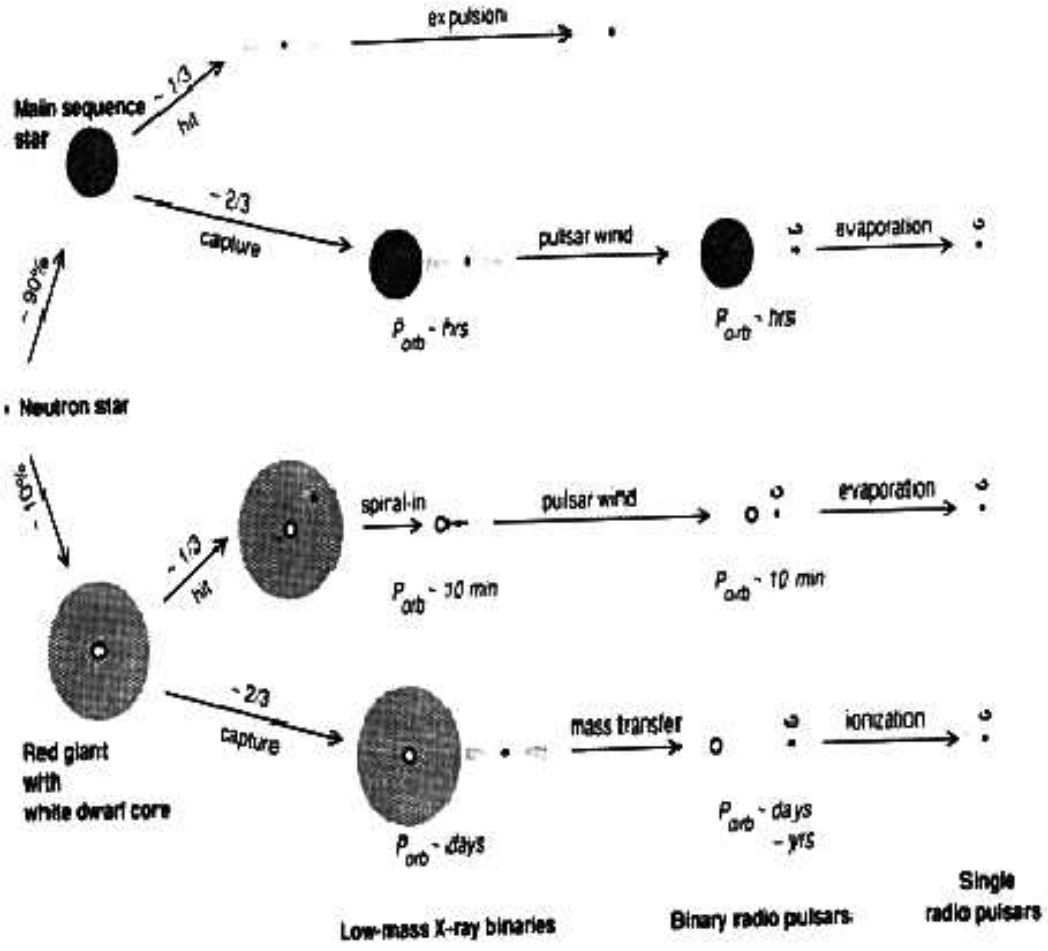

Figure 9: Formation and evolution of neutron stars in a globular cluster by tidal capture and direct collision (from Bhattacharya & Van den Heuvel 1992).

in a binary system with the companion being a low mass white dwarf. Here we can state that the most of the millisecond pulsars and radio binaries are formed from such LMXB systems. LMXBs are not as violent as the HMXBs as there is no supernova explosion involved and thus the binary system is preserved. Also the mass transfer is entirely due to Roche lobe overflow and this creates an accretion disk which is efficient in recycling the old neutron star. However we should not forget the fact that a LMXB itself has initially evolved from a normal binary system which has survived a supernova explosion.

6 Binary and Millisecond Pulsars in Globular Clusters

About 20 % of the observed binary and millisecond pulsars have been located in globular clusters. This has led to a theory that globular clusters are the breeding grounds of binary and millisecond pulsars. Fig. 9 shows the various possible scenarios of pulsar evolution in a globular cluster.

Two mechanisms have been proposed for the formation of neutron star binaries neutron stars in globular clusters. They are tidal capture and exchange collisions. Globular clusters are large groups of closely packed stars. There may be around 10^5 to 10^6 stars

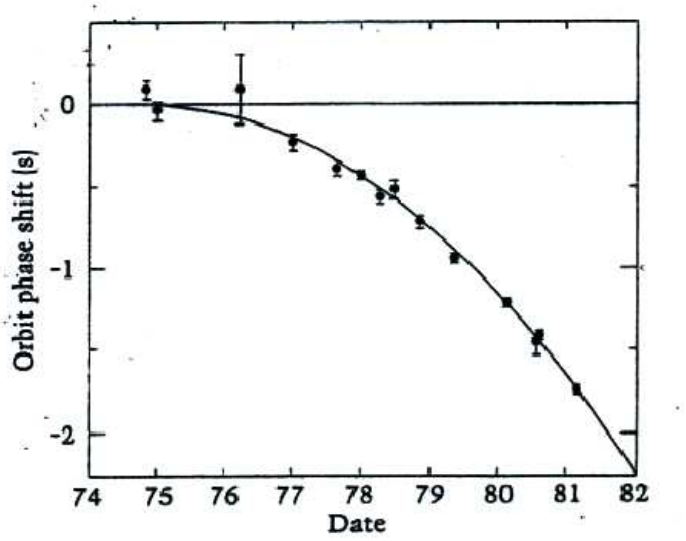

Figure 10: The relative positioning of the two members of a binary system can be quantified by the phase of the orbit, measured in seconds. As the orbit shrinks and the two members move faster round each other this phase changes. The phase change was measured for PSR 1913+16 and is given here as plotted by Taylor and Weisberg in 1984 (from Narlikar 1999).

in such clusters. In such a highly dense region there is a high probability of physical interaction between stars. A neutron star may close in on a target star and capture it into an orbit around itself thus forming a binary. This process is called tidal capture. The second method of binary formation is by exchange collisions. A neutron star may come across a binary at a distance close to the semi-major axis of the binary. This may lead to the formation of a three body system with complicated orbits of each body around the others. But this system would not be stable and after some time the lightest of the three is ejected. Thus the neutron star may take the place of an ordinary star. When mass starts to be transferred in the newly formed binary it becomes an X-ray binary. Direct collision of a neutron star with another star may also result in the formation of pulsars but in this case an individual pulsar will be formed. Collision of a neutron star with one member of a binary may also give rise to new binaries. Also two binary systems may interact with each other giving rise to complicated situations at first but ultimately leaving behind a new pair. Some single millisecond pulsars have also been observed in globular clusters. The reason for their origin is that they evolved from narrow LMXBs and after recycling the pulsar wind evaporates the very small companion.

7 Applications

The most important and interesting application of the binary pulsar is as a tool to prove the general theory of relativity and the existence of gravitational radiation. How is this possible? According to Einstein's General Theory of Relativity, in a binary system the two masses should emit gravitational waves. If they emit gravitational waves then they should loose energy and thus momentum and this should be compensated by shrinking of the orbit. This fact has been observed in binary pulsars. Fig. 10 shows the data from a binary pulsar PSR 1913+16. One can see from the figure that the pulsar follows the

curve predicted by the general theory of relativity almost exactly thus proving both the theory of relativity and the existence of gravitational waves.

8 Conclusions

Most of the binary and millisecond pulsars have probably originated through spin up recycling.

There are two classes of binaries from which binary radio pulsars evolve they are the high mass X-ray binaries HMXBs and the low mass X-ray binaries LMXBs.

The high occurrence of single radio pulsars may find an explanation in efficient evaporation of their companions due to the pulsar wind as well as collisions of neutron stars and cluster binaries. Disruption of the binary after a supernova explosion may be another reason.

Millisecond Pulsars and pulsars in globular clusters are products of evolution of LMXBs.

Acknowledgments

I would like to thank Alessandro Romeo for his constant support and encouragement without which this work would not have been possible. I also thank Oscar Agertz for his help in editing this document.

References

Lyne A.G., Graham-Smith F., 1998, *Pulsar Astronomy* Cambridge University Press

Lorimer D.R., 2001, Living Reviews in Relativity, 4, 5

Bhattacharya D., van den Heuvel E.P.J., 2001, Phys. Rep., 203, 1-124

Narlikar J.V., 1999, *Seven Wonders of the Cosmos*, Cambridge University Press

Gravitational-Wave Astronomy

Daniel Persson

Göteborg University
SE-41296 Göteborg, Sweden
(gu99dape@dd.chalmers.se)

*

Abstract

During the past few years, the sensitivity of gravitational-wave antennas has increased considerably and it is actually becoming realistic to believe that we are at the dawning of what will evolve into a completely new field of research: *Gravitational-Wave Astronomy*. In the spirit of this belief, I review the status of current research about theoretical and observational properties of gravitational waves. For the paper to be as self-contained as possible, I include a detailed derivation of Einstein's field equations in the weak-field limit, and show how these can be used to obtain wave-like solutions in general relativity. Two of the main projects concerned with the detection of gravitational waves will also be examined: LIGO (Large Interferometer Gravitational-wave Observatory) and LISA (Large Interferometer Space Antenna). Besides detecting known astrophysical sources, such as supernovae and binary pulsars, there is also a slight possibility that these antennas could detect a background of gravitational radiation analogous to the Cosmic Microwave Background (CMB). If it exists, this background necessarily originates from the Planck epoch and thus would provide new insight into the extreme physics that governed the very early universe.

1 Introduction

A few years after Einstein published his original paper on General Relativity, he realized that the theory predicted the existence of gravitational waves. Since then, a lot of research has been pursued in order to understand the properties of these waves, and, if possible, to detect them. The first observational breakthrough came in 1974 when Hulse and Taylor (they were later awarded the Nobel prize for their discovery) found that the decrease in angular velocity of the binary pulsar PSR1916+13 (see, Hulse & Taylor 1975) was in excellent correspondence with the predicted energy loss due to gravitational radiation (the theoretical calculation was first made already in the 50s, see Bondi 1957). In the 1960s, the astronomer John Weber built the first antenna with the purpose of directly detecting these elusive waves. He was, unfortunately, unsuccessful and even to this date

*Hot Topics in Astrophysics 2002/2003, Alessandro B. Romeo, Christoffer Petersson, Daniel Persson & Oscar Agertz (Eds.), Chalmers University of Technology and Göteborg University, 2003.

there has been no direct detection of gravitational radiation. There are however many interesting projects going on that will hopefully change this in a near future. Antennas are in operation in USA (LIGO), Italy (Virgo), Germany (GEO600), Japan (TAMA300) and Australia (ACIGA). There are also plans for the launch of a space-based observatory (LISA) in 2011. All of theses projects look very promising and, combined, should be able to probe a large part of the spectrum. In this paper I will restrict myself to the LIGO (Laser Interferometer Gravitational-wave Observatory) and LISA (Laser Interferometer Space Antenna) projects.

There are many similarities between the current state of gravitational-wave science and the field of neutrino physics in the mid 1950s (see, Hughes 2002): the theory is well defined and makes physical predictions, there is extensive indirect evidence, but a convincing direct detection has so far escaped us.

What are gravitational waves in a physical sense, and why are scientists so eager to detect them? The simple answer to the first part of this question is that they are ripples in spacetime, traveling at the speed of light (see Fig. 1). In general relativity, the gravita-

Figure 1: An artists impression of gravitational waves, or ripples in spacetime. (from LIGO homepage)

tional force is due to a curvature of spacetime and the emittance of a gravitational wave would therefore, intuitively, correspond to a propagating disturbance in this curvature. In a way, this also answers the second part of the question because it is precisely the fact that a gravitational wave is really a non-stationary gravitational field that makes them nearly undetectable. The gravitational force is by far the weakest of the forces of nature and thus interacts very weakly with ordinary matter. The same goes for a gravitational wave; it simply doesn't care about the antennas that we construct to halt its passage through earth. Although this sounds like a bad property, it is exactly this that makes gravitational waves so fascinating and causes scientists all over the world to engage in the growing field of *Gravitational-Wave Astronomy*. Because of the waves' ignorance of the contents in the universe, we can use them to probe areas which are impossible to reach with conventional (electromagnetic) astronomy. For example, the interior of a collapsing star has so far been shielded to us because of extensive scattering in the outermost regions of the star. Gravitational waves, on the other hand, will travel freely from the very center of the star, straight to our antennas on earth, thus providing us with information that would otherwise be lost forever.

When examining the theory of gravitational radiation, one finds many similarities with the Maxwellian theory of electrodynamics, which describes the propagation of electromagnetic waves. There are, of course, important differences, some of which I will take a closer look at later. One of the most profound examples I have already touched upon; gravitational waves are oscillations in spacetime *itself*, and electromagnetic waves are field

oscillations that propagate *through* this spacetime. With this in mind, it might actually be more fitting to compare gravitational waves to waves propagating over the surface of the ocean (see Hughes 2002). There is also an important property to notice about the wavelength of the waves. An electromagnetic wave, typically has a wavelength that is smaller than the size of the emitting source, making it possible to create an image of this source. In contrast, this is not possible with gravitational radiation because the wavelength is often larger than the emitting system. Gravitational-wave astronomy is therefore about observing large areas of the sky, instead of the tiny regions that conventional astronomy covers.

This paper is a review of what I consider to be the most interesting features of this new field of research. For readers who have a basic knowledge of general relativity but has no experience of gravitational-wave theory, I begin with a thorough derivation of the Einstein field equations in the weak-field limit and explain how to obtain a wave-like solution. The remainder of Sect. 2 will then continue to examine some of the important properties of these solutions, and how to interpret them in a physically relevant way. In Sect. 3, I take a closer look at some observational outlooks, focusing on one earth-based and one space-based antenna. Sects. 4 and 5 are concerned with the intriguing possibility of directly and indirectly probing the physics of the very early universe, using gravitational radiation.

2 Theory

Because of the vast distances to astrophysical sources, we are mainly interested in the behaviour of gravitational waves far away from their origin. This implies that the waves are propagating through vacuum in a nearly flat spacetime[1]. Therefore I will start with examining how general relativity behaves in the weak-field limit.

2.1 The Weak-Field Limit

The natural starting point is the full field equations of general relativity

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi T_{\mu\nu} \tag{1}$$

where $R_{\mu\nu}$ is the Ricci tensor, R is the Ricci scalar, $g_{\mu\nu}$ is the metric tensor and $G_{\mu\nu}$ is the Einstein tensor. We will now study how these equations simplify when the metric is given by the Minkowski metric $(-+++)$ plus a small perturbation, i.e

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \tag{2}$$

where $|h_{\mu\nu}| << 1$. Before we begin, it is interesting to look at what symmetries we can use to help us in the calculations. It turns out that there are two types of transformations (see e.g. Wald 1984; Weinberg 1972; Schutz 1985) that leave Eq. (1) invariant: *background Lorentz transformations* and *Gauge transformations*.

[1]One might argue that the antennas are sitting in the gravitational field of the surrounding solar system and thus the waves reaching us are not traveling through flat spacetime. This effect is however completely negligible compared to the enormous gravitational fields at the generating source.

For background Lorentz transformations we borrow the Lorentz transformation matrix from special relativity

$$(\Lambda^{\alpha'}{}_{\beta}) = \begin{pmatrix} \gamma & -v\gamma & 0 & 0 \\ -v\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3}$$

Even though we are not in a flat spacetime we can use this transformation to obtain some interesting properties for $h_{\mu\nu}$. When we apply this transformation to $g_{\mu\nu}$ we are really fixing a particular coordinate system where this transformation is valid. In other words, the field equations will not, in general, be invariant under Eq. (3) but we can use it as a special case to obtain information about the weak gravitational field. So, applying the Lorentz transformation to the metric yields the following expression

$$\begin{aligned} g_{\alpha'\beta'} &= \Lambda^{\mu}{}_{\alpha'}\Lambda^{\nu}{}_{\beta'}g_{\mu\nu} \\ &= \Lambda^{\mu}{}_{\alpha'}\Lambda^{\nu}{}_{\beta'}\eta_{\mu\nu} + \Lambda^{\mu}{}_{\alpha'}\Lambda^{\nu}{}_{\beta'}h_{\mu\nu} \\ &= \eta_{\alpha'\beta'} + h_{\alpha'\beta'}. \end{aligned} \tag{4}$$

We see that the perturbation, $h_{\mu\nu}$, transforms exactly like a tensor in special relativity, giving us an idea of how to treat the concept of waves in spacetime. In this sense, $h_{\mu\nu}$ is a tensor field acting on a flat manifold and thus creating propagating ripples which we interpret as gravitational waves. All the tensors in general relativity will then be defined only in terms of $h_{\mu\nu}$, as we will later see.

Next we consider the other important symmetry, namely the gauge invariance of general relativity (see, e.g., Wald 1984; Weinberg 1972). This means that Eq. (2) remains unchanged by a coordinate transformation of the form

$$x^{\mu} \to x^{\mu'} = x^{\mu} + \epsilon^{\mu}(x) \tag{5}$$

where we have to impose some kind of condition for the change-vector, ϵ^{μ}. At first glance it might seem obvious that we want ϵ^{μ} to be 'small'. But what exactly does that mean in this case? Let us take a look at the transformation properties of these new coordinates. The Lorentz transformation becomes

$$\Lambda^{\mu'}{}_{\nu} = \frac{\partial x^{\mu'}}{\partial x^{\nu}} = \frac{\partial x^{\mu}}{\partial x^{\nu}} + \frac{\partial \epsilon^{\mu}}{\partial x^{\nu}} = \delta^{\mu}{}_{\nu} + \epsilon^{\mu}{}_{,\nu} \tag{6}$$

and similarly, the inverse transformation becomes

$$\Lambda^{\mu}{}_{\nu'} = \delta^{\mu}{}_{\nu} - \epsilon^{\mu}{}_{,\nu}. \tag{7}$$

Applying these transformations to the metric gives

$$\begin{aligned} g_{\alpha'\beta'} &= \Lambda^{\mu}{}_{\alpha'}\Lambda^{\nu}{}_{\beta'}\eta_{\mu\nu} + \Lambda^{\mu}{}_{\alpha'}\Lambda^{\nu}{}_{\beta'}h_{\mu\nu} \\ &= \{\text{Neglecting terms of second order} \end{aligned}$$

$$
\begin{aligned}
& \text{in } \epsilon \text{ and products of } \epsilon \text{ and } h\} \\
= \; & (\delta^{\mu}{}_{\alpha} - \epsilon^{\mu}{}_{,\alpha})(\delta^{\nu}{}_{\beta} - \epsilon^{\nu}{}_{,\beta})\eta_{\mu\nu} \\
= \; & (\delta^{\mu}{}_{\alpha} - \epsilon^{\mu}{}_{,\alpha})(\eta_{\mu\beta} - \epsilon_{\mu,\beta}) \\
= \; & \eta_{\alpha\beta} + h_{\alpha\beta} - \epsilon_{\alpha,\beta} - \epsilon_{\beta,\alpha}. \qquad (8)
\end{aligned}
$$

It is now clear that the gauge transformation, Eq. (5), induces a transformation on $h_{\mu\nu}$ given by

$$h_{\mu\nu} \rightarrow h'_{\mu\nu} = h_{\mu\nu} - \epsilon_{\mu,\nu} - \epsilon_{\nu,\mu} \qquad (9)$$

and the constraint on ϵ^{μ} must therefore be that $\epsilon_{\beta,\alpha}$ is small, i.e

$$|\epsilon_{\beta,\alpha}| << 1. \qquad (10)$$

As long as Eq. (10) is fulfilled we are certain that the field equations remain invariant. We will now use these symmetry properties to rewrite Eq. (1) in terms of the perturbative field, $h_{\mu\nu}$.

We start with the Riemann tensor, originally written as

$$R^{\alpha}{}_{\beta\mu\nu} = \Gamma^{\alpha}{}_{\beta\nu,\mu} - \Gamma^{\alpha}{}_{\beta\mu,\nu} + \Gamma^{\alpha}{}_{\sigma\mu}\Gamma^{\sigma}{}_{\beta\nu} - \Gamma^{\alpha}{}_{\sigma\nu}\Gamma^{\sigma}{}_{\beta\mu}. \qquad (11)$$

Since we are in a region of very small curvature we can neglect terms of second order in Γ and the Riemann tensor simplifies to

$$R^{\alpha}{}_{\beta\mu\nu} = \Gamma^{\alpha}{}_{\beta\nu,\mu} - \Gamma^{\alpha}{}_{\beta\mu,\nu}. \qquad (12)$$

To see how Eq. (12) can be rewritten in terms of $h_{\mu\nu}$ we must first have the relationship between the Christoffel symbols and the metric. This is called the "affine connection" and is given by

$$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2}g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta}). \qquad (13)$$

Using the following symmetry properties of $g_{\mu\nu}$;

$$g_{\alpha\beta} = g_{\beta\alpha}$$

and

$$g_{\alpha\beta,\mu\nu} = g_{\alpha\beta,\nu\mu},$$

we can write the Riemann tensor in terms of $g_{\mu\nu}$

$$
\begin{aligned}
R^{\alpha}{}_{\beta\mu\nu} \; = \; & \frac{1}{2}g^{\alpha\sigma}(g_{\sigma\beta,\nu\mu} + g_{\sigma\nu,\beta\mu} - g_{\beta\nu,\sigma\mu}) - \\
- \; & \frac{1}{2}g^{\alpha\sigma}(g_{\sigma\beta,\mu\nu} + g_{\sigma\mu,\beta\nu} - g_{\beta\mu,\sigma\nu}) \\
= \; & \frac{1}{2}g^{\alpha\sigma}(g_{\sigma\nu,\beta\mu} - g_{\sigma\mu,\beta\nu} + g_{\beta\mu,\sigma\nu} - g_{\beta\nu,\sigma\mu}). \qquad (14)
\end{aligned}
$$

We now lower the first index and write

$$\begin{aligned} R_{\alpha\beta\mu\nu} &= g_{\alpha\lambda}R^{\lambda}{}_{\beta\mu\nu} \\ &= \frac{1}{2}(g_{\alpha\nu,\beta\mu} - g_{\alpha\mu,\beta\nu} + g_{\beta\mu,\alpha\nu} - g_{\beta\nu,\alpha\nu}) \\ &= \{g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}\} \\ &= \frac{1}{2}(h_{\alpha\nu,\beta\nu} + h_{\beta\mu,\alpha\nu} - h_{\alpha\mu,\beta\nu} - h_{\beta\nu,\alpha\mu}) \end{aligned} \tag{15}$$

were we used the fact that $\eta_{\mu\nu,\alpha\beta} = 0$. We can find the Ricci tensor in the weak-field limit by contracting the Riemann tensor over the first and third indices

$$R_{\mu\nu} = \Gamma^{\alpha}{}_{\alpha\mu,\nu} - \Gamma^{\alpha}{}_{\mu\nu,\alpha} \tag{16}$$

where the Christoffel symbols are simplified to

$$\begin{aligned} \Gamma^{\alpha}{}_{\mu\nu} &= \frac{1}{2}(\eta^{\alpha\beta} + h^{\alpha\beta})[h_{\beta\mu,\nu} + h_{\beta\nu,\mu} - h_{\mu\nu,\beta}] \\ &= \frac{1}{2}\eta^{\alpha\beta}(h_{\beta\mu,\nu} + h_{\beta\nu,\mu} - h_{\mu\nu,\beta}) + \bigcirc(h^2), \end{aligned} \tag{17}$$

so the first derivatives become

$$\Gamma^{\alpha}{}_{\mu\nu,\alpha} = \frac{1}{2}\eta^{\alpha\beta}(h_{\beta\mu,\nu\alpha} + h_{\beta\nu,\mu\alpha} - h_{\mu\nu,\beta\alpha}) \tag{18}$$

$$\Gamma^{\alpha}{}_{\alpha\mu,\nu} = \frac{1}{2}\eta^{\alpha\beta}(h_{\beta\alpha,\mu\nu} + h_{\beta\mu,\alpha\nu} - h_{\alpha\mu,\beta\nu}). \tag{19}$$

The Ricci tensor can now be expressed in terms of $h_{\mu\nu}$ only;

$$\begin{aligned} R_{\mu\nu} &= \frac{1}{2}\eta^{\alpha\beta}(h_{\beta\mu,\nu\alpha} + h_{\beta\nu,\mu\alpha} - h_{\mu\nu,\beta\alpha} - h_{\beta\alpha,\mu\nu} - h_{\beta\mu,\alpha\nu} + h_{\alpha\mu,\beta\nu}) \\ &= \frac{1}{2}(h^{\alpha}{}_{\mu,\nu\alpha} + h^{\alpha}{}_{\nu,\mu\alpha} - \Box h_{\mu\nu} - h_{,\mu\nu} - h^{\alpha}{}_{\mu,\alpha\nu} + h_{\alpha\mu,}{}^{\alpha}{}_{\nu}) \\ &= \frac{1}{2}(h^{\alpha}{}_{\mu,\nu\alpha} + h^{\alpha}{}_{\nu,\mu\alpha} - h_{,\mu\nu} - \Box h_{\mu\nu}), \end{aligned} \tag{20}$$

where $\Box = \eta^{\alpha\beta}\partial_\alpha\partial_\beta$ is the D'Alembertian in flat spacetime (see, e.g., Chakrabarty 1999). In Cartesian coordinates we recognize this as the four-dimensional wave operator, or Laplacian; $-\frac{\partial^2}{\partial t^2} + \nabla^2$. By contracting once again we find an expression for the Ricci scalar

$$\begin{aligned} R = \eta^{\mu\nu} R_{\mu\nu} &= \frac{1}{2}\eta^{\mu\nu}(h^{\alpha}{}_{\mu,\nu\alpha} + h^{\alpha}{}_{\nu,\mu\alpha} - h_{,\mu\nu} - \Box h_{\mu\nu}) \\ &= \frac{1}{2}(h^{\alpha}{}_{\mu,}{}^{\mu}{}_{\alpha} + h^{\alpha\mu}{}_{,\mu\alpha} - \eta^{\mu\nu} h_{,\mu\nu} - \Box h) \\ &= \frac{1}{2}(2h^{\alpha\mu}{}_{,\mu\alpha} - 2\Box h) \\ &= h^{\alpha\mu}{}_{,\mu\alpha} - \Box h, \end{aligned} \tag{21}$$

where we have defined the trace $h = h^{\alpha}_{\alpha} = \eta^{\alpha\beta} h_{\alpha\beta}$. The curvature of spacetime is now completely described by the perturbative field, $h_{\mu\nu}$, acting on a flat background. The weak-field expression for the Einstein tensor becomes

$$\begin{aligned} G_{\mu\nu} &= R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R \\ &= \frac{1}{2}(h^{\alpha}{}_{\mu,\nu\alpha} + h^{\alpha}{}_{\nu,\mu\alpha} - h_{,\mu\nu} - \Box h_{\mu\nu}) - \frac{1}{2} g_{\mu\nu}(h^{\alpha\mu}{}_{,\mu\alpha} - \Box h) \\ &= \frac{1}{2}(h^{\alpha}{}_{\mu,\nu\alpha} + h^{\alpha}{}_{\nu,\mu\alpha} - h_{,\mu\nu} - \Box h_{\mu\nu}) - \frac{1}{2}\eta_{\mu\nu}(h^{\alpha\mu}{}_{,\mu\alpha} - \Box h) + \bigcirc(h^2) \\ &= \frac{1}{2}(h^{\alpha}{}_{\mu,\nu\alpha} + h^{\alpha}{}_{\nu,\mu\alpha} + \eta_{\mu\nu}\Box h - \Box h_{\mu\nu} - \eta_{\mu\nu} h^{\mu\nu}{}_{,\mu\nu}). \end{aligned} \tag{22}$$

At this stage it would be very convenient if we could somehow use our gauge freedom to further simplify this expression. It turns out that, without changing the properties of Eq. (22), we are allowed to impose what is called the "harmonic gauge condition" (see, e.g., Weinberg 1972), i.e

$$h^{\mu}{}_{\nu,\mu} = \frac{1}{2} h_{,\nu}. \tag{23}$$

To see that this is true we might assume that there is some $h_{\mu\nu}$ that does not satisfy Eq. (23). Then, according to the transformation properties of $h_{\mu\nu}$, Eq. (9), we can always find some $h'_{\mu\nu}$ that does. In fact $h'_{\mu\nu}$ satisfies Eq. (23) if ϵ is defined by the following condition

$$\Box\epsilon = h^{\mu}{}_{\nu,\mu} - \frac{1}{2} h_{,\nu}, \tag{24}$$

ensuring that we can safely use Eq. (23) to simplify our equations. What is really done is that we decide to work in a particular coordinate system that makes life slightly simpler. So, applying Eq. (23) to Eq. (22) yields

$$\begin{aligned} G_{\mu\nu} &= \frac{1}{2}(h^{\alpha}{}_{\mu,\nu\alpha} + h^{\alpha}{}_{\nu,\mu\alpha} - \eta_{\mu\nu} h^{\mu\nu}{}_{,\mu\nu} + \eta_{\mu\nu}\eta^{\alpha\beta} h_{,\alpha\beta} - \eta^{\alpha\beta} h_{\mu\nu,\alpha\beta}) \\ &= \frac{1}{2}(\frac{1}{2} h_{\nu\mu} + \frac{1}{2} h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu} h_{,\nu\nu} + \eta_{\mu\nu}\eta^{\alpha\beta} h_{,\alpha\beta} - \eta^{\alpha\beta} h_{\mu\nu,\alpha\beta}) \\ &= \frac{1}{2}(h_{,\mu\nu} - \frac{1}{2}\eta_{\mu\nu} h_{,\nu\nu} + \eta_{\mu\nu} h_{,\alpha\alpha} - \eta^{\alpha\beta} h_{\mu\nu,\alpha\beta}) \\ &= \frac{1}{2}(-\Box h_{\mu\nu} + \frac{1}{2}\eta_{\mu\nu}\Box h). \end{aligned} \tag{25}$$

To make the final expression as neat as possible we define what is called the "trace reverse"[2] of $h_{\mu\nu}$ (see, e.g., Schutz 1984) as

$$\overline{h}^{\alpha\beta} = h^{\alpha\beta} - \frac{1}{2}\eta^{\alpha\beta}h, \tag{26}$$

so that $\Box\overline{h}_{\mu\nu} = -\Box h_{\mu\nu} + \frac{1}{2}\eta_{\mu\nu}\Box h$.

We have now achieved the goal that we set out for and the field equations in the weak-field limit can simply be written in the form

$$\Box\overline{h}_{\mu\nu} = -16\pi T_{\mu\nu}. \tag{27}$$

2.2 A Plane-Wave Solution

In this section we will take a look at the physically interesting solutions to Eq. (27) and examine some of their properties. I will omit most of the details of the derivations since they are not necessary for grasping the remainder of this paper. Interested readers are referred to standard textbooks in general relativity for more details (see, e.g., Weinberg, Wald).

When solving equation Eq. (27) we are, of course, interested in the case where the gravitational field is weak but not stationary. This problem has therefore many similarities with electrical sources in electrodynamics. Thus, we interpret $T_{\mu\nu}$ as a gravitational source, generating a gravitational wave, $h_{\mu\nu}$. The expression for the generation of gravitational waves can be computed using Green's functions (see, e.g., Arfken & Weber 2000; Jackson 1999) but here I will simply state the result, which is given by (see Carroll 1997):

$$\overline{h}_{\mu\nu}(t,\mathbf{x}) = 4G\int d^3\mathbf{y}\frac{T_{\mu\nu}(t-|\mathbf{x}-\mathbf{y}|,\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}. \tag{28}$$

From the point of view of gravitational-wave astronomy, it is more interesting to study the behaviour of waves propagating far away from the source. We must therefore solve Eq. (27) in vacuum, i.e. when $T_{\mu\nu} = 0$. The equation then reads

$$\Box\overline{h}_{\mu\nu} = 0. \tag{29}$$

For clarity we can choose to work in Cartesian coordinates and Eq. (29) becomes

$$\Big(-\frac{\partial^2}{\partial t^2} + \nabla^2\Big)\overline{h}_{\mu\nu} = 0. \tag{30}$$

This is a four-dimensional wave-equation with the standard plane-wave solution (see, e.g., Chakrabarty 1999)

$$\overline{h}_{\mu\nu} = A_{\mu\nu}e^{ik_\alpha x^\alpha}, \tag{31}$$

[2]The name originates from the fact that $\overline{h} = -h$.

where the amplitude, $A_{\mu\nu}$, is a constant, symmetric, second rank tensor and k_α is the constant wave-vector. To find the constraints that ensures that Eq. (31) is a solution we check if it satisfies Eq. (30):

$$\begin{aligned}\Big(-\frac{\partial^2}{\partial t^2}+\nabla^2\Big)\overline{h}_{\mu\nu} &= \Box\overline{h}_{\mu\nu}\\ &= \eta^{\alpha\beta}\partial_\alpha\partial_\beta\overline{h}_{\mu\nu}\\ &= \eta^{\alpha\beta}\partial_\alpha\partial_\beta(A_{\mu\nu}e^{ik_\beta x^\beta})\\ &= \eta^{\alpha\beta}\partial_\alpha(ik_\beta A_{\mu\nu}e^{ik_\alpha x^\alpha})\\ &= -\eta^{\alpha\beta}(k_\alpha k_\beta A_{\mu\nu}e^{ik_\rho x^\rho})\\ &= -\eta^{\alpha\beta}k_\alpha k_\beta\overline{h}_{\mu\nu}\\ &= -k_\alpha k^\alpha\overline{h}_{\mu\nu}. \end{aligned} \tag{32}$$

We see that for Eq. (29) to be satisfied we must have $k_\alpha k^\alpha\overline{h}_{\mu\nu}=0$. For realistic situations we cannot have all components of $\overline{h}_{\mu\nu}$ zero so the relevant condition for Eq. (31) to be a solution is

$$k_\alpha k^\alpha = 0. \tag{33}$$

Physically this means that k_α is tangent to the world-line of a photon and thus the waves propagate at the speed of light (see Carroll 1997).

We must now try to find a unique solution. This is done by eliminating all components of $A_{\mu\nu}$ that has no physical meaning. Again, we consider the symmetries of the solution. Since the currently discussed form of the field equations, Eq. (30), is valid under the harmonic gauge condition, Eq. (23), we should also consider how this affects the solution. We can rewrite Eq. (23) in terms of the trace reverse of $h_{\mu\nu}$ as (see, e.g., Schutz 1984)

$$\overline{h}^{\mu\nu}{}_{,\mu} = 0. \tag{34}$$

Inserting the expression for $\overline{h}_{\mu\nu}$ we get

$$\begin{aligned}\overline{h}^{\mu\nu}{}_{,\mu} &= \partial_\mu\overline{h}^{\mu\nu}\\ &= \partial_\mu(A^{\mu\nu}e^{ik_\alpha x^\alpha})\\ &= iA^{\mu\nu}k_\mu e^{ik_\alpha x^\alpha}\\ &= 0 \end{aligned} \tag{35}$$

which is only true if

$$A^{\mu\nu}k_\mu = 0. \tag{36}$$

This equation states that $A_{\mu\nu}$ must be orthogonal to k_μ. All the possible symmetries

have not yet been used. In the former section we saw that the following coordinate transformation;

$$x^\mu \to x^{\mu'} = x^\mu + \epsilon^\mu(x),$$

left the field equations invariant. In Eq. (9) we saw how this transformation induced another transformation on $h_{\mu\nu}$:

$$h_{\mu\nu} \to h'_{\mu\nu} = h_{\mu\nu} - \epsilon_{\alpha,\beta} - \epsilon_{\beta,\alpha}.$$

Now, let us see how $\overline{h}_{\mu\nu}$ transforms under Eq. (9):

$$\overline{h}_{\mu\nu} \to \overline{h}'_{\mu\nu} = h'_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h', \tag{37}$$

where $h'_{\mu\nu}$ is given by Eq. (9) and the trace, h', is given by

$$h' \equiv h'^{\mu}{}_{\mu} = h - \epsilon^{\nu}{}_{,\nu} - \epsilon_{\nu}{}^{,\nu}.$$

Inserting this into Eq. (37) yields

$$\begin{aligned}
\overline{h}'_{\mu\nu} &= h_{\mu\nu} - \epsilon_{\mu,\nu} - \epsilon_{\nu,\mu} - \frac{1}{2}\eta_{\mu\nu}(h - \epsilon^{\nu}{}_{,\nu} - \epsilon_{\nu}{}^{,\nu}) \\
&= h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h - \epsilon_{\mu,\nu} - \epsilon_{\nu,\mu} + \eta_{\mu\nu}\epsilon^{\mu}{}_{,\mu} \\
&= \overline{h}_{\mu\nu} - \epsilon_{\mu,\nu} - \epsilon_{\nu,\mu} + \eta_{\mu\nu}\epsilon^{\alpha}{}_{,\alpha}.
\end{aligned} \tag{38}$$

and the harmonic condition becomes

$$\begin{aligned}
\overline{h}'^{\mu\nu}{}_{,\nu} &= \overline{h}^{\mu\nu}{}_{,\nu} - \epsilon^{\mu,\nu}{}_{,\nu} - \epsilon^{\nu,\mu}{}_{,\nu} + \underbrace{\partial_\nu(\eta^{\mu\nu}\epsilon^{\alpha}{}_{,\alpha})}_{=0} \\
&= \overline{h}^{\mu\nu}{}_{,\nu} - \epsilon^{\mu,\nu}{}_{,\nu} \\
&= 0.
\end{aligned} \tag{39}$$

We know that the first term on the right-hand-side of Eq. (39) is zero so for $\overline{h}'_{\mu\nu}$ to satisfy Eq. (34) we need that $\epsilon^{\mu,\nu}{}_{,\nu} = 0$. But we can rewrite this expression slightly as

$$\begin{aligned}
\epsilon^{\alpha,\mu}{}_{,\mu} &= \eta_{\mu\nu}\epsilon^{\alpha}{}_{,\mu\nu} \\
&= \eta_{\mu\nu}\partial_\mu\partial_\nu\epsilon^\alpha \\
&= \Box\epsilon^\alpha \\
&= \Big(-\frac{\partial^2}{\partial t^2} + \nabla^2\Big)\epsilon^\alpha \\
&= 0
\end{aligned} \tag{40}$$

and we see that ϵ^μ also satisfies a wave-equation, which we can use to further reduce our degrees of freedom. The solution to Eq. (40) is, in analogy with Eq. (31),

$$\epsilon_\mu = B_\mu e^{ik_\alpha x^\alpha}, \tag{41}$$

where k_α is the usual wave-vector and B_μ are constant vector-coefficients to be determined. The final step towards a completely physically relevant description of the wave is therefore to choose the coefficients in a clever way. To see how to do this we insert the solution Eq. (40) into Eq. (38) (see Schutz 1984) and obtain

$$A'_{\mu\nu} e^{ik_\alpha x^\alpha} = A_{\mu\nu} e^{ik_\alpha x^\alpha} - ik_\nu B_\mu e^{ik_\alpha x^\alpha} - ik_\mu B_\nu e^{ik_\alpha x^\alpha} + ik_\alpha \eta_{\mu\nu} B^\alpha e^{ik_\alpha x^\alpha}. \tag{42}$$

After dividing out the exponential factors we get

$$A'_{\mu\nu} = A_{\mu\nu} - ik_\nu B_\mu - ik_\mu B_\nu + ik_\alpha \eta_{\mu\nu} B^\alpha. \tag{43}$$

We can now choose the coefficients, B_μ, to obtain the last two restrictions on $A_{\mu\nu}$. It is convenient to choose B_μ so that $A_{\mu\nu}$ becomes traceless (see, e.g., Weinberg 1972; Schutz 1984; Caroll 1997)

$$A^\mu{}_\mu = 0 \tag{44}$$

and also

$$A_{\mu\nu} U^\nu = 0, \tag{45}$$

where $\vec{U}$ is some fixed four-velocity that we can choose freely.

The conditions we have imposed are called, under a common name, the "transverse traceless gauge" (TT gauge) and since, as long as we remain in this gauge, $A_{\mu\nu}$ is traceless we can drop the 'bar' on the field, $h_{\mu\nu}$, so that

$$\overline{h}_{\mu\nu} = h_{\mu\nu}. \tag{46}$$

No symmetric freedom remains to simplify our equations, which means that the components that are left must be physically important. To get an idea of what these components are, we consider a background Lorentz transformation (see, e.g., Chakrabarty 1999; Schutz 1984; van Holten 1997), analogous to Eq (3), where the vector $\vec{U}$ is the time basis vector $U^\nu = \delta^\nu{}_0$. Together with Eq. (45) this implies that $A_{\mu 0} = 0$ for all values of μ. Remaining in this frame, we can choose the axes so that the wave is traveling in the z-direction, i.e. $k^\mu = (\omega, 0, 0, \omega)$. According to Eq. (36) this imposes yet another constraint on $A_{\mu\nu}$, $A_{\mu z} = 0$ for all values of μ. All the non-zero components of $A_{\mu\nu}$ are now; A_{xx}, A_{yy}, A_{xy} and A_{yx}. But since $A_{\mu\nu}$ is trace-less and symmetric it must be that $A_{xx} = -A_{yy}$ and $A_{xy} = A_{yx}$. The remaining components of $A_{\mu\nu}$ can thus be written in matrix form as

$$A_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & A_{xx} & A_{xy} & 0 \\ 0 & Axy & -Axx & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{47}$$

Physically this means that any gravitational wave, traveling in the z-direction[3], must be a superposition of two plane waves, characterized by the coefficients A_{xx} and A_{xy}.

2.3 Polarization of Gravitational Waves

In this section I completely omit mathematical details and simply describe the results that are relevant for understanding the problems of detecting gravitational waves.

The polarization of the waves is closely related to the fact that we only have two relevant coefficients, A_{xx} and A_{xy}, to describe the wave. This immediately implies that we must also have two possible polarization states. To understand the polarization of these two parts we consider a ring of particles in the xy-plane (see Carroll 1997). Suppose first that the wave has the following properties; $A_{xx}\neq 0$ and $A_{xy} = 0$. The particles will then oscillate back and forth in a 'plus-like' manner (see Fig. 2) and we therefore rename this state as, $A_{xx} = A_{+}$. The other case, $A_{xy}\neq 0$ and $A_{xx} = 0$, then corresponds to a

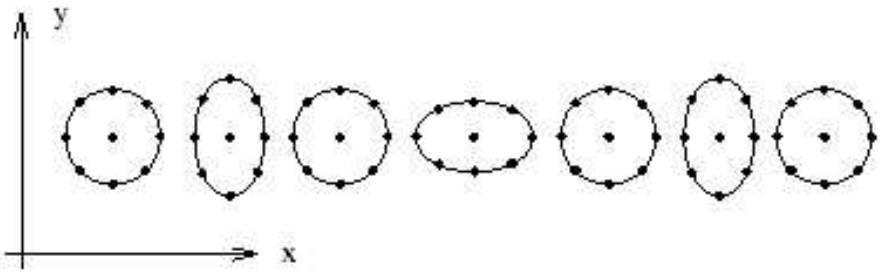

Figure 2: The behaviour of a ring of particles when interacting with a gravitational wave of polarization A_{+} ($A_{xx}\neq 0$ and $A_{xy} = 0$) (from Carroll 1997).

'cross-like' oscillation (see Fig. 3) and we call this state $A_{\times}$. What is interesting to notice

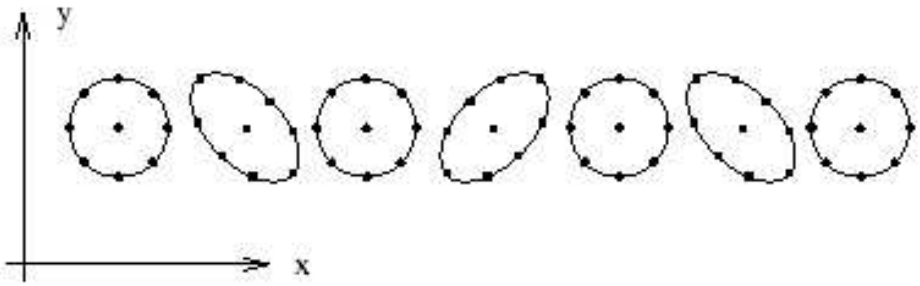

Figure 3: The behaviour of a ring of particles when interacting with a gravitational wave of polarization $A_{\times}$ ($A_{xy}\neq 0$ and $A_{xx} = 0$) (from Carroll 1997).

[3]The analysis is, of course, similar for a wave traveling in any spatial direction.

here, is that the two polarizations are rotated $\pi/4$ relative to each other[4]. As we will see in Sect. 3 this is important for the construction of gravitational-wave antennas. For a wave that is a superposition of both polarization modes, i.e. $A_{xx} \neq 0$ and $A_{xy} \neq 0$, the ring of particles will undergo a complete rotation in either direction, depending on the sign of the polarizations (see Fig. 4).

Another interesting feature of the polarization appear when one tries to quantize the wave (see Carroll 1997). When quantizing a field theory, the polarization angle is related to the spin of the quantized particle. The spin of the particle is, in general, given by $S = \frac{2\pi}{\theta}$, where θ is the angle under which the polarization modes are invariant. The electromagnetic wave is invariant under a rotation of 2π so the photon has spin 1. Equivalently, the gravitational wave is invariant under a rotation of π so the corresponding particle must be mass-less (since the wave propagates with the speed of light) and have spin 2. This, predicted, but not yet detected particle, is called the *graviton*.

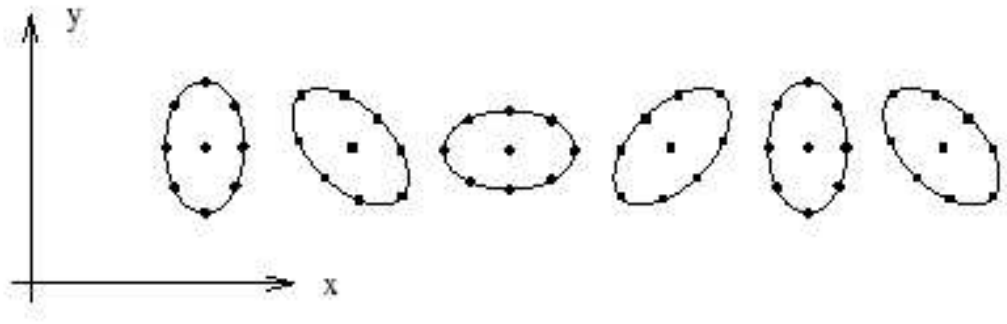

Figure 4: The behaviour of a ring of particles, interacting with a gravitational wave that is a superposition of the two polarization states, A_+ and $A_\times$ (from Carroll 1997).

2.4 Intensity of Radiation

There are two main problems with the detection of gravitational radiation; gravitational waves interact very weakly with matter and the emitted power is much lower than that of electromagnetic radiation. For the moment we shall focus on the second problem, and try to understand why this is so.

In electrodynamics, the main contribution to the emitted power comes from dipole radiation. The energy emitted per unit time by an oscillating electric dipole, d, is given by (see Rezzolla 2003)

$$L_{e.d} = \frac{2}{3} q^2 a^2, \tag{48}$$

where $d \equiv qx$ and $\ddot{d} \equiv q\ddot{x}$. Here x is the direction of oscillation and q is the electrical charge. Now we are interested in doing the same calculation for gravitational radiation. In this case we have instead an oscillating mass-dipole. Suppose that we have N number of point-like particles with mass m_A. Then the total mass-dipole moment is

[4]The corresponding polarization angle in electrodynamics is $\pi/2$. This difference in polarization is due to the fact that the gravitational wave is described by a second rank tensor, $h_{\mu\nu}$, and the electromagnetic wave is described by a vector potential, A_μ (see Schutz 1984).

$$\vec{d} \equiv \sum_{A=1}^{N} m_A \vec{x}_A \tag{49}$$

and we can calculate the momentum by taking the first derivative of $\vec{d}$

$$\dot{\vec{d}} \equiv \sum_{A=1}^{N} m_A \dot{\vec{x}}_A = \vec{p}. \tag{50}$$

This expression seems to indicate that it is straight-forward to calculate the dipole-moment of the gravitational radiation. There is one problem though: conservation of momentum demands that

$$\dot{\vec{p}} \equiv \ddot{\vec{d}} = 0, \tag{51}$$

telling us that there can be no mass-dipole emission in the form of gravitational waves, i.e. $L_{m.d} = 0$. By solving the complete weak-field equations, Eq. (27), it is possible to show that the gravitational wave generated by an astrophysical source is proportional to the second derivative of the quadrupole moment (see Carroll 1997), i.e.

$$\overline{h}_{ij}(t, \mathbf{x}) = \frac{2}{3R} \frac{d^2 q_{ij}(t_r)}{dt^2}, \tag{52}$$

where q_{ij} is the quadrupole moment, R is the radius of the source and $t_r = t - R$. The first-order contribution to gravitational radiation is therefore in the form of quadrupole emission. We already got a hint about this behaviour in the section about polarization. There we saw that the waves only produced a shear in the particle distribution and no average translational motion, thus indicating a quadrupole emission. This fundamental difference between electromagnetic and gravitational radiation originates from the fact that the center of charge in an electromagnetic source can oscillate freely but an oscillating center of mass violates conservation of momentum. The quadrupole moment is, in general, smaller than the dipole moment, and gravitational radiation is therefore much weaker than electromagnetic radiation.

3 Detection of Waves from Astrophysical and Cosmological Sources

There are three approaches to an experimental verification of gravitational waves. First, we can build antennas in the hope of detecting radiation from astrophysical sources, such as stellar collapse, black hole collisions etc. Second, we can hope to detect the possible background of relic gravitational waves originating from the very early universe, giving us a great opportunity to verify predictions from the standard model and other competing theories. Last, it might also be possible to find indirect evidence of gravitational emission from the inflationary epoch by studying its affect on the Cosmic Microwave Background. In this section, we will focus on the first two approaches and leave the third for Sect. 4.

It is an exciting time for astronomers because it might not be too long before an entirely new window on the universe opens up (see, e.g., Thorne 1995B). The sensitivity of gravitational-wave antennas has increased enormously during the last few years, and it

can no longer be considered a dream to engage in the field of gravitational-wave astronomy. There are many limitations to the current research in 'classical' astronomy, mainly because of extensive scattering of the light that we are so eager to measure. When we study the violent processes in the universe we often see just the outermost regions because the inner part is hidden behind a wall of scattered photons. And as if this is not enough, the light that actually does escape is very often prevented from reaching us because of the vast atomic and molecular dust clouds that lies in its path. The same problem arises when we try to observe far back in time towards the big bang. At a time when the universe was about 300 000 years old we encounter another obstacle; the microwave background radiation. This radiation comes from the time when electrons and protons united to form neutral atoms and thus released the incredible amount of electromagnetic radiation that was formerly trapped in clouds of charged particles. We can therefore never hope to observe further back than this with conventional astronomy. It is well known though that the gravitational force acts weakly on low-mass objects, and this causes the gravitational waves to interact extremely weakly with ordinary matter that might stand in the way on their passage through space. This is both a blessing and a curse because this is exactly what makes gravitational waves so hard to detect.

There are currently many projects in progress for building antennas that are sensitive enough to observe this radiation. We will focus on two of these that have a somewhat different approach but both are very promising: the ground-based LIGO (Laser Interferometer Gravitational-Wave Observatory) (see LIGO homepage) and the space-based LISA (Laser Interferometer Space Antenna) (see LISA homepage) antennas. Following is a review of the theory behind these antennas and what they are designed to observe.

3.1 LIGO

LIGO is a collaboration between countries all over the world. Currently there are two antennas at work in USA but the plans are to build a world-wide interferometry network with additional antennas in Europe, Japan and Australia. The two working antennas are situated in Hanford and Livingston, a distance of 3000 km apart. The large distance is needed in order to minimize the possibility of coincident environmental disturbances. Both sites are equipped with highly sensitive sensors to rule out false detections.

A schematic image of the antennas can be seen in Fig. 5. It consists of widely separated test masses that will oscillate upon interaction with the wave. Suppose, for simplicity, that the axes of the antenna lie along the x- and y-axes in Fig. 1. If the wave travels along the z-axis and hits the antenna, the test masses will oscillate in the same plus-like manner as the polarized wave. This will cause a difference in length of the two arms (the arms are initially of the same length, L), L_1 and L_2, given by (see Hughes 2002)

$$\delta L(t) \equiv L_1(t) - L_2(t). \tag{53}$$

For the simplified case of a '+'-polarized wave, this equation becomes

$$\delta L(t) = A_+(t)L, \tag{54}$$

where L is the initial length of the two arms (see Thorne in Hawking & Israel 1988 for a full derivation of Eq. (53)). In reality, it is more likely though that the wave is a

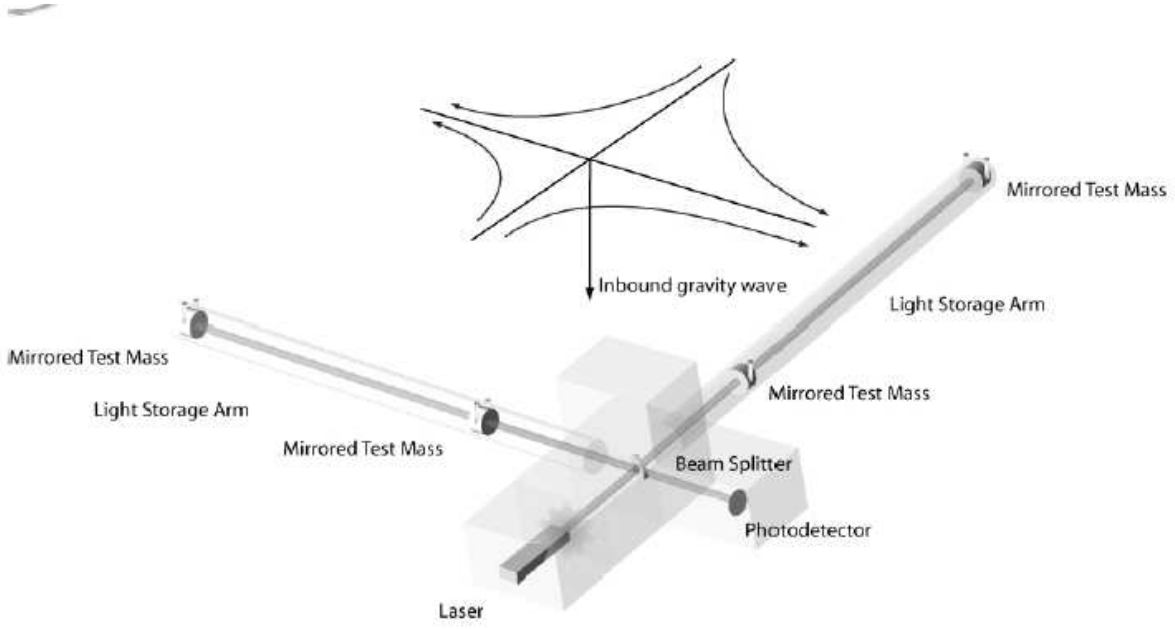

Figure 5: The ground-based LIGO antenna. The two storage arms are initially at the same length, L, but an incoming wave will induce a distance shift, $\delta L(t)$ between the arms. The distance shift is measured by laser beams that oscillate coherently between the test masses if no wave is present. As the wave hits the antenna it will change the oscillations of the lasers (from LIGO homepage).

superposition of both polarizations and the general form of the displacement can then be written as

$$\delta L(t) = [F^{+}A_{+}(t) + F^{\times}A_{\times}(t)]L \equiv A(t)L, \tag{55}$$

where F^{+} and $F^{\times}$ are the antenna response functions, which can be seen as weight functions describing how the polarization depend on the position and orientation of the source relative to the detector (see Hawking & Israel 1988). To get an idea of what sensitivity the antenna must have we need the order of magnitude of the amplitude, A. By Eq. (52) and Eq. (31) we see that

$$A \sim \frac{\ddot{q}}{R} \tag{56}$$

where $\ddot{q}$ is approximately given by

$$\ddot{q} \simeq 2Mv^2 \simeq 4E_{kin}^{ns}. \tag{57}$$

Here v is the internal velocity of the source and E_{kin}^{ns} is the non-spherical part of the internal kinetic energy of the source (see Hughes 2002). LIGO will, among other things, observe the coalescence of binary neutron star systems so let us take this as an example. For this process, we have $E_{kin}^{ns}/c^2 \sim 1\,M_{\odot}$. To get the dimensions right we must change back from natural to SI-units and we get

$$A \sim \frac{G}{c^4}\frac{\ddot{q}}{R} \sim 10^{-21} - 10^{-22}. \tag{58}$$

Inserting the largest limit into Eq. (55) yields

$$\delta L(t) = 10^{-21} L, \tag{59}$$

which tells us that the sensitivity must be extremely high. For every kilometer of baseline, L, we need to detect a distance shift of better that 10^{-18}m. Now might be a good time to ask the question: How is it possible to reduce the noise in the system so that such a weak signal can be measured? This is, of course, a complex question and a lot of research have been spent on solving this problem. But there is one basic principle to rely on when discussing noise reduction and that is that the incoming wave acts in a coherent way and the disturbing noise is completely random. This fact makes it possible to average over many vibrations and thus, hopefully, filter out the weak signal.

The different sources of gravitational waves are categorized by the range of frequencies in which they emit radiation. LIGO is mostly designed to observe within the *high-frequency* band[5] ($1\,\mathrm{Hz} \leq f \leq 10^4\,\mathrm{Hz}$), where it should be possible to detect such diverse phenomena as compact binaries, stellar core collapse and stochastic backgrounds (see, e.g., LIGO homepage; Thorne 1995A). We will look a little closer at stochastic backgrounds in Sect. 3.3.

The measurements by LIGO started already during the summer 2002. This was basically a test run and in February this year the second run started with a Eq 10 times higher sensitivity. The measurements will last until April 14. LIGO has also started collaborations with interferometers in Germany (GEO) and Japan (TAMA) who will be observing jointly.

3.2 LISA

Ground-based antennas can never hope to observe in the *low-frequency* band ($10^{-5}\,\mathrm{Hz} \leq f \leq 1\,\mathrm{Hz}$) because of surface vibrations on earth (see Hughes 2002). This is the reason for the planned launch of the space-based antenna LISA in 2011. The project is a collaboration between NASA (the National Space Agency) and ESA (the European Space Agency). In principle, LISA is designed similarly to LIGO but because of a big difference in scale they will not function in quite the same way. LISA arranges its test-masses in an equilateral triangle, orbiting the sun at about 1 Au (see Figs. 6 and 7). The length of the arms is 5×10^6 km, vastly greater than the arms on LIGO. Because of perturbations in the orbits of the planets these lengths are not constant but varies by about 1% over time-scales of months. This is not a problem though because the distance shifts that LISA will measure are of the order of 10^{-12} m and occurs on much shorter time-scales. Because of the large distances it is not possible to use the same kind of interferometry as in LIGO (Fabry-Perot interferometry). Diffraction would spread the laser beams as they propagate between the spacecrafts. Instead one uses *Michelson* interferometry. Interested readers are referred to the LISA homepage. Some of the sources that will be observed with LISA are binary star systems within the Milky Way, coalescing binary black holes and stochastic backgrounds.

[5]This frequency band roughly coincides with the audio band of the human ear. It is therefore possible to convert incoming waves to sound and detect them without fancy equipment.

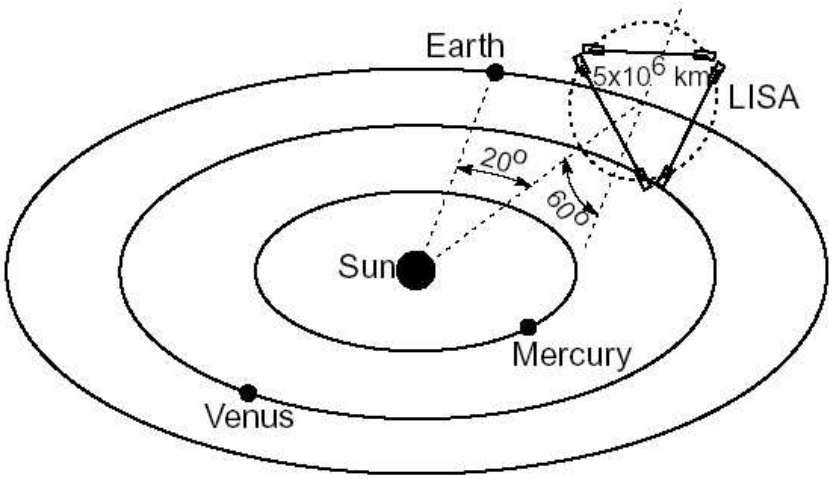

Figure 6: The LISA antennas will orbit the sun at about the same distance as the earth. The test-masses are ordered in an equilateral triangle with a distance of 5×10^5 m between the spacecrafts. The configuration is based on a Michelson interferometer (from Hughes 2002).

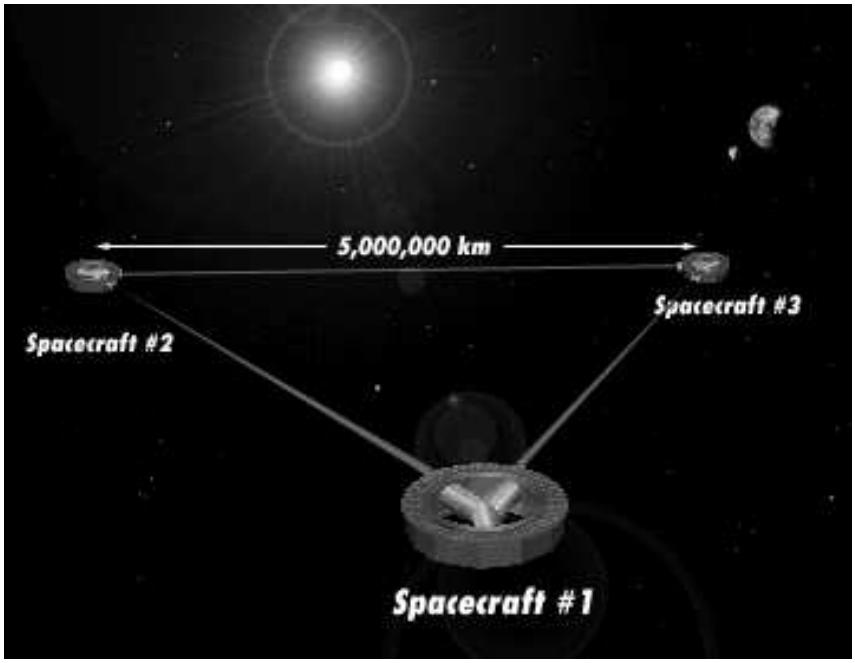

Figure 7: The LISA antennas (see Fig. 6 for more info) (from LISA homepage).

4 Relic Gravitational Waves

A particularly interesting feature of this new field of research[6] is the possible existence of a Cosmic Gravitational-wave Background (CGB), i.e. the gravitational analogy to the CMB (see, e.g., Grishchuk 2000 & 2002; Giovanni 1999; de Araujo et al. 2000) . Since the gravitational field interacts extremely weakly with other energy components in the universe, the gravitons decoupled already at the Planck epoch (10^{-43} s after the big bang). A detection of this background would therefore immensely enhance our understanding of the physics that governed the early universe. Because of limitations on the energy scales in current particle accelerators, this might also be the only way to verify unified particle theories, such as String/M-Theory. There are many processes that produce a spectrum

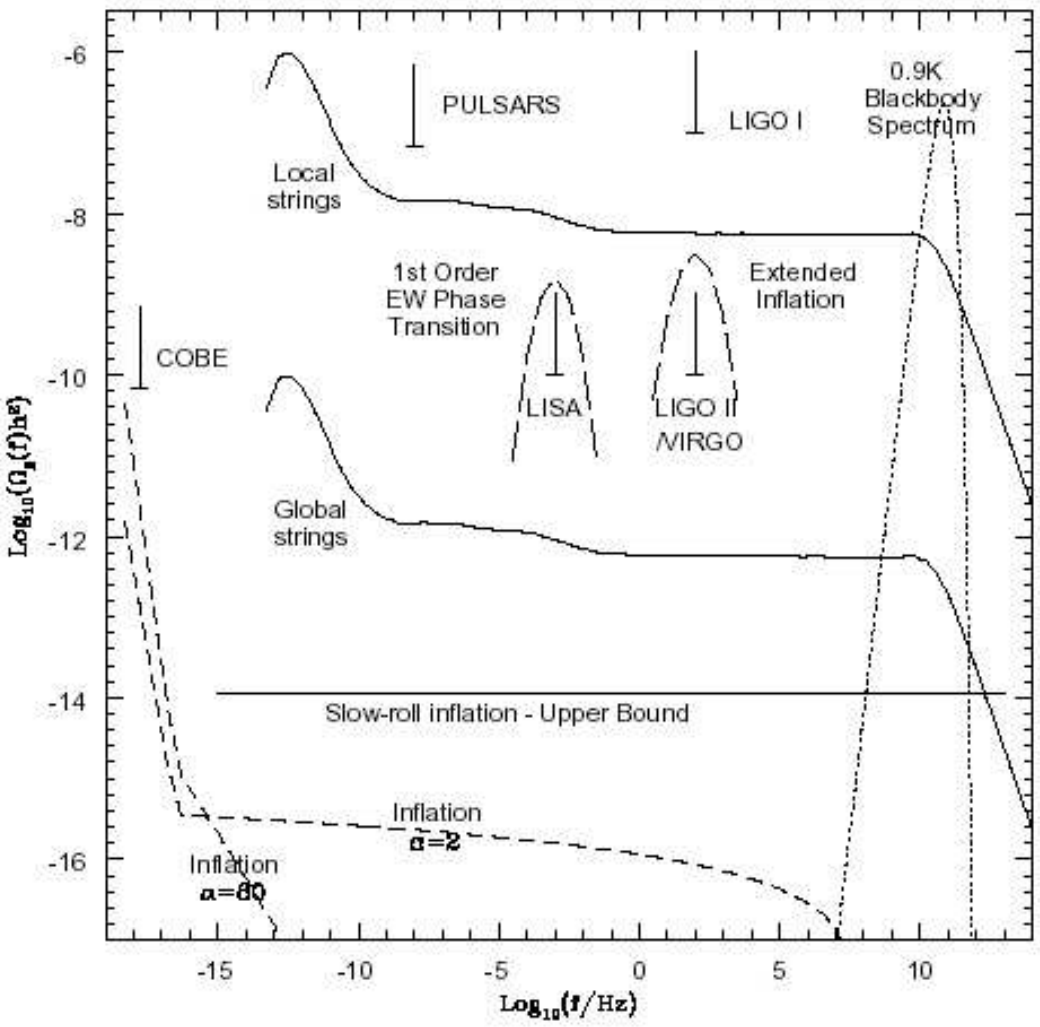

Figure 8: LISA and LIGO sensitivities compared to the expected intensity of potential sources of a CGB. Cosmic strings and first-order phase transitions may produce detectable signals (from Battye & Shellard 1996).

of gravitational waves, but the question is if any of these are strong enough that current

[6] At least to this author.

detectors have sensitivities within the right range (see Fig. 8). Those that are most frequently mentioned in the literature include: waves produced during inflation, waves emitted from early universe phase transitions and waves generated through the decay of topological defects.

4.1 Amplification of Vacuum Fluctuations During Inflation

The theory of inflation extends the standard big bang model with the addition of a short period where the universe undergoes an extremely rapid exponential growth (see, e.g., Islam 2001; Narlikar 1993). This process solves a number of problems in standard cosmology and also provides a possible origin for structure formation in the universe. It is now widely believed that this period of expansion also generated an intense gravitational radiation that might be detectable today as a stochastic background (see, e.g., Liddle 1999).

Before inflation, the size of the universe was very small and quantum fluctuations in the primordial plasma was an important part of the total energy content. When the inflationary expansion started, these fluctuations where amplified to macroscopic scale and energy was released in the form of gravitational waves (see, e.g., Hogan 1998). A possible background could have originated both from fluctuations in the inflaton field and from quantum fluctuations in the total gravitational field of the universe itself.

From the point of view of detection, we are interested in the amplitude of these primordial waves. It turns out that waves originating from fluctuations in the inflaton could possibly be within the LISA frequency band but waves generated by fluctuations in the graviton field are not. For a more detailed discussion of gravitational radiation from inflation, see Grishchuk 2002.

4.2 Phase Transitions in the Early Universe

According to the Theory of Grand Unification (GUT), the known symmetries of particle physics have evolved from a larger symmetry group by a series of spontaneous symmetry breakings. If this is correct, then the universe have gone through a number of phase transitions during its evolution. This behaviour is well known in condensed matter systems where phase transitions are a common feature. It was not however until 1966 that this was first discussed in a cosmological context. Nambu suggested the existence of topological defects (see Sect. 4.3) and phase transitions in the universe. Today, these ideas are an integrated part of modern cosmology and, as we will see, might have profound consequences for the interpretation of a detected CGB.

In particle theories, symmetry breaking is caused by a scalar field, ϕ, commonly called the Higgs field, and its associated potential, $V(\phi)$. The symmetry breaking occurs when the vacuum (groundstate) expectation value is non-zero and does not exhibit all the symmetries of the corresponding Hamiltonian. To get an idea of how this works we can take a look at a simple model, originally proposed by Goldstone in 1961. The model starts by assuming a Lagrangian of the form (see Vilenkin & Shellard 1994)

$$\mathcal{L} = (\partial_\mu \overline{\phi})(\partial^\mu \phi) - V(\phi), \tag{60}$$

where the potential is given by the famous "mexican hat"-potential

$$V(\phi) = \frac{1}{4}\lambda(\overline{\phi}\phi - \eta^2)^2, \tag{61}$$

where λ and η are positive constants. The field ϕ is a function of position only, $\phi = \phi(x)$, and is invariant under a global transformation[7] that simply changes the phase of the field by an amount α

$$\phi(x) \rightarrow e^{i\alpha}\phi(x). \tag{62}$$

The potential has a minima in the form of a circle with radius $|\phi| = \eta$ at the bottom of the hat. This corresponds to the vacuum state with a non-zero expectation value given by

$$\langle 0|\phi|0\rangle = \eta e^{i\theta}, \tag{63}$$

where θ is an arbitrary phase. It is clear that the transformation, Eq. (61) will change this phase from θ to $\theta + \alpha$ and the groundstate is no longer invariant under Eq. (61). To obtain a state of unbroken symmetry we have to move to the metastable local maximum of the potential where $\langle 0|\phi|0\rangle = 0$. If the system would end up in this unbroken state, any small perturbation would cause the field to roll into the broken state which has a lower energy. We therefore say that the symmetry is spontaneously broken.

A similar process might have occurred in the early universe with ϕ being some scalar field permeating all of spacetime. Suppose that the corresponding potential had two distinct minima and thus two distinct phases (see Fig. 9). If the field ends up in the

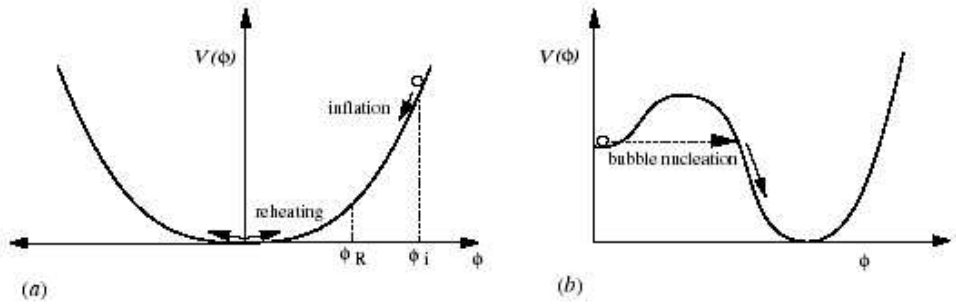

Figure 9: A potential with two minima, corresponding to the true and false vacuum state of the scalar field. Through quantum tunneling, the field undergoes a phase transition to the lowest minima, or the true vacuum state (from Battye & Sheppard 1996).

minima with the highest energy it is said to be in the "false" vacuum-state. Through quantum tunneling, the field can then move down to the low-energy minima, the "true" vacuum, where the field has a zero expectation value. This results in the creation of "bubbles" containing the lower phase as the phase transition occurs (see Fig. 10). As the field move into the lower phase a vast amount of energy is released in the form of gravitational waves. Also, when the phase-bubbles collide, the release of energy is strong

[7]This transformation is an element of the continuous group $U(1)$ and the name "global" simply refers to the fact that α is independent of position x.

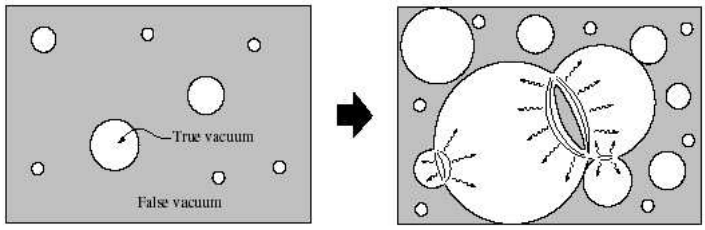

Figure 10: Bubble creation as the phase transition from the false to the true vacuum occurs. As the bubbles collide, large amounts of energy is emitted in the form of gravitational radiation (from Battye & Sheppard 1996).

enough to accelerate matter close to the speed of light. Depending on the rate of expansion during the phase-transition, the size of the bubbles vary and thus the amplitude of the generated waves vary. If the phase transition was first-order (discontinuous) then the expected intensity of the background can be detected by LISA (see, e.g., Hogan 1998; Ungarelli & Vecchio 2001). On the other hand, if the phase transition happened to be second-order (continuous), the resulting background can not be detected with current antennas.

4.3 Cosmic Strings and Other Topological Defects

The known topological defects that might be of interest for cosmology include cosmic strings, domain walls, monopoles and textures (see e.g. Hogan 1998; Vilenkin & Sheppard 2000)[8]. I will mainly focus on the first two, leaving the interested reader to investigate the others further. Perhaps one of the most familiar examples of a topological defect occurs

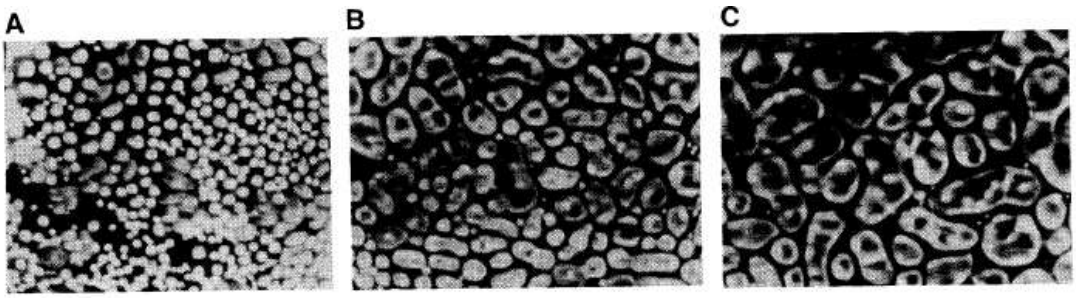

Figure 11: The nucleation of bubbles in a liquid crystal, leading to the formation of a string network (from Bowick et al.).

in ferromagnets, when the overall symmetric spin-configuration splits up into smaller domains with different magnetizations. The process decreases the total energy of the system. At the domain boundaries, domain walls appear and by this phenomenon one concludes that these walls must have been present before the splitting. Another interesting phenomenon, also well-known in laboratories on earth, is the formation of string vortices in liquid crystals (see Bowick et al. 1994) (see Fig. 11).

[8]The title of this section is borrowed from the book with the same title by A. Vilenkin and E. P. S. Sheppard 1994.

If the crystal undergoes a phase transition where the temperature rapidly drops below the critical temperature, then the change in phase will occur independently in different parts of the crystal. This implies that in each of these volumes there will be a symmetry breaking process completely independent of what is going on in other parts. This discrete separation of symmetry breakings can result in the formation of a network of defects, or a network of string vortices. Similar behaviour is found when symmetries are broken in particle physics theories and thus also in cosmological theories. Now let us take a closer look at what effects this behaviour could have on the early universe.

Cosmic strings (see, e.g., Battye et al 1997) could have formed if a scalar field, shortly after the big bang, got trapped in the false vacuum (see last section). The defects appear because of the rapid expansion and its effect on the interactions between various energy fields (see, e.g., Hogan 1998). The strings can be thought of as defects in the very geometry of spacetime itself and after formed they can move, stretch, change shape, be excited, form connections and break up into smaller objects called loops. They more or less behave like particles and thus they can also decay. Since they are, essentially, deformations of spacetime, it is not hard to imagine that when decaying, the main energy release is in the form of ripples in spacetime, i.e. gravitational waves. Cosmic strings might have formed as late as during the electroweak symmetry breaking [9] but for the emitted radiation to be strong enough for detection it is more relevant to focus on strings formed at the time of grand unification (10^{-35} s after the big bang (10^{15} GeV)). Strings formed during this epoch would have been massive enough to produce a detectable CGB and also to affect the anisotropy of the CMB (see Sect. 5).

5 Polarization of the CMB due to Gravitational Radiation

Although the microwave background is remarkably isotropic and homogeneous, there are still small temperature deviations on all angular scales. This anisotropy has been closely mapped by the COBE satellite and even more recently by the WMAP project (see WMAP results). The WMAP results show a large increase in accuracy compared to the COBE data and reveals lots of information about the state of the universe at the time of recombination. The spectrum of angular anisotropy might also be very interesting from the viewpoint of gravitational waves (see, e.g., Allen & Koranda 1994; Buonanno 2003; WMAP results nr 25), even though the CMB is electromagnetic in origin. The reason for this is that a primordial emission of gravitational radiation might have affected the formation of the small perturbations shown in the spectrum, thus creating a possible indirect detection of the CGB (see, e.g., Liddle 1999; Crittenden et al 1993). If this is correct, the temperature anisotropy has two different contributions; scalar perturbations in the inflaton field and tensor perturbations in the graviton field. To distinguish between these two components directly is very difficult but might be possible by noting that the contributions have different polarizations. Therefore it might not be enough to study the WMAP temperature anisotropy but instead to perform additional observations of the CMB polarization. Let us take a brief look at how such a procedure would work. If we could somehow measure an average "polarization vector" all over the sky, then this would correspond to a vector field permeating a sphere around us (see, e.g., Kamionkowski

[9] The breaking of the symmetry between the weak and electromagnetic forces occurred 10^{-11} s after the big bang. The corresponding energy scale is 10^3 GeV.

1998). Any vector field can be written as the gradient of a scalar function plus the curl of a vector field. Contributions from scalar perturbation have no 'handedness' and thus produces no curl, but as we saw in Sect. 2.3 gravitational waves do have a handedness and so must produce a curl. The curl- and curl-free components can be decomposed by taking the divergence and curl of the vector field. In this way one can determine whether or not there is a gravitational-wave contribution to the CMB temperature anisotropy and if there is, find its magnitude.

The data from the WMAP observations are still being analyzed and so far there has been no indication of a tensor contribution, but there is still a lot of data processing to be done. If it turns out that there is no contribution from gravitational waves, then this will impose serious constraints on the various inflationary scenarios.

6 Conclusions

There is no doubt that a direct detection of gravitational radiation would be a great triumph for physics, and would answer many questions that have haunted scientists for years and probably also questions that no one has even asked yet. General relativity would have passed yet another test of its validity, and perhaps even the hardest test so far.

The very notion of actually probing energies as high as the Planck energy is extremely interesting from a particle physics point of view. The limitations of current particle accelerators makes it very difficult to verify predictions from unifying theories. Therefore a detection of a stochastic background of gravitational radiation would be an enormous breakthrough, not only for astronomy, but also for elementary particle physics.

Some of the antennas have only just completed their first science runs and most have not even started yet. The sensitivities of current antennas will increase even more and a detection of some kind is to be expected during the upcoming 5-10 years. There are already plans underway for the next generation laser-antennas (see Cornish & Larson 2001), both on earth and in space. Should there be no detection whatsoever, it will surely be useful in its own way because it will rule out numerous predictions from particle and cosmological theories and perhaps force us to reevaluate Einstein's theory of gravity.

Acknowledgments

All results presented in this paper are due to the people listed in the references. I do not take any credit other than for the occasions where I have expressed some of my own thoughts and opinions. I would like to thank Alessandro for providing an interesting and inspiring course.

References

Allen B., Koranda S., 1994, Phys. Rev., D50, 3713, (astro-ph/9404068)

de Araujo J. C. N., Miranda O. D., Aguiar O. D., 2000, Phys. Rev., D61, 124015, (astro-ph/0004395)

Arfken G. B., Weber H-J., 2000, *Mathematical Methods for Physicists*, Harcourt/Academic Press, 5th Edition

Battye R. A., Caldwell R. R., Shellard E. P. S., 1997, in *Topological Defects in Cosmology*, Eds. Melchiorri. F., Signore. M., (astro-ph/9706013)

Battye R. A., Shellard E. P. S., 1996, Class. Quant. Grav., 13, A239-A246, (astro-ph/9610196)

Battye R. A., Shellard E. P. S., 1999, [astro-ph/9604059]

Bondi H., 1967, Nature, 179, 1072

Bowick M. J., Chandar L., Schiff E. A., Srivastava A. M., 1994, Science, New Series, 263, 5149

Buonanno A., 2003, [gr-qc/0303085]

Carroll S. M., 1997, [gr-qc/9712019]

Chakrabarty I., 1999, [physics/9908041]

Cornish N. J., Larson S. L., 2001, Class. Quant. Grav., 18, 3473-3496, (gr-qc/0103075)

Crittenden R., Davis R. L., Steinhardt P. J., 1993, Astrophys. J., 417, L13-L16, (astro-ph/9306027)

Giovanni M., 1999, [hep-th/9912480]

Grishchuk L. P., 2000, Lect. Notes Phys., 562, 167-194, (gr-qc/0002035)

Grishchuk L. P., 2002, in *A Relativistic Spacetime Odyssey*, Eds. Ciufolini. I., Dominici. D., Lusanna. L., World Scientific, (gr-qc/0202072)

Hawking S. W., Israel W., 1988, *Three Hundred Years of Gravitation*, Cambridge University Press

Hogan C. J., 1998, in *Second International LISA Symposium on Gravitational Waves*, Ed. Folkner. W., (astro-ph/9809364)

Hughes S. A., 2002, Annals Phys., 303, 142-178, (astro-ph/0210481)

Hulse R., Taylor J., 1975, *Astrophys. J.*, 324

van Holten J. W., 1997, Fortsch. Phys.,45, 439-516, (gr-qc/9704043)

Islam J. N., 2001, *An Introduction to Mathematical Cosmology*, Cambridge University Press, 2nd Edition

Jackson J. D., 1999, *Classical Electrodynamics*, John Wiley & Sons

Kamionkowski M., 1998, Lectures given at *The Early and Future Universe* CCAST workshop, Beijing, June 22-27, (astro-ph/9809214)

Liddle A. R., 1999, in proceedings of ICTP summer school in high energy physics, (astro-ph/9901124)

LIGO homepage, `http://www.ligo.caltech.edu`

LISA homepage, `http://lisa.jpl.nasa.gov`

Narlikar J. V., 1993, *Introduction to Cosmology*, Cambridge University Press, 2nd Edition

Rezzolla L., 2003, Lectures given at the Summer School on Astroparticle Physics and Cosmology, ICTP, July 2002, (gr-qc/0302025)

Schutz B., 1985, *A First Course in General Relativity*, Cambridge University Press

Thorne K. S., 1995A, [gr-qc/9506084]

Thorne K. S., 1995B, [gr-qc/9506086]

Ungarelli C., Vecchio A., 2001, Phys. Rev., D64, 121501, (astro-ph/0106538)

Vilenkin A., Shellard E. P. S., 2000, *Cosmic Strings and Other Topological Defects*, Cambridge University Press

Wald R. M., 1984, *General Relativity*, University of Chicago Press

Weinberg S., 1972, *Gravitation and Cosmology - Principles and Applications of the General Theory of Relativity*, John Wiley & Sons

WMAP results, 2003, [astro-ph/0302207-25]

Is the Fine-Structure Constant Really Constant?

Christoffer Petersson

Göteborg University
SE-41296 Göteborg, Sweden
(gu99chpe@dd.chalmers.se)

*

Abstract

The question of whether the fundamental constants of nature vary with space-time is important to both physics and philosophy. The focus will in this paper be on the fine-structure constant, which has a value corresponding to the strength of the electromagnetic force. The reason why this question recently has been seriously revived, is that observational results from the Keck/HIRES telescope has, in the last couple of years, been indicating a controversial trend in the fine-structure constant measurements. This trend, coming from quasar absorption spectra analysis, appears to prefer a smaller fine-structure constant in the cosmic past. Currently, 128 absorption systems, in the redshift range $0.2 < z < 3.7$, have been analyzed by Webb, Murphy, Flambaum and collaborators and given rise to a 5.7σ detection. But in order to determine whether this detection is due to an actual variation, or an undiscovered systematic effect, more independent observations are needed.

1 Introduction

Questions concerning varying constants was asked independently by Milne (1936, 1937) and Dirac (1937, 1938). Both considered a varying gravitational constant, G. The fundamental constant which will be considered here is the fine structure constant, which has the observationally vital property of being dimensionless.

The four fundamental forces in nature (gravity, electromagnetism, weak and strong interaction) express themselves with fundamental constants, or coupling constants. These constants can be considered as natural standards which everything else can be measured against. So, if anything is changing, the change can be detected from measurements and comparisons with these natural standards. For example, if one of two objects in a system becomes larger, it can be detected by comparing the sizes of these two objects. The comparison is done in form of a dimensionless ratio between their sizes. From this, one can tell that their size relative to one another have changed but not which of the object that is responsible. But, if instead everything in the system suddenly becomes larger this ratio comparison becomes meaningless. Suppose, for example, one tries to measure

*Hot Topics in Astrophysics 2002/2003, Alessandro B. Romeo, Christoffer Petersson, Daniel Persson & Oscar Agertz (Eds.), Chalmers University of Technology and Göteborg University, 2003.

whether c, which is a dimensionfull constant expressed in meters per second, has varied or not. If one meter is defined as the distance light travels in $1/c$ seconds, the change in c would not be detectable, since the length of one meter would also change (see, e.g., Duff 2002; Flambaum 2002). Another problem arises when one uses the old definition of the meter which was: One tenth of a million of the the distance from the equator to the north pole. It is in this case impossible to be certain whether the speed of light has changed or the size of the earth. The conclusion is that one needs natural standards in order to detect a variation of a quantity. But what if these natural standards themselves are changing? The key point, when one is considering varying constants, is that one is only interested in measuring a change in dimensionless constants. Thus to be able to detect a variation of a fundamental constant, it must have the form of a ratio of other constants, which change in a disproportional way.

The coupling constant of the electromagnetic force, is called the *fine structure constant*, α. Its value is equal to $1/137.03599958$ (see Mohr & Taylor 2000), with an experimental uncertainty of only 0.00000052. α is a dimensionless combination of three other fundamental constants:

$$\alpha = \frac{e^2}{\hbar c}. \tag{1}$$

This means that α, by definition, combines electromagnetism, quantum mechanics and relativity. The value of α can be obtained in a number of ways, for instance via experiments concerning the *quantum hall effect* and *neutron scattering*. The most accurate method comes however from experiments on the magnetic moment of the electron, which gives a result with an accuracy of a few parts per billion.

The rest of this paper is organized as follows. In Sect. 2, the atomic theory behind the two ways of determining a possible change in α, is examined. In Sect. 3, the observational methods and the analysis concerning quasar absorption spectra are reviewed, ending with current results and a summary of potential systematic effects. In the fourth and final section, there will be some conclusions and a discussion concerning various variations in α and future prospects.

2 Atomic Theory

There are, in practice, two ways to determine whether or not α has varied in time. The first approach was based on a comparison of the observed transition wavelengths of alkali doublets (see Cowie & Songaila 1995; Ivanchik, Potekhin & Varshalovich 1999; Varshalovich, Potekhin & Ivanchik 2000).

The second method, which offers about an order of magnitude improvement in precision, was founded by Dzuba, Flambaum and Webb (1999) and uses the relativistic correction to the energy levels of different ions.

This section is a review of the basic theory behind these two methods, starting from the Bohr model, and ending with a summary of constraints concerning the variability of α.

2.1 The Bohr Model

In the Bohr model, the energy levels are characterized by their principal quantum numbers, n, and given by

$$E_n = -\frac{me^4 Z_a^2}{2\hbar^2 \nu^2} \tag{2}$$

where m is the electron mass, Z_a is the charge seen by the external electrons, Z_a=1 for a neutral atom and Z_a=2 for a single charged ion, and ν is the effective principal quantum number representing the interaction between the external electrons, also called *screen effect*. For a hydrogen-like atom ν=n, the principal quantum number, and Z_a=Z, the ordinary nuclear charge (see, e.g., Haken & Wolf 2000).

2.2 Fine-Structure

In 1917, Sommerfelt introduced α in order to explain the observed splitting of the principal quantum number energy states of the Hydrogen atom. And together with the Dirac relativistic theory of the electron, physicists were able to extend the Bohr model by adding a correction term to the model. This additional term, which originates from special relativity, is called the *spin-orbit coupling* and takes into account, the interaction of the magnetic field created by the moving electron and the intrinsic magnetic moment of the electron, the *spin*. This coupling of the orbital moment, $\vec{L}$, and the spin of the electron, $\vec{S}$, gives rise to the total moment of the electron, $\vec{J}$=$\vec{L}$+$\vec{S}$, and produces a splitting of each energy level in the atomic spectra, creating a *fine-structure* of the atom. Heavy atomic nuclei, being strong central potentials, will attract the orbiting electron more than a light nuclei, and hence make the relativistic correction more significant.

It is also possible to compute a *hyperfine-structure* using the coupling of the total moment of the electron with the moment of the nucleus. But since its induced variation are in the order of a thousand times smaller than the variation induced by the fine-structure, it will be ignored.

There are basically two ways of using the fine-structure in order to detect a shift in wavelengths and a possible varying value of α: the alkali doublet method and the many multiplet method.

2.3 The Alkali Doublet Method

One interesting and useful fact about the fine structure is that, after the splitting, the energy difference between the new levels are proportional to α^2.

The alkali atoms have only one electron in the external layer, in their neutral state. α can therefore be calculated when a level divides into two, creating a *doublet*, due to the the fine structure splitting. If λ_1 is the wavelength of the corresponding transition from the $s_{1/2}$ state to the $p_{3/2}$ state, $s_{1/2}\rightarrow p_{3/2}$, and λ_2 is the wavelength of $s_{1/2}\rightarrow p_{1/2}$ of a alkali-like doublet then

$$\frac{\lambda_1 - \lambda_2}{\bar{\lambda}} \propto \alpha^2, \tag{3}$$

where $\bar{\lambda}$ is the average wavelength of λ_1 and λ_2 (see, e.g., Webb et al. 2001).

The problem with this *alkali doublet method* or AD method is that it is inefficient. The s ground-state has the largest relativistic correction, since it is closest to the nucleus and is therefore most sensitive to changes in α. However, since the ground-state is common to both doublet-transitions, measurements are done relative to the the same ground state and the relativistic shift for this state is not included in the AD method (see Fig. 1).

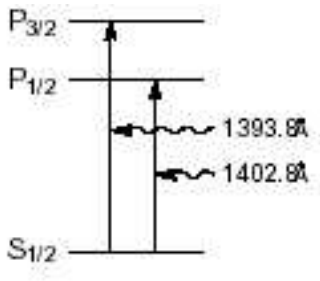

Figure 1: The transitions of the Si IV alkali doublet. The inability of the AD method to include the common s state, which have the maximum relativistic correction, can also be seen (from Murphy et al. 2002).

2.4 The Many Multiplet Method

In order to realize how the new improved method works, one starts by examining the ordinary relativistic correction to the energy levels of a transition in a hydrogen-like atom, obtained as an expectation value of the relativistic perturbation:

$$\Delta_n = -\frac{me^4 Z_a^2 (Z^2\alpha)^2}{2\hbar^2 n^3}\left(\frac{1}{j+\frac{1}{2}} - \frac{3}{4n}\right) \tag{4}$$

where j is the total momenta of the electron. Since Δ_n is only significant near the nucleus, the electron density must, in this region, have the following proportional relation:

$$\left|\Psi(r < \frac{a}{Z})\right|^2 \propto \frac{Z}{n^3 a^3} \tag{5}$$

where a is the Bohr radius and a/Z is the typical size of a hydrogen like atom (see Dzuba, Flambaum & Webb 1999). However, in astrophysical observations, one also observes atoms and ions, which have many electrons. One then needs a more appropriate formula for Δ_n. It is therefore useful that the electron density, for an electron in the external level near the nucleus, when $n >> 1$, obeys the following proportionality:

$$\left|\Psi(r < \frac{a}{Z})\right|^2 \propto \frac{Z_a^2 Z}{\nu^3 a^3} \tag{6}$$

One is now able to multiply Δ_n by the electron density in the many electron ion case divided by the electron density in the hydrogen like ion case and obtain the relativistic correction term for the many electron ion:

$$\Delta_n = \frac{E_n (Z\alpha)^2}{\nu^2}\left(\frac{1}{j+\frac{1}{2}} - \frac{Z_a}{Z\nu}(1 - \frac{Z_a}{4Z})\right) \tag{7}$$

where E_n is the energy level calculated earlier.

It has been shown by Dzuba, Flambaum and Webb (1999) that the relativistic correction, of atoms with one external electron, is approximately described by:

$$\Delta_n \approx \frac{E_n (Z\alpha)^2}{\nu^2}\left(\frac{1}{j+\frac{1}{2}} - C(j,l)\right) \tag{8}$$

where l is the orbital electron momenta and $C(j,l)$ is a coefficient derived from many-body calculations which include electron-electron correlations. In order to generate a complete set of electron orbitals and to obtain the correct value of Δ_n, a relativistic Hartree-Fock Hamiltonian is used in repeated calculations for different values of α. Since the attraction of a single electron to the nucleus increases the relativistic correction and hence makes its radius decrease, the nuclear potential screened by the core electrons also becomes smaller. The reason why the many-body effect, $C(j,l)$, has a negative sign is because it decreases the binding energy of this electron. In most atoms and ions used in the observations, $C(j,l)$ is, to a first approximation, equal to 0.6, for the lower energy levels, s and p orbitals.

Eq. (8) shows not only a α^2-dependence but also a Z^2-dependence. The latter implies that the correction is smaller in light atoms than in heavy ones. One is also able to deduce that the maximal correction is given when the total electron momenta, j, is minimized, j=1/2 which corresponds to the states $s_{1/2}$ and $p_{1/2}$. This means that with this method one is actually able to compare both heavy ($Z \sim 30$) and light ($Z \leq 10$) atoms and also $s \rightarrow p$ and $d \rightarrow p$ transitions in heavy atoms where the frequency shifts can be of opposite sign. So, compared to the AD-method, where only the fine structure splitting in the excited states within one multiplet were included, this method includes the total relativistic shift of frequencies in all transitions, including the ground state (see Fig. 2).

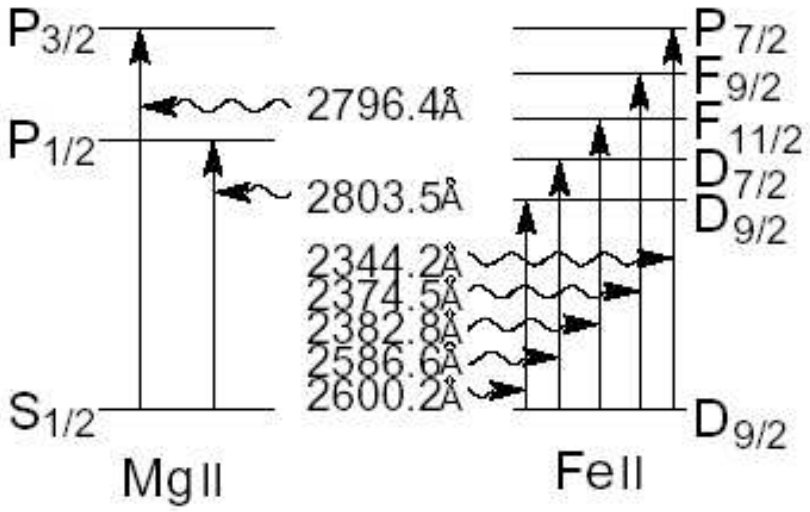

Figure 2: MgII/Fe II transitions are shown as an example of how the number of transitions and also the statistics are increased, making the systematic errors decrease (from Murphy et al. 2002).

The many multiplet method is consequently better from a statistical point of view, because it furnishes more measures since basically all atomic and ionic lines in different frequency ranges and redshifts are now of interest. This method also minimizes the risk of systematic errors since it is possible to exclude any spectral line to avoid effects like line blending and calibration errors.

2.5 The Frequency Dependence on α

The energy for fine structure multiplet levels are

$$E = E_0 + Q_1((\frac{\Delta\alpha}{\alpha})^2 - 1) + Q_2((\frac{\Delta\alpha}{\alpha})^4 - 1) + K_1(\vec{L}\vec{S})(\frac{\Delta\alpha}{\alpha})^2 + K_2(\vec{L}\vec{S})^2(\frac{\alpha_z}{\alpha})^4, \quad (9)$$

where E_0, Q_1 and Q_2 describe the position of the configuration center and where K_1 and K_2 describe the level splitting of a particular configuration (see Dzuba, Flambaum & Webb 1999). To find K_1 and K_2, the experimental fine structure intervals are fitted and numerically calculated. To find the dependence of the configuration center on α, the coefficients Q_1 and Q_2, numerical calculations for different values of α are used. $\Delta\alpha/\alpha$ is defined as $\Delta\alpha/\alpha = (\alpha_z - \alpha_0)/\alpha_0$, where α_z and α_0 are the values of α at the a certain redshift z and in the laboratory. The value of $\Delta\alpha/\alpha$ will from here on representing the varying of α in time. The final result can be represented in the following equation for the rest frequency, ω_z, of any transition observed in the quasar absorption spectra at a redshift z:

$$\omega_z = \omega_0 + q_1\Big[(\frac{\alpha_z}{\alpha})^2 - 1\Big] + q_2\Big[(\frac{\alpha_z}{\alpha})^4 - 1\Big] \quad (10)$$

where ω_0 is the laboratory frequency of the considered transition, $q_1 = Q_1 + K_1(\vec{L}\vec{S})$ and $q_2 = Q_2 + K_2(\vec{L}\vec{S})^2$.

The q-coefficients now contain all the relativistic corrections and measures how sensitive each transition frequency is to a change in α. It is important to notice that an error in the q_1 and q_2-coefficients will not infer an artificial non-zero value of $\Delta\alpha/\alpha$. In order to measure, or at least to place an upper limit on the variation of α, one now uses these data to fit the absorption systems in a quasar spectra.

If the q_2-part in Eq. (10) is approximated to zero, verified from the fact that only $\Delta\alpha/\alpha << 1$-values are considered, the relativistic corrections are now all represented in one coefficient, q:

$$\omega_z = \omega_0 + q\Big[(\frac{\alpha_z}{\alpha})^2 - 1\Big]. \quad (11)$$

The value of the q coefficients for typical transitions used in the MM method are plotted against wavelength in Fig. 3, divided into low and high redshift subsamples.

The small q-coefficients corresponding to the Mg-transitions are here used as *anchors* which large shifters as Fe II transitions can be measured against. The more complex high z sample, consisting of Fe and Al transitions, is more resistant to systematic errors like compression of the system, since wavelength distortions will here have a more varied effect on $\Delta\alpha/\alpha$.

For a given variation in α, the MM method produces wavelength shifts for different species which are able to vary, sometimes by 2 orders of magnitude, and also be of opposite signs, which helps to reduce the systematic effects.

Measurements of ω_0 with high precision have been carried out with the usage of *Fourier transform spectrographs* (see Pickering et al. 2000; Griesmann & Kling 2000) in order to exploit the high sensitivity in the calculations of the q-coefficients.

2.6 Variability Constraints on α

Constraints on how much α is allowed to vary can for example come from laboratory measurements consisting of comparisons between two clocks, a Hg^+ atomic clock and

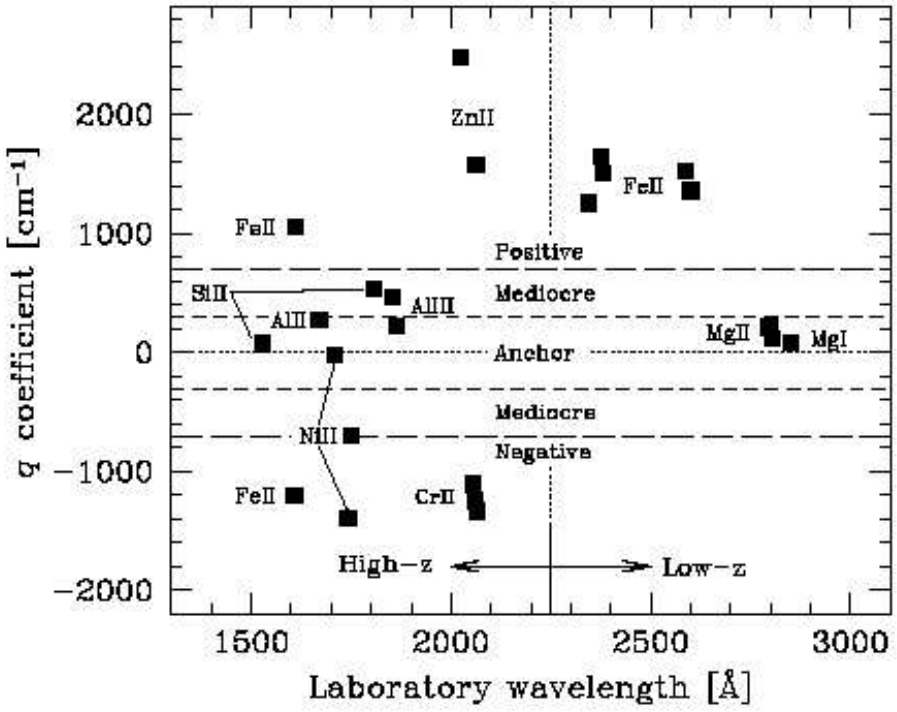

Figure 3: The distribution of q-coefficients of the MM method transitions are divided into sections of high and low z, and also anchor, mediocre and positive/negative values of q (from Murphy et al. 2003).

a hydrogen maser with different frequency-dependencies on α, which are sensitive to extremely small variations in α (see Sortais et al. 2001).

There is also a limit on the α variability coming from the Oklo uranium mine, situated in the African state of Gabon (see Damour & Dyson 1996). Terrestrial Uranium is usually found with a certain set of isotopic abundances of ^{238}U, ^{235}U and traces of ^{234}U, but where only ^{235}U is useful for the nuclear power process. But samples from the Oklo-mine revealed abundances which were very similar to those from todays used nuclear fuel. The conclusion was that there must have been a natural reactor in Oklo. It was also possible to determine the relative concentrations of two isotopes of the element samarium, which also was found in the mine, and compare it with the concentration found in unprocessed natural samarium.

The reason why there was a difference between the concentrations was that the reactor had, for about 1.7 Billion years, transformed one isotope into the other. This transformation needs a fine-tuned set of coincidences, which involves for example the electromagnetic force and the value of the fine-structure constants. The coincidences necessary for the reaction to occur, must have been about the same 1.7 billion years ago, as they are today. This places strong constraint on the possible change in the nuclear potential and on the variability of α. A more formal way to put it is that the Oklo-constraint arises essentially from the fact that the Q-value of the nuclear reaction is very sensitive to a variation in α since it contains an electromagetic contribution. The constancy of Q follow from the observations concerning the reaction rates now and at Oklo-time are the same. This is again a comparison of two clocks at different epochs.

The same argument can be applied to radioactive dating since lifetimes τ of a radioactive nuclei depend on Q and the He-nuclei decay rate, which depend exponentially on α

(see Barabash 2002). In the process Re→Os+e+ν one is able to compare the half-life, $\tau_{1/2}$, from laboratory experiments and the value inferred from Re/Os measurements in ancient meteorites, dated by other ratios, which are not significantly affected by a α-variation.

In order to get an estimate on the variability of α at earlier times one examines the *Cosmic Microwave Background* (CMB). Since the matter/radiation decoupling happened at a time when the temperature was low enough for atoms to form, there is an α and binding energy-dependence and hence a constraint from about 300 000 years after the big bang.

To achieve a constraint from even earlier times one examines the *Big Bang Nucleosynthesis* (BBN) and at the primordial ^{4}He, which is a measure of the neutron density at the weak interaction freeze-out tuned by the neutron/proton mass difference.

The results, which will be shown below, are all within the upper limits, set by these constraints, except for the one at Oklo. This constraint is however satisfied if α is assigned a non-linear evolution.

3 Observations and Results

The idea of using quasar absorption spectra of high redshift gas clouds, situated in front of a quasar, in order to constrain any possible α-variation over cosmological time-scales, was tested already in the 50's and 60's (see Bahcall, Sagent & Schmidt 1967). To get an idea of how a spectrum of this kind could look like, how it is analyzed and how the data is evaluated, some of the basics will here examined, ending with a summary of the results and the possible systematic errors.

3.1 Quasar Absorption Spectra

Quasars are supermassive black holes, generating light-sources, which are able to outshine entire galaxies of stars.These objects can be observed from great distances and analyzed via the spectra they produce. Since the universe is expanding, great distances means that these objects are ancient, often about 13 billions years old, making these spectra indicators of what the universe looked like then.

The spectra of quasars generally has some particular features . First of all, they contain peaks at certain wavelengths from, for example, excited neutral hydrogen transitions in the surroundings of the quasar (see the schematic spectrum in Fig. 4 at 4900Å). These emission features, called *Lyman alpha peaks*, are situated at some wavelength which can be compared to the wavelength the same hydrogen transition is situated at in the laboratory. In this way, the distance to the quasar can be computed since the redshift parameter z is defined as $1 + z = \omega_0/\omega_z$, where ω_0 is the laboratory wavelength and ω_z is the observed wavelength at redshift z.

Another characteristic feature of these spectra comes from the fact that the light from the quasar passes through a number of interstellar gas clouds which are somewhere between about 6 and 11 billion light years from us, corresponding to a redshift range of about $z \sim 1 - 3$, in the line of sight. Atoms in the gas clouds absorb parts of the light emitted by distant quasars. These absorption features, represented in the spectra by dips, are then compared to the corresponding transition wavelength obtained in the laboratory. The absorption lines from these gas clouds are substantially narrower than emission lines and it is therefore possible to compare the exact wavelength location of the dip to the theoretical location and obtain tight constraints on a α-variation.

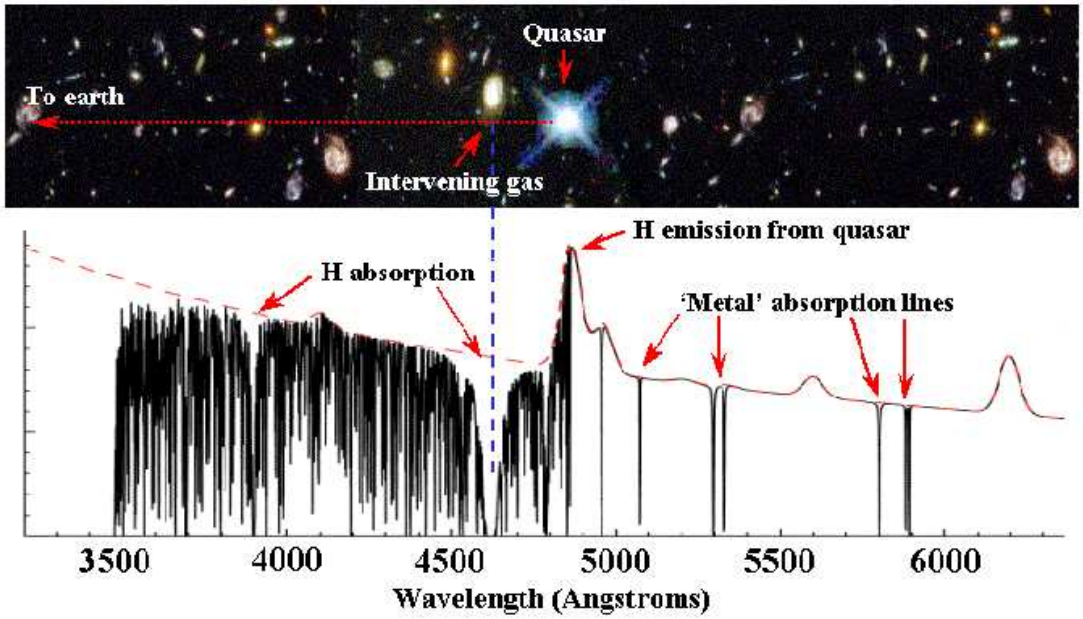

Figure 4: This absorption spectra is due to quasar light being absorbed in different interstellar gas clouds situated at six to eleven light-years away, in the line of sight between us and the quasar. These clouds are generally composed of well-studied elements like hydrogen, which give rise to the Lyman-alpha forest, and metals such as magnesium, iron, silicon and aluminum. The resulting spectrum consequently presents absorption features coming from these various elements and emission features from excited hydrogen (from Murphy 2002, private communication).

Typical elements contained in these clouds are, except for hydrogen which give rise to the *Lyman-alpha forest* (see Fig. 4 at 3500-4800Å), well studied elements such as magnesium, iron, silicon and aluminum.

3.2 Spectrum Analysis

To give examples of typical absorption features used in the analysis of an absorption spectrum, one could for instance look at the Mg II doublet at 2796Å and 2806Å (see the top two panels of the schematic absorption features in Fig. 5). These absorption features are located at redshift z=1, which means that the observed doublet wavelengths are redshifted by a factor $(1+z)$, to 5593Å and 5607Å (see Churchill 2001).

Fe II absorption lines at 2344Å, 2382Å and 2600Å can be seen in the bottom three panels in Fig. 5. The value of the redshift is, as mentioned above, normally obtained from typical emission peaks, making it possible to calculate the wavelengths where the absorption features should be found (see the black dip in Fig. 5). If α has not varied over cosmic time, then these dips should occur at precisely these redshifted wavelengths. But, if α has varied, the observed dip (red) should be located at a smaller or larger wavelength depending on the relativistic correction of the transition.

Suppose one assumes an α-variation of 1 part in 10000, which is about 10 times more than what is actually measured. The Mg II doublet dips would still be found essentially at the original wavelengths, because of their small q-values, which means that other transitions, with significant shifts, can be compared to them. Examples of such shifters are the Fe II lines which are shifted towards shorter wavelengths.

Examples of transitions which are shifted towards longer wavelengths are the Cr II

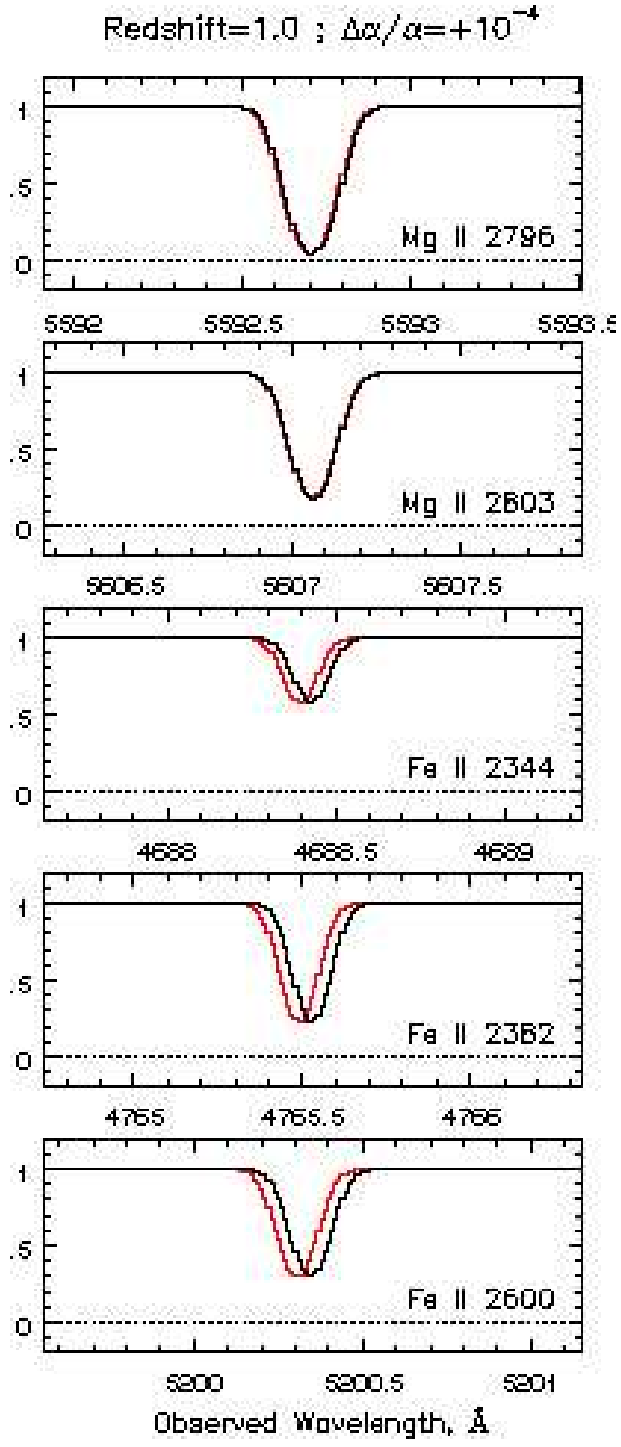

Figure 5: The transitions for Mg II 2796, Mg II 2803, FeII 2344, FeII 2382, and Fe II 2600, measured in a gas cloud at z=1. If there would be no evolution in α, the observed absorption features (red dip) would be located where the predicted dip (black) is. Here an α-variation of 1 part in 10000 is assumed yielding a wavelength shift of the transitions, redshifted by $1+z$, and found at slightly smaller wavelengths (from Churchill 2001).

doublet at 2056Å and 2066Å.

With the many multiplet method it is possible to measure the relative shifts of all observed transitions available and the fact that shifts toward both shorter and longer wavelengths are measured, reduces the risk of data calibration or analysis errors.

3.3 Statistical Analysis

To interpret the experimental data one needs statistical methods. The first thing to do is to estimate the errors. In this experiment, the most important limiting condition is the spectral line broadening, which is due to the effect of random velocities of the absorbing atoms within the gas clouds. To evaluate the broadening and to estimate the quality of the measurements, one introduces a velocity width parameter, **b**. The velocities of these atoms, thought of as gas-particles in thermal equilibrium, follows the *Maxwell-Boltzmann distribution* of velocities. To calculate a realistic value of **b** one adds a parameter for the turbulent motion of the atoms, $\mathbf{b}_{turb}$, and obtains the simple equation

$$\mathbf{b} = \frac{2k_BT}{M} + \mathbf{b}_{turb}, \tag{12}$$

where k_B is the Boltzmann constant, T is the temperature and M is the atomic mass (see, e.g., Murphy et al. 2003).

Since the exact conditions in the cloud are not known T and $\mathbf{b}_{turb}$ are considered to be free parameters together with z and $\Delta\alpha/\alpha$. These four parameters can now be thought of as free parameters of a *multiparameter fit*, using a nonlinear, least-square method, in order to obtain the best values of these four numbers for each absorption system.

To get the length of the error bars, σ, one uses the relation $\mathbf{b}_i = \sqrt{2}\sigma_i$ to obtain the mean value

$$\left\langle \frac{\Delta\alpha}{\alpha} \right\rangle = \sum \left(\frac{\left\langle \frac{\Delta\alpha}{\alpha}_i \right\rangle}{\sigma_i^2} \right) \frac{1}{\sum \frac{1}{\sigma_i^2}} \tag{13}$$

And the value of χ^2 is then given by

$$\chi^2 = \frac{1}{N} \sum \frac{\left(\frac{\Delta\alpha}{\alpha}_i - \left\langle \frac{\Delta\alpha}{\alpha} \right\rangle \right)^2}{\sigma_i^2}. \tag{14}$$

χ^2 will now be close to 1 if one chooses the length of the error bars correctly. Hence, the closer χ^2 is to 1, the more consistent the system of measurements is, including the derived free parameters. The first and second derivatives of χ^2 is then used, with respect to each free parameter, to give a natural weighting for the estimate of $\Delta\alpha/\alpha$.

Voigt profiles are after that fitted to every MM transition, for each quasar absorption system. To obtain the best fitting value of $\Delta\alpha/\alpha$, χ^2 is minimized simultaneously for all the velocity components. From the diagonal terms of the final covariance matrix, the 1σ error can be derived. To get a estimate of the errors and also of the $\Delta\alpha/\alpha$ value as a function of cosmic time, *Monte Carlo simulations* are used (see Webb et al. 2001).

3.4 Results

The AD method was first used to constrain possible variations of α, using galaxy emission spectra (see Savedoff 1956). 16 Si IV ADs were analyzed by Varshalovich, Potekhin and Ivanchik (2000), and the result was

$$\frac{\Delta\alpha}{\alpha} = (-4.6 \ ^{+}_{-} 4.3_{stat} \ ^{+}_{-} 1.4_{sys}) \times 10^{-5} \tag{15}$$

where the uncertainty in laboratory measurements of the Si IV transitions gave rise to the systematic error term. After improved laboratory wavelength measurements and extensions of the Si IV sample, the systematic error term was shown to be negligible, $\Delta\alpha/\alpha$=$(-0.5 \ ^{+}_{-} \ 1.3_{stat}) \times 10^{-5}$, which currently is the strongest constraint on $\Delta\alpha/\alpha$ from the AD method (see Murphy et al. 2001).

Three independent samples from the 10-meter Keck/HIRES telescope/spectrograph on Mauna Kea, Hawaii, have now been analyzed. In the first, one Mg II and Fe II multiplets in 30 quasar absorption systems was used, in the redshift range $0.5 < z < 1.6$ (see Churchill & Vogt 2001). Using the MM method for the first time, this analysis gave rise to the significant indication of a negative $\Delta\alpha/\alpha$ in the cosmic past (see Webb et al. 1999).

Damped Lyman Alpha absorbers were investigated in the second sample (see Prochanska & Wolfe 1999). Atomic transitions of Ni II, Zn II, Cr II and Al II were here used, and they produced a similar non-zero detection but now in the redshift range $1.8 < z < 3.5$.

By 2001, a full analysis was completed (see Murphy et al. 2001) and the MM method had now been applied to 49 absorption systems in the redshift range $0.5 < z < 3.5$ and a 4.1σ significant evidence was saying that α was smaller in the cosmic past, $\Delta\alpha/\alpha$=$(-0.72 \ ^{+}_{-} \ 0.18) \times 10^{-5}$.

In the third and most recent investigation (see Murphy et al. 2002) the number was increased to 128 absorption systems, in the redshift range $0.2 < z < 3.7$, showing a 5.7σ detection of a smaller α in the past, $\Delta\alpha/\alpha$=$(-0.57 \ ^{+}_{-} \ 0.10) \times 10^{-5}$. The result of these observations are shown in Fig. 6.

3.5 Potential Systematic Effects

During the period between the first observations and today serious efforts have been done in order to find and estimate the possible systematic errors. Up until now no significant systematic effect have been found. This section is a summary, based on the paper from Murphy, Webb and Flambaum (2003), of potential systematic effects, beginning with the two most important: atmospheric dispersion and isotopic abundance evolution.

Atmospheric Dispersion

When the spectrograph is not exactly perpendicular to the horizon, it is not parallel to the atmospheric dispersion direction, which means that the light coming from the quasar will be dispersed in a way depending on its wavelength, resulting in a slightly modified absorption spectra.

Since the Keck/HIRES telescope did not have an image rotator during the observations of 77 out of 128 absorption system, there was a risk that these systems could have been affected by these effects. This would have given rise to a compressing of the spectra yielding a mimicking of a negative α variation at low z. The result of comparing the two types of spectra is shown in the top panel of Fig. 7 and reveals that there is no significant difference between the corrected and uncorrected results. The trend of a varying α is however not significantly affected by this effect.

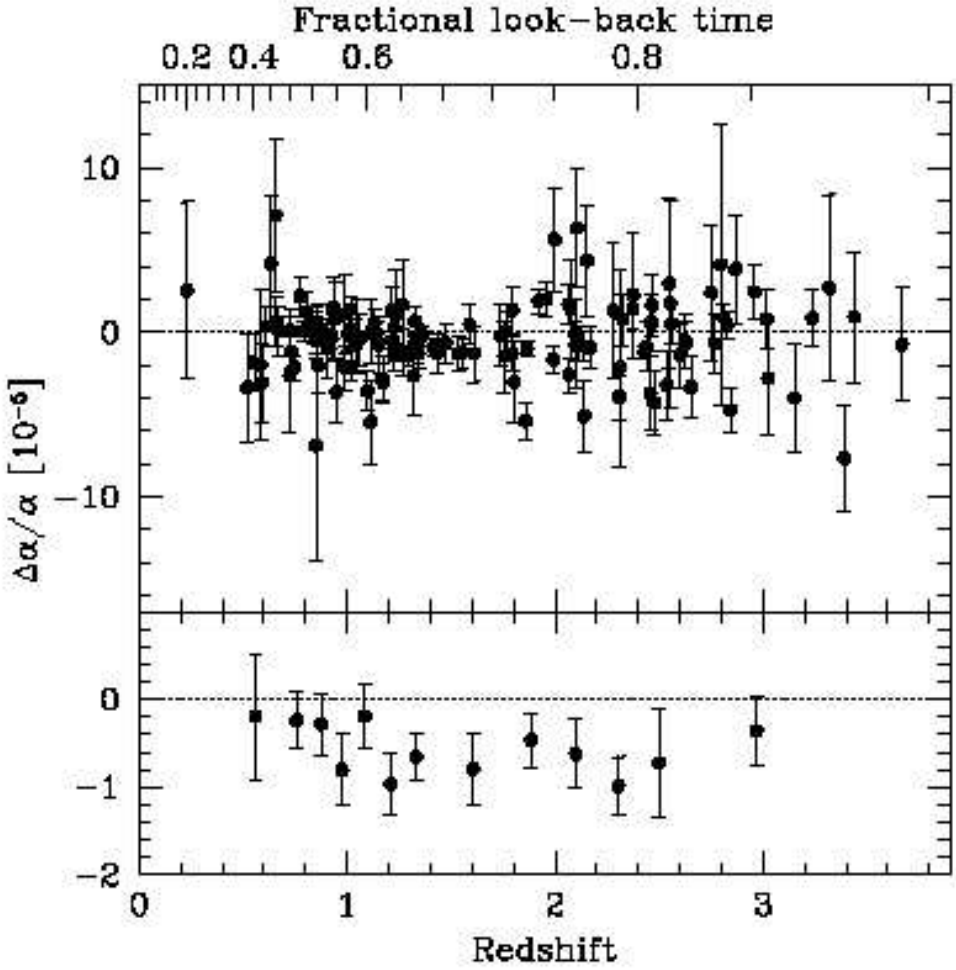

Figure 6: The result of the $\Delta\alpha/\alpha$ points are plotted against the redshifts of the absorption system, or equivalently, the fractional look-back time. The upper panel show the distribution of $\Delta\alpha/\alpha$-points for all 128 quasar absorption system with 1σ errors and in the lower, $\Delta\alpha/\alpha$ is binned and the weighted mean $\Delta\alpha/\alpha$ is shown with 1σ error (from Murphy et al. 2002).

Isotopic Abundance Evolution

The problem with isotopic abundance evolution comes from the fact that it is impossible to know how a particular element in a quasar is divided into its isotopes. Since the observed wavelength for one peak of an element is an average value depending on the proportion of the different isotopes. And because different isotopes have different masses, the wavelength of the emitted photon is affected.

Isotopic ratios corresponding to those found on earth are normally fitted to the absorption lines of the spectrum, which could yield a false value of $\Delta\alpha/\alpha$ if the actual isotopic ratios where different. Absorption lines from, for example, Mg and Si are suggested, from galactic observations (see Gay & Lambert 2000), to be totally dominated only by the isotopes ^{24}Mg and ^{28}Si. If only these isotopes were considered in the fit instead of the terrestrial ratio, there would still not be a significant change in the results. The magnitude of the effect from the removal of the weaker Mg and Si isotopic components is shown in Fig. 7.

As seen in the figure, this correction makes most of the observed average $\Delta\alpha/\alpha$-points become even more negative.

Differential Isotopic Saturation

For every absorption line, there is a composite of absorption lines coming from all the isotopes of each species. The composite laboratory wavelength values are only applicable in the linear part of the *curve of growth*, which corresponds to the *optically thin regime*. However, when the column density increases, the line centroid shifts in wavelength since the strongest isotope will begin to saturate. In light atoms, this mass isotopic shift is the dominant splitting and proportional to $\omega_0/(M)^2$. This is the reason why the light atoms or ions, with small q-coefficients, has the largest wavelength separations between the isotopes. To get an estimate of the maximal magnitude of these corrections, one assumes that all isotopes are saturated and that the line centroid lies at the unweighted mean isotopic wavelength. The magnitude would then be of order $\Delta\alpha/\alpha \approx -1 \times 10^{-5}$. This means that the effect of this correction would actually strengthen the claim of a negative $\Delta\alpha/\alpha$. And also, for every unsaturated transition added to the sample, the correction will be reduced. The weighted mean $\Delta\alpha/\alpha$ for the more complex structure of the high z sample is least sensitive to this effect which in the overall measurements can not have a significant effect on $\Delta\alpha/\alpha$.

Hyperfine Structure Effects

The composite wavelength of an absorption line will be unchanged by the hyperfine splitting, which means that the center of the hyperfine components remains at the same place when the splitting increases. The hyperfine splitting of energy levels only occur in species with either odd proton or neutron number. Since the species with these properties used in the measurements are either low abundant or have a small magnetic moment, it is quite unlikely that this effect could be of importance. There are, in addition to this, a possibility that the hyperfine-level populations in these low density gas clouds are not equal, since the equilibrium between levels are maintained predominantly by CMB photons. This potential lack of thermal equilibrium does not seem to have important effect on $\Delta\alpha/\alpha$ in these measurements.

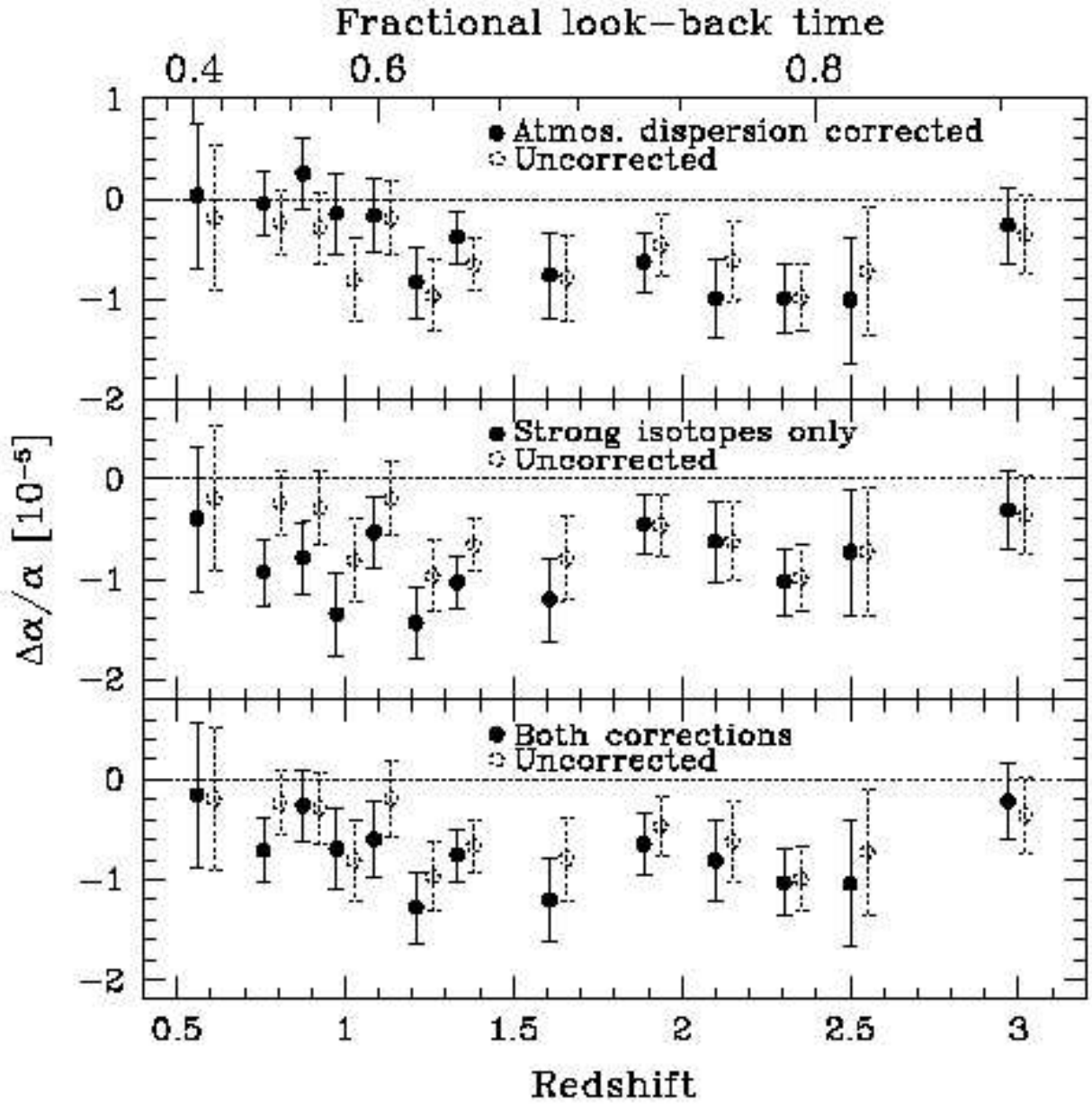

Figure 7: In the top panel the solid points with corresponding error bars represent the results where the effect of atmospheric dispersion has been taken into account in contrast to the uncorrected dotted ones. In the middle panel the results, corrected for isotopic abundance evolution, are compared with the uncorrected ones. And in the lower panel the result from the combined corrections of these, the two most important systematic errors, are shown (from Murphy et al. 2002).

Laboratory Wavelength Errors

If the laboratory values of ω_0 are wrong, it would lead to errors in single determinations of $\Delta\alpha/\alpha$. The accuracy from laboratory measurements are typically of a magnitude which implies that the maximum error in $\Delta\alpha/\alpha$ is at least an order of magnitude below the observed deviation from zero. The offset in wavelengths are degenerate with redshift, which means that it is impossible to introduce an apparent deviation in $\Delta\alpha/\alpha$, coming from laboratory wavelength errors.

Wavelength Mis-Calibration

A Thorium-Argon (ThAr) lamp was used as a comparison in the wavelength calibration of the quasar *Charge Coupled Device* (CCD) images, before and after the exposures. Since the laboratory wavelengths of the ThAr lines are known to greater accuracy than the values of ω_0, the only systematic effect possible is if there had been a ThAr line mis-identification, which could have led to a mis-calibration over large wavelength regions. The magnitude of this effect is estimated to be at least an order of magnitude less than the actual quasar absorption results. The effect of mis-calibration of the CCD, yielding a shift in α, can not be greater than $\Delta\alpha/\alpha \approx 10^{-7}$. Hence, calibration errors can not have a substantial effect on $\Delta\alpha/\alpha$.

Temperature Variation During Observations

Since the refractive index of air in a spectrograph depends on the surrounding temperature, a systematic mis-calibration of the wavelength scale in the quasar frame could occur if the quasar spectrum was calibrated with only a single ThAr exposure, but taken at a different temperature. This effect could give rise to an overestimation of the wavelength separation, but not an artificial non-zero $\Delta\alpha/\alpha$ value. The ThAr exposures were however taken close to the time of the quasar observations, in order to minimize these kinds of effects. In addition to this, *Image Header Information* were used to calculate temperature differences for all samples, indicating that this effect was unable to mimic a shift in $\Delta\alpha/\alpha$.

Line Blending

Errors coming from signal-to-noise and spectral resolution considerations and the profile fits of velocity structure were reduced because a great number of lines were fitted simultaneously. The problem is if the absorption system is not deconvolved into the actual number of velocity components. If unresolved lines are present, it could yield a shift of the fitted line wavelengths, corresponding to the velocity components of one or multiple transitions.

The way to estimate the magnitude of this effect is to do *line removal*, which means that one removes one or several transition(s) from the fit and examines if $\Delta\alpha/\alpha$ varies (see Fig. 8). For every transition removed, the weighted mean of $\Delta\alpha/\alpha$ and its associated 1σ error bar is compared before and after the line removal. In Fig. 8, no significant deviations are seen after the line removal, indicating that no particular transition is systematically blended. Thus, the results can not be affected by this effect.

Instrumental Profile Variations

The absorption line centroid can be wrongly situated if the *instrumental profile* (IP) of the spectrograph shows intrinsic asymmetries. And if these asymmetries in some way vary

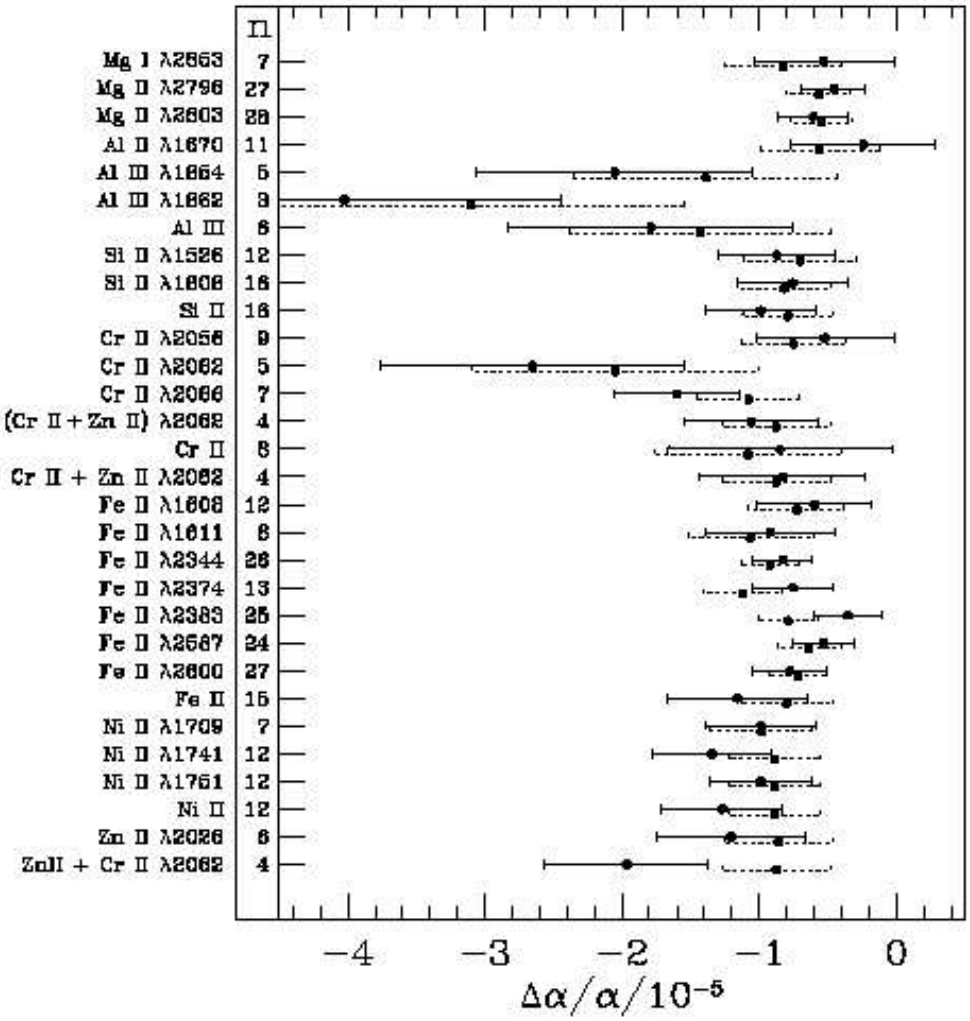

Figure 8: The dotted error bar corresponds to the weighted mean of $\Delta\alpha/\alpha$ before line removal and the solid error bar after the removal. To the left one can see which transition that has been removed and how many systems, n, this removal was possible in. Although a detailed analysis is required, due the fact that these quantities are in different ways dependent of each other, there has not been any extreme deviations found (from Murphy et al. 2002).

as a function of wavelength, this effect could yield an artificial non-zero value of $\Delta\alpha/\alpha$. This effect is however shown, in the observations done by HIRES, as a by-product of the wavelength calibration error investigation, to be negligible.

Heliocentric Velocity Variation

The change in heliocentric velocity can at most be in the order of 0.1 km/s, during a typical exposure of one hour. This change has the effect of smearing out the spectral features. This velocity space smearing can however in these measurements of $\Delta\alpha/\alpha$ be ignored, since it will be absorbed into the redshift parameter used in the fit to the spectrum.

Magnetic Fields

If strong magnetic fields were present in the medium between the galaxies, it could result in a shift of ionic energy levels and correlated values of $\Delta\alpha/\alpha$ in regions close by. The large number of quasars studied are however spread over almost the entire sky and the measured field strengths are typically in the order of μG in these regions, which is almost 9 orders of magnitude lower than what is required for a significant effect.

Kinematic Effects

In these observations, the structure of the absorption velocity features are assumed to be the same in different species. It seems however unlikely that one particular species would be red or blueshifted with respect to another when averaged over a large sample of absorption systems. Velocity dispersions due to, for example, expansion or asymmetric rotation would yield a greater observed scattering of the $\Delta\alpha/\alpha$ points than what should be expected from the $\Delta\alpha/\alpha$ error estimates.

The scattering, which is due to kinematic effects, in individual $\Delta\alpha/\alpha$ points are randomized over a large sample. Since high z points exhibit more scatter, probably because the precision is limited by systematics, the need for more absorption systems is greater than the need for increased signal-to-noise.

4 Conclusions and Discussion

In this section the question of whether or not there are spatial variations of α, or some preferred temporal evolution, will be addressed.

Also some basics of why higher dimensional theories have a varying α as a natural consequence and some of the effects a varying α would yield. There are theories where c is varying (see, e.g., Magueijo 2000), for example, in order to explain the *horizon problem*. In addition to this there are also theories where e is varying (see, e.g., Bekenstein 1982), due to a spacetime variation of a scalar field. Sandvik, Barrow and Magueijo (2001) has suggested a theory, based on Bekenstein's ideas, which predicts a variation of α, sensitive to whether it is the radiation, matter or Λ-dominated era.

4.1 Spatial or Temporal Variations in α?

If the measured α-variation was of spatial origin, one would expect some additional scattering in the raw values of $\Delta\alpha/\alpha$. The scatter at low z is however consistent with the error bars, which means that there is no evidence indicating that any spatial variations in α

exist for $z < 1.8$, at least not at the $\Delta\alpha/\alpha \approx 10^{-5}$-level (see Murphy, Webb & Flambaum 2003).

The distribution of $\Delta\alpha/\alpha$, is shown, in galactic coordinates, in Fig. 9. For comparison, the CMB pole and anti-pole are also shown. All information concerning redshifts is here lost, since all absorption systems in the line of sight toward the quasar are represented by a single weighted mean of $\Delta\alpha/\alpha$.

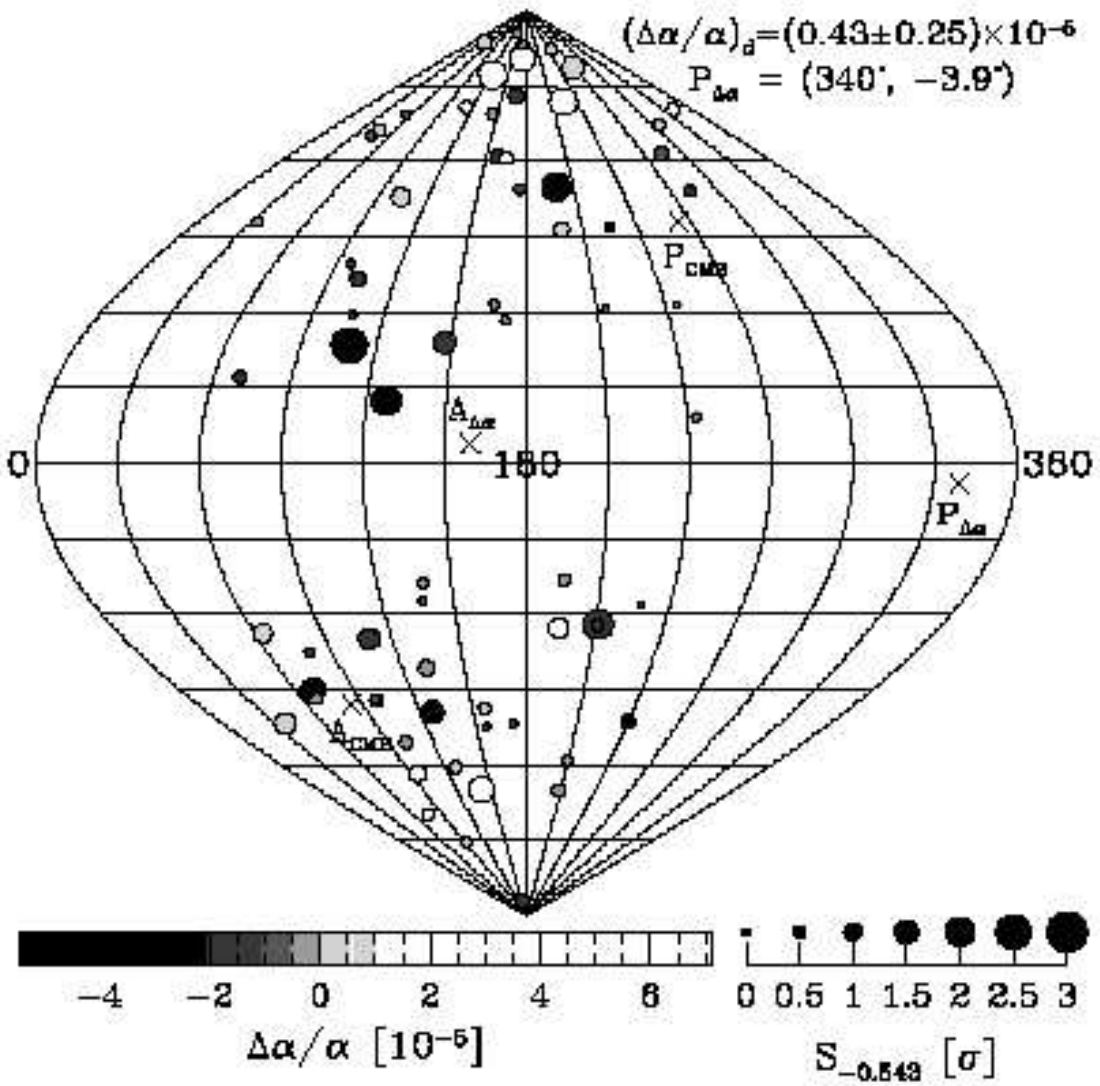

Figure 9: The spatial distribution of $\Delta\alpha/\alpha$. The value of $\Delta\alpha/\alpha$ is given by the grey-scale and the size of the points scales with the significance with respect to the overall weighted of the sample where $\Delta\alpha/\alpha = (-0.543 \pm 0.116) \times 10^{-5}$ (from Murphy, Webb & Flambaum 2003).

In the question concerning temporal variations, the data do not discriminate very well between a constant, linear or perhaps an oscillating evolution of $\Delta\alpha/\alpha$ with time. At the moment, the data do however prefer a linear evolution with time, but only to the confidence level where one would get more of a preference half the time by chance alone. This statement assumes that $\Delta\alpha/\alpha = 0$ at $z = 0$, which can be verified by the terrestrial constraints mentioned earlier. But, if one suspects spatial variations, as well as temporal, these constraints can not be applied and compared with the α evolution.

4.2 Extra Dimensions

Fundamental constants appear, in the ordinary laws of physics, not to be allowed to vary. But, what if the laws of nature are in reality laws of a world with more spatial dimensions than the three spatial one normally considers?

Theories trying to unify the four forces, like superstring theories and M-theory, which are defined in ten and eleven dimensions respectively, has the property of preferring varying coupling constants (see, e.g., Kolb, Perry &Walker 1986).

If space possess more spatial dimensions than the normal three, the constants of the physical laws would only be true constants in its higher dimensional world. In these theories, the coupling constants, like G and α, turn out to be dependent on the size of the additional dimensions, commonly thought of as compactified, according to Kaluza-Klein theoy, to the Planck length (10^{-35}m), and also on scalar fields introduced by the theory, like the dilaton in superstring theory (see, e.g., Marciano 1984).

Suppose the world consists of, for example, eleven dimensions and these dimensions are allowed to vary in time, in analogy with the hubble-expansion. The strength of the electromagnetic force would then only be constant if you measure it in all dimensions, but would be varying if you measure it in fewer. The observed variation of strength would be in proportion to the expansion/contraction of the additional compactified dimensions.

Since the variation of α is either zero or very slight, $\Delta\alpha/\alpha = 0,57 \times 10^{-5}$, the rate of this variation would be at least a billion times slower than the observed Hubble expansion.

4.3 Effects of a Varying α

A variation in α and the electromagnetic interaction would have drastic consequences in many areas of physics. Atoms, as we know them, only exists at places and times where α is very close to 1/137. If α was larger than this, the electrons would be pulled into the nucleus and if it was smaller, the atoms would not be held together strongly enough and thus unable to exist.

This implies that stars and humans could have formed if not α had the value or very close to the value it has today. This naturally leads one to the *anthropic principle* (see, e.g., Barrow 2000) and also to the question of whether the *equivalence principle* (see, e.g., Moffat 2001) is valid or not.

Even if the full extent of the physical and philosophical effects of a conclusive evidence of a varying α is impossible to estimate, there is no question physics will need some revision.

4.4 Future Prospects

The top priority for the future is to obtain more independent samples, since all analysis has been done on data from the Keck/HIRES telescope. If data from observations done for instance with VLT/UVES or Subaru/HDS telescope agree, it will be an important confirmation of the present results.

Observations focused on identifying new mm band molecular rotational absorption systems and H I 21-cm systems are currently ongoing. Since these systems have different potential systematic errors, they are important to compare the optical results with, especially in order to negating the line of sight velocity difference.

To summarize, if the effect indicated in the data is due to a varying α and not an undiscovered systematic effect, there will be a physical revolution, probably more contro-

versial than the one due to the Supernova type 1a indications of an accelerating Hubble expansion (see Perlmutter et al 1998; Riess et al 1998).

Acknowledgments

I would like to thank Alessandro Romeo for giving this wonderful course and being a true source of inspiration. I am also most grateful to Michael Murphy for valuable discussions and for providing me some of his research material.

References

Bahcall J. M., Sargent W. L. W., Schmidt M., 1967, ApJ, 149, L11

Barabash A. S., arXiv:nucl-ex/0210011

Barrow J. D., 2000, *The Universe that Discovered Itself*, Oxford U.P. , London and New York

Churchill C. W., 2001, *The Quasar Absorption Line Fine-Stucture Experiment* (`http://www.astro.psu.edu/users/cwc/fsc.html`)

Churchill C. W., Vogt S. S., 2001, AsJ, 122, 679

Cowie L. L., Songalia A., 1995, ApJ, 453, 596

Damour T., Dyson F., 1996, Nucl. Phys., B480, 37

Dirac P. A. M., 1937, Nature, 139, 323

Dirac P. A. M., 1938, Proc. Roy. Soc. Lond., A165, 199

Duff M. J., arXiv:hep-th/0208093

Dzuba V. A., Flambaum V. V., Webb J. K., 1999, Phys. Rev. Lett., 82, 888

Flambaum V. V., arXiv:astro-ph/0208384

Gay P.L., Lambert D. L., 2000, ApJ, 533, 260

Griesmann U., Kling R., 2000, ApJ, 533, 260

Haken H., Wolf H. C., 2000, *The Physics of Atoms and Quanta*, Springer, Berlin and New York

Ivanchik A. V., Potekhin A. Y., Varshalovich D. A., 1999, A&A, 343, 439

Kolb E. W., Perry M. J., Walker T. P., 1986, Phys. Rev., D33, 869

Magueijo J., 2000, Phys. Rev., D62, 103521

Marciano W. J., 1984, Phys. Rev. Lett., 52, 489

Milne E. A., 1936, Proc. Roy. Soc. Lond., A156, 62

Milne E. A., 1936, Proc. Roy. Soc. Lond., A158, 324

Moffat J. W., arXiv:astro-ph/0109350

Mohr P. J., Taylor B. N., 2000, Rev. Mod. Phys., 72, 351

Murphy M. T., Webb J. K., Flambaum V. V., 2003, submitted to MNRAS

Murphy M. T., Webb J. K., Flambaum V. V., Curran S. J., 2002, arXiv:astro-ph/0209488

Murphy M. T., Webb J. K., Flambaum V. V., Churchill C. W., Prochanska J. X., 2001, Mon. Not. Roy. Soc., 327, 1223

Perlmutter S. et al., 1998, ApJ, 517, 565

Pickering J. C. et al., 2000, Mon. Not. Roy. Soc., 319, 163

Prochanska J. X., Wolfe A. M., 1999, AsJ, 121, 369

Riess A. G. et al., 1998, ApJ, 116, 109

Savedoff M. P., Nature, 1956, 178, 689

Soratis Y. et al., 2001, Physica Scripta, T95, 50

Varshalovich D. A., Potekhin A. Y., Ivanchik A. V., 2000, AIP Conf. Proc. 506, *X-ray and Inner-Shell Processes*, Argonne National Laboratory

Webb J. K. et al., 1999, Phys. Rev. Lett., 82, 884

Webb J. K. et al., 2001, Phys. Rev. Lett., 87, 091301

Webb J. K. et al., 2002, arXiv:astro-ph/0210531

Planet Formation

Katarina Karlsson

Göteborg University
SE-41296 Göteborg, Sweden
(gu97kaka@dd.chalmers.se)

*

Abstract

Our knowledge of the solar system has lead to a model for its formation that predicted that planets should be common around other stars before we had made any detections of extrasolar planets. The last ten years have been interesting we both have got images of discs around young stars where planets are believed to form and we have made the first detections of extrasolar planets. Here the main properties of the solar system and its formation by core accretion is summarised. The possibility of giant formation disc instability is briefly discussed as it is not thought to be the most likely model for formation of giants in our solar system but could have been important for some extrasolar planets. An introduction to detected extrasolar planets is given and a discussion of their formation by comparing them with our solar system. In the end some speculations are introduced about of what we might learn by future observations and how observations of young stellar objects could determine if some planets are formed by disc instabilities are made.

1 Introduction

Our model of the formation of planetsis based on the study of one planet system - our own. Already in the mid 18:th century before we had knowledge about the outer planets Uranus and Neptune, the asteroids and where comets come from and are made of, Kant and Laplace suggested that the planets had formed in a rotating disc around the sun. Later when we learnt more about the lives of stars, a rotating gas disc seemed like a natural product around forming stars. Also planets like in our planet system might be able to grow in such discs. Therefore planets were thought to be common around other single stars even before we had any detections of extrasolar planets. In the 1980s then gas discs were found around forming stars by radio astronomy, and in the 1990s Hubble Space Telescope took the first images of these discs where we think planets are growing (see Kenyon 2000; Schilling 1999). 1995 the first extrasolar planet was detected around another sun. Less then ten years have passed since then and we now have detected 102 (see exoplanets.org 2003) extrasolar planets, of which some form planet systems. The

*Hot Topics in Astrophysics 2002/2003, Alessandro B. Romeo, Christoffer Petersson, Daniel Persson & Oscar Agertz (Eds.), Chalmers University of Technology and Göteborg University, 2003.

detection methods we have today reveals worlds which are very different from our own but can their formation be explained in the same model as our solar system? and what kind of planets are the most common ones around other stars? Still we belive based on the planet formation theory that terrestrial planets probably should be common around stars similar to our sun but our current detection methods are far from being able to detect them. In the next section the formation of the planets in our solar system is described. Then follows a presentation of detected extrasolar planets properties, a discussion of how they might have formed, what we can learn from them and some speculations about the future.

2 Planet Formation

Before going into the formation of planets we will look at the result of planet formation, our solar system to see what clues it gives us to planet formation. Then follows a short introduction to how starbirth leads to the protoplanetary disc that surrounds forming stars and the most important properties of these discs. The growth of planets in protoplanetary discs is usually divided up into tree stages: the growth of grains to planetesimals, planetesimals growth to protoplanets and formation of the final planets from protoplanets. Although they are described separately here they do not necessarily happen clearly separated in the disc or at the same time in the hole disc. The core accretion directly leads to the formation of terrestrial planets in the inner part of the disc and a similar growth but with gas accretion added to the more massive protoplanets in the outer solar system leads to the giant planets (see Lissauer 1993; Kenyon 2000; Pater & Lissauer 2001). The formation theory of the outer solar system is not in detail as well understood as the evolution leading to the terrestrial planets. The main problem for forming a Jupiter is that it has to capture a lot of gas before the gas disc disappears. Current theories gives Jupiter formation really in the upper limit of possible formation times. An alternative model is that of disc instabilities (see Boss 1997, 2002) which has its main advantage in that it gives fast formation of gas giants. Disc instabilities are briefly discussed in the end of this section.

2.1 Clues to the Formation of Our Solar System

One could writea lot about our solar system, but this very short introduction will focus on a few main properties which gives us clues about its evolution.

At a first overview of the planets of our solar system, they can roughly be divided into two kinds. First the inner solar system with the terrestrial planets Merkury, Venus, Earth and Mars at mean distances from the sun 0.4AU, 0.7AU, 1AU and 1.5AU respectively. These are small rock planets with masses from 0.1-1$M_\oplus$ (earth being the largest). Then comes the outer solar system with the giant planets Jupiter at 5AU, Saturn at 10AU, Uranus at 30AU and Neptune at 30AU. They can in turn be divided up in the gas giants Jupiter and Saturn and the ice giants Uranus and Neptune. All the four giants are believed to have dense cores of ice and rock of about 10$M_\oplus$ but Jupiter and Saturn has total masses 317 $M_\oplus$ and 96 $M_\oplus$ and an abundance of H and He of 90% and 80% respectively while the same numbers for Uranus and Neptune are 15 $M_\oplus$, 17$M_\oplus$ and 5-20% H and He. Other objects in the system are outside Mars orbit the asteroids, that is thought to be material that could not form a 5th terrestrial planet because it was to close to Jupiter, the Kupier belt at 35-500AU (see Kenyon 2001) and Orts cloud outside 10^4AU.

The planets orbit the sun in prograde direction on near circular orbits lying in the plane of the suns equator, except for Pluto having an orbit with eccentricity 0.25 and inclination 17°. In the sun-planets system most, 98%, of the angular momentum lies in the planets orbits while they have only 0.2% of the systems mass.

Except from Venus, Uranus and Pluto the planets rotates around their own axis in the same direction as their orbit around the sun with obliqueness $< 30°$.

All the planets and some of the asteroids are differentiated with heavy elements sunken into their cores. They must sometime in their past had very warm interiors and since this seems to be a rule for all the planets such a past should be a phase in their formation.

Looking at surfaces of some planets and satellites in the solar system gives us a hint of a violent past while they are covered with craters. If the rates of impacts on the planets always had been the same as today the age of the solar system would not have been enough to form al the craters we see.

2.2 Starbirth and Protoplanetary Discs

The beginning of a planet system starts with the birth of a star. Stars are born in big interstellar clouds mostly made of hydrogen and helium and often as in the case for our sun and important for us and our planets: a smaller fraction of heavier elements produced in a star that has ended its life in a supernova explosion spreading the elements in interstellar space where it can be recycled in the formation of a new star. A star starts to form in the cloud when it by some perturbation gets so dense that the pressure cant stop it from collapsing under its own gravity. As the cloud contracts its rotation velocity speeds up to conserve angular momentum.

The process forming the star will also lead to the formation of a rotating disc around it. The matter near the centre of the cloud will feel the strongest gravitational forces and collapse faster then the outer part of the cloud so the collapse will go from the inside and out. The gas in the central part has low angular momentum and can easily fall into the centre. The angular momentum for the outer parts is high and matter there can't lose their rotational velocity, the centrifugal force stops it from falling radially into the centre. Gravity then forces it to fall rotating into the plane of the forming suns equator. The gas stops when it comes to the equatorial plane where it meets gas falling in from other side of the eguator, and forms a rotating disc. The released gravitational energy goes into heating the disc. The rotation in the disc follows nearly keplerian orbits so the velocity will decrease outwards. At orbits close to each other in a viscous disc, the difference in velocities will lead to collisions that makes the inner particles lose some energy making them spiral inwards while the outer gain in angular momentum moves outwards. The total effect of this will be that matter moves inwards in the disc and angular momentum spreads outwards.

To try and estimate the mass of the formed protoplanetary disc one assumes that the gas in the disc was a mix of elements with the same abundances as in the sun today. Then adding hydrogen and helium to the planets compositions as they are today to get a minimum value for the mass needed in the original disc to form our planet system. This minimum mass protoplanetary disc gives an estimated value of $0.02M_{\odot}$. This should really be seen as a minimum value for the protoplanetary disc's mass, lots of gas and matter has probably been lost during the evolution of the disc and its original mass could have been as much as about the same mass as the sun. The lifetime of the gas in the disc gives an important upper time limit for the formation of the gas giants which has to catch their gas before the gaseous disc disappears. The upper limit of this comes from

studies of young forming stars (see Schilling 1999; Weinberg, Becklin & Schneider 2003) and our theory of star formation. The sun is thought to have gone into a phase in its evolution where its luminosity was about 20-30 times larger then today and it then would have produced a stellar wind that ought to have sweep away all gas remaining in the disc. This would have happened when the disc was about 10^6 - 10^7 years old. The gas could have been lost even before this by photoevaporation by UV light from massive stars born near the sun so 10^7 is an upper limit for the formation of the gas giants.

Planets do not start to grow directly from the gas in the disc. The matter which grows into planets must condense out of the gas to grow grains that then by collisions grows into planetesimals, see the next section. The composition the of the growing planets then depends on what elements could condense out of the gas at the radius where they grow, which is depending on the temperature. The temperature in the disc decreases outwards, in the inner part where we today find the terrestrial planets, mostly metals and rock materials could condense. Between Mars and Jupiter an important temperature limit must have been, the one when it's cold enough (about 150K) for water and other ices to condense. The growing giantst' cores then not only could get other composition then the terrestrial, there where also more matter to grow from not only rock and metals but also ice. Ice flakes would also have been about 3 times as abuntant as rock in the disc (see Bennett, Donahue, Schneider & Voit, 1999).

2.3 Growth of Grains

When small drops of matter condense out of the gas they start to grow into grains by mutual collisions to eventually become kliometer sized objects. When the grains grows from centimeter to kilometer sized objects they will go through a dangerous phase where they are in a big risk of spiralling in towards the sun and being evaporated again. What happens is that the gas density in the disc decreases outwards so there will be a pressure gradient pointing outwards and thus the gas will feel a pressure force directed outwards which help the gas a little in balancing the inwards directed gravitational force so the gas will not have to move at keplerian velocity around the sun but about 0.5% slower then the keplarian speed. The smallest grainst' motion is still coupled to the motion of the gas and for them this is not a problem but when they grow they decouple from the gas and need to move at keplarian velocity in order not to spiral inwards. Growing grains moving faster then the gas will then run in a headwind of gas making them lose velocity which leads to the grains spiraling inwards. The effect of the gas headwind depends on their ratio of surface area to mass. While the area grows with the square of the radius and the mass grows with the cube of the radius, this ratio will go down as they grow. When they become large enough (about 1km sized), the effect on their motion from the headwind of gas will be neglectable. The strongest effect is when they meet about their own mass in gas in one orbit which at 1AU happens when they about 1m in size.

A 1m object at 1AU would spiral in to the sun in 100yr so the growth from centimeter sized grains to kilometer sized planetesimals must have gone fast for the surviving ones. The variations of velocities and spiralling in caused by the gas could be an important help in raising the collision rate for the grains thus makes them grow faster then they would have done without the gas.

2.4 Planetesimals

The growth of planetesimals to protoplanets goes through a phase called runaway growth when the largest planetesimals get their effective collision cross sectioned enlarged by their large mass gravitational effect on their neighbours that they will fast will grow by consuming everything in their reach.

The growth rate of the planetesimals is dependent on the mass density of objects in the disc ρ_s, their physical surface area, and the areas gravitational enhancement factor F_g

$$\frac{dM}{dt} = 2\pi\rho_s v R F_g \tag{1}$$

$$F_g = 1 + (\frac{v_e}{v})^2 \tag{2}$$

where v is the relative velocity between the masses m_1 and m_2 and v_e is their escape velocity from each other, given by

$$v_e = (\frac{2G(m_1 + m_2)}{r})^{\frac{1}{2}}. \tag{3}$$

When $v > v_e$ is the growth rate relatively slow but when a planetesimal grows large enough v_e will become much larger then v and the planetesimals runaway growth will begin. The gravitational enhancement factor can, during this phase, grow as large as 1000 but not much larger, while an object of that size stir up the velocities among its neighbours so to that degre that it by increasing v while growing stops F_g from becoming much larger. The runaway growth ends when the planetesimal has emptied the zone of matter that it can perturb into a colliding orbit with itself. In a minimum mass protoplanetary disc, this would lead to protoplanets with a minimum mass of $6\mathrm{M}_{moon}$ at 1AU and of $1\mathrm{M}_{\oplus}$ at the distance of Jupiter. The runaway growth produces radialy separated large protoplanets with low eccentricities while smaller objects remaining in the disc may have been perturbed into high eccentric orbits by the objects of the runaway growth.

2.5 Terrestrial Planets

The runaway growth phase made protoplanets with low eccentricities of the protoplanets to grow into the planets we know today they must have gravitationally perturbed each other into colliding orbits. During this phase the gravitational interactions scattered them and their velocity, making the relative velocities of the objects high and the growth into planets slow compared to the runaway growth phase. This last phase is characterised by collision of large objects that may lead either to accretion or to the break up of the protoplanets into smaller objects. A growing protoplanet like the one becoming our Earth is thought to have collided with several moon sized objects and one in the size of Mars. Our moon is thought to have formed in a big collision throwing up matter from the surface of the Earth at the impact that then got into orbit around the Earth and eventually formed the moon.

Simulations of the growth of protoplanets into terrestrial planets predicts the forming of two to five terrestrial planets (see Wetherill 1995). If the simulations has the gravitational effect of a Jupiter object at 5AU it gives the largest terrestrial planet at around 1AU, if Jupiter is removed the peak in mass spreads outwards. Simulations also show that the asteroids probably is a result of their proximity to Jupiter. Jupiter perturbs theire orbits to much for a planet to form and is also thought to scatter lots of matter

from this part out of the solar system. The total mass of the asteroid belt is today only a small fraction of the mass of the Earth. If we had not had Jupiter maybe a 5th terrestrial planet would have formed here.

2.6 Giants

As mentioned above the gas giants Jupiter and Saturn must have formed in 10^7 years, before the gaseous disc was lost. Giants start their growth in a similar way as the terrestrial planets but here there is also ices in the solid material for the growing planetesimals so planetesimals and protoplanets grow larger. The gravitational interaction on the gas from the sun is less out here and already the escapevelocity for an object of mass $1M_\oplus$ is enough for starting a slow gas accretion but still the accretion of solids dominates. The planetesimals goes through a phase of runaway growth, like the terretrial planets, before the growth rate off solids starts to slow down. As their mass increases the gas accretion rate increases and when they reach a mass off about $10M_\oplus$ the gas accretion rate exceeds that off solids and the planets enters a runaway gas accretion phase that last until the planet cleaned up all the gas in its feeding zone or the gas disc is cleared away from the disc. Jupiter is supposed to be able to form in about 10-20Myr in this way. The time here is not so well know as we don't know the surface density of the protoplanetary disc. When the runaway gas accretion was ended Jupiter was 100 times larger then today and a slow process of contraction and radiation of the released gravitational energy started until it shrinked to it's current size. Jupiter is the fastest giant to form as the growth rate is dependent on the surface density that decreases outwards in the protoplanetary disc. The period is longer for planets growing further out which gives them slower collision rates that also increases the growth time. That Jupiter formation is just in the limit of the lifetime of the gaseous disc is not so unreasonable as its gas to mass rate is higher then for Saturn that maybe had not so much time to catch all the gas during its runaway gas accretion phase. Uranus and Neptune never reached runaway gas accretion before the gas was lost.

2.7 Disc Instabilities and Formation of Giant Planets

The core accretion model described above is now the mainly accepted theory for the formation of all the planets in our solar system. Core accretion is not the only way planets can form in a disc, another model is that of planet formation by disc instabilities. This model fails to explain some major aspects of the solar system such that: the planets enhancement in condensable matter and small objects like asteroids, comets and moons. It is not anymore thought of as a likely explanation of the hole system but when it comes to giants it is still a possible alternative scenario to core accretion.

The main advantage with the disc instability model is that it is fast. The core accretion model predict Jupiter formation at times that are in the upper limit of the lifetime of the gaseous disc. Instead of forming a Jupiter planet in Myrs it forms gas giants in kyrs. For the disc to become gravitational unstable it has to be thin and have a surface density much higher then the minimum protoplanetary disc, but formation of Jupiter planets in Myrs timescales also requires relatively large surfacedensities which gives a disc that comes near the limit of being gravitationally unstable (see Boss 2001, 2002; Boss, Wetherill & Haghighipour 2002) .

Planet forms in a gravitational unstable disc by perturbations that locally makes a part of the disc dens enogh that it starts to collapse under its own gravity. It is possible

that the gas giants could have formed this way but it does not explain why they are enhanced in heavier elements compared to solar abundances and it is Jupiter that fits this formation model best. It does also not easily explain why the four giants have about the same mass in their cores. The ice giants does not have enough gas to be explained only by disc instabilities on possibility is that they formed by core accretion and the gas giants by disc instabilities. Or Uranus and Neptune could also had been formed by disc instabilities but they lost most of there gas by photo evaporation of extreme ultraviolet radiation light if there were massive stars formed near the sun. Such stars could have evaporated the gas in the disc outside 10AU in only 10^5yr. This would then explain the gas composition of the unprotected Uranus and Neptune far out in the disc and why Saturn at 10AU lost some of it gas compared to Jupiter at 5AU. In this model the four planets would originally had formed with masses $1M_J$, $2M_J$, $2M_J$ and $2.5M_J$. One argument for this model is that the cores of the giant planets might be smaller then we previous though which would make the gas capture slower in the core accretion model. Still core accretion is the most probable explanation also for the giant planets in our solar system. We know it must have occurred both inside (terrestrial planets and asteroids) and outside (kupier belt objects, Pluto) the giant planets and it gives the best explanation for the elemental composition of the planets and their interiors. But we now know of more planets then those in our solar system and because detected extrasolar planets are giants, the kind of planets that disc instabilities are good at forming, it is still an interesting model for planet formation as will be seen in the next section.

3 Extrasolar Planets

Our current theory of planet formation was developed only knowing about the planets in the system we live in. Even though we did not know if there where planets around any other stars it seemed probable that planets should be common while planet formation seems like a likely evolution around a forming star and we know how stars evolve pretty well. Then in 1995 the first planet around another star was detected! And it was followed by more detections of extrasolar planets so today we know of 102 extrasolar planets some of them in planet systems. These almost ten years we had of finding planets outside our own system has showed us worlds that are very unlike the one we know. That we only have find these kind of worlds is an effect of that the detection method can't find planetary systems like ours yet (see Karlsson 2002), we have just come to the limit where a Jupiter orbiting a nearby star should possibly be detectable. But these planets found, can we explain them with the theory developed by looking at our solar system and what can we learn about planet formation from them and future detections of extrasolar planets?

3.1 What Have We Found?

Because of the detection model our sample of detected extrasolar planets is biased to short periods and high masses. We have found planets of about 0.1 to 10 Jupiters mass but those with lower mass lying closer to their sun then Jupiter. $1M_J$ at 5AU is about the limit of our resolution today and no real Jupiter detected yet. Detected planets tend to be more common at small masses and high periods in the detectable range of these parameters. It's not necessarily unlikely that Jupiterlike planets could be common (see Lineweaver, 2002). Other features we found for detected extrasolar planets is that unlike in our solar system they have eccentricites equally spread over $0 < e < 0.7$ except for those

planets very close to the sun. Among those orbiting at very short periods the eccentriceas are low probably due to tidal forces from the sun. There also seem to be a pile up of short period planets at 0.05AU and none are found closer to the sun also non of the biggest planets are among these. A property in the stars that we have seen planets around is that they have higher metalicity (abundance of elements heavier then helium) then the average star metalicity. Our sun has a higher metalicity then the average and most star with detected planets have even higher metalicity then the sun.

3.2 Extrasolar Planets and the Formation of Our Solar System

We have now found giants that seem to be very unlike the ones we know of. No model for planet formation predicts giant planets close to the sun. But our models are based on a system where no such planets are found. Either they are formed there in some way we today don't have any good understanding of or they have formed further out and then moved inwards. That planets could migrate both outwards and inwards was thought of before we detected any extrasolar planets but it does not seem like it happened in our system. Migration did not predict giant planets close to the sun because the migration speed is excpected to increase with decreasing radius so migration needs a method to stop the planets from going all the way to the sun. They might stop by coming into the zone less then 0.1AU from the star which is clean from gas. This might be an explanation for the pile up at 0.05AU mentioned above but does not explain the giants found a bit further out. Also non of the heaviest planets are found here so either migration is very uneffective for these or it is so effective that they, if they migrate, don't stop but goes all the way into the sun.

If many giants are formed in the disc they could gravitationally perturb each other so that planets could both be thrown out of the system or inwards towards the sun. It seems likely that this has happened looking at the spread in eccentricities we see in extrasolar planets. Simulations with the core accretion model otherwise gives planets with low eccentricities.

That we find so many giants even some heavier then Jupiter gives us also some problems with the timelimit for forming gas giants. For forming Jupiter in our solar system we needed that the suns gasdisc was in the upper limit of its excepted lifetime. Disc instabilities had the advantage that it forms giants fast and probably could occur in a thin disc if it's density is high enough. It is still possible that disc instabilities is the way some of the extrasolar planets we se today have formed even if it never occurred in our solar system. One thing maybe speaking against the formation of extrasolar planets by disc instabilities is that it seems like these giants are common around stars with high metalicity. The disc instability formation is very insensible to the metalicity of the disc and can not explain this. Maybe high metalicity in the disc help giants forming by core accretion like in our solar system.

3.3 The Future

More observation of extrasolar planets will give us more clues to their formation. If we find that big giants are the common kind of planets then this would speak for a formation by disc instabilities as most discs are not believed to live long enough for runaway gas accretion. Specially not in starforming regions near young massive star with extreme ultraviolet radiation killing the gas of the discs. If it is found that giant planets really just form around high metalicity stars then it's more probable that the high metalicity play an

important role in core accretion then that these planets has formed by disc instabilities. One observational test to see if giant planets form by disc instabilities will be with by NASA Space Interferometer Mission after 2010 (see Boss 2002). Observation in star forming regions to measure the wobbling young stellar objects will do if a massive planet orbits around them. If these movements are seen already at young stellar objects being about 0.1-1Myr the planets must have formed by disc instabilities. If on the other hand signs for big planets are only seen on those being 10-20Myr or older it would be a sign of planet formation by core accretion.

We have no evidence for extrasolar terrestrial planets. We have the oposite indications for some stars: they are not likely to occur in those systems with a giant planet near the sun specially not if we think that these planet must have formed further out and then moved inwards. A Jupiter passing by Earths orbit on its way towards the sun would probably have been catastrophic for the likelihood of earth to form. If it had not been accreted by Jupiter it would have been badly disturbed in its orbit or even thrown away from the solar system. Still we think that the process of forming terrestrial planets should be something common to occur around other stars and terrestrial planets are probably common.

If it shows that giant gas planets like Jupiter only form around high metalicity stars it would mean that giants are not common. Which would mean that our solar system is not among the most normal ones. Maybe then terrestrial planets are the norm but in systems without gas giants that could mean that even more terrestrial planets could have formed at least out to the asteroid belt. Terrestrial planets in a system without a Jupiter could also have negative effects for our search for life around other stars as we think that Earth have been protected from bombardment of objects from the outer solar system by Jupiter.

Planet formation occurs on much to long timescales for us to really see a planet form but we know today much more then we did ten years ago we now both have pictures from Hubble Space Telescope of the discs stars being born where planets form and we are detection the result of planet formation around other stars. The detection of extrasolar planets has just begun and to summaries the future in a few words: it looks interesting.

Acknowledgements

I would like to direct my thanks to Beatrice Sundbäck for her help with correction of the english language and to Alessandro Romeo for all the encouraging enthusiastic support during the Hot Topics course.

References

Bennett, Donahue, Schneider, Voit, 1999, *The cosmic perspective*, Addison Wesley

Boss, 1997, Science, 276, 1836

Boss, 2002, Earth and Planetary Science Letters, 202, 513

Boss, Wetherill & Haghighipour, 2001, Icarus, 156, 291

Karlsson, 2002, Hot Topics in Astrophysics 2001/2002, Chalmers University of Technology and Göteborgs University

Kenyon, 2000, astro-ph/0010036, to appear in the Proceedings of Galactic Structure, Stars, and the Interstellar Medium, a NASA Legacy meeting

Kenyon, 2001, astro-ph/0112120, invited review for PASP, March 2002

Lineweaver, 2002, astro-ph/0201003, conforms to version accepted for publication in "Astrobiology"

Lissauer, 1993, ARA&A, 31, 129

Pater, Lissauer, 2001, Planetary Sciences, Cambridge University Press

Schilling, 1999, Science, 286, 66

Weinberg, Becklin, Schneider, 2003, Scientific Frontiers in Reasearch on Extrasolar Planets ASP Conference Series, 294

Wetherill, 1996, Icarus, 119, 219

What Life Can Exist on the Newly Discovered Extra-Solar Planets?

Eddie Berntsson

Göteborg University
SE-41296 Göteborg, Sweden
(gu99edbe@dd.chalmers.se)

*

Abstract

The aim of this paper is to analyze the implications of finding life on the newly found extra-solar planets. We will also examine some of these planets more closely and also try to determine whether or not life could have formed there. Which kind of stars that are most appropriate to host a habitable planetary system is obviously also of great importance, especially if we one day will be forced to leave Earth.

1 Introduction

For more than 300 years, the people of Earth have asked the question: Are we alone in the Universe? This paper will try to answer parts of this question, such as: Which kinds of species are able to exist out there? Some of the newly detected planets have been found to be quite hot due to the heat from inner contraction. Life can however still exist there, but only at certain atmospheric levels. To get an overview of the topics that will be dealt with, I would like to present this outline of the paper:

- What has been discovered?
- Candidates for life
    1. What is the habitable zone?
    2. Planets with orbits partially within the habitable zone
    3. Planets with orbits entirely within the habitable zone
- A look at some specific planets and planetary systems
- Planets around dangerous stars - no candidates for life

*Hot Topics in Astrophysics 2002/2003, Alessandro B. Romeo, Christoffer Petersson, Daniel Persson & Oscar Agertz (Eds.), Chalmers University of Technology and Göteborg University, 2003.

One thing we have to keep in mind is that no planets have been discovered visually, only indirectly. We should also keep in mind that any speculation about life-forms on these planets are based on the life-forms we know of today. New technical instruments such as the Hubble Space Telescope as well as new and bigger radio-antennas have been enormously helpful in detecting smaller objects in the sky than we could before. If we, for some reason, need to leave Earth in the future, it could be good to know which new planets that are most suitable for the continuation of life.If we look at planetary systems of older stars, we must be aware of the possibility of disasterous end stages.

2 What We Have Discovered

During the seventies, we got better and more powerful instruments to observe the sky. We also got better computers, so the observational data of the sky could be analyzed faster and better. As early as 1972, Pieter van de Kamp at Sproul Observatory reported a new discovery. Around the star ϵ Eridani, only 10.7 light-years away, something with a mass of about 5% of the solar mass, was orbiting at a distance of 7.7 AU with a period of 25 years. This was the first discovery of an extra-solar planet. Since then, more than 91 planet systems with 105 possible planets have been reported. The stars, which host these planets, are all main sequence stars. They are typically type F, G and K stars, since it is easiest to observe planets around these. The majority, or 55%, of the discovered planets are big gas planets. They are orbiting close to the star, less than 20 times the distance between the Earth and the Moon, or 5% of the distance between the Earth and the Sun.

There is a problem with the big eccentricity of these planets. The eccentricity exceeds that of Pluto and Mercurius (see Eq. (2)). Some other planets are orbiting around pulsar stars, and later we will discuss life on one of these planets. We will see that only 21 systems have orbits crossing the habitable zone, and only four of them are considered stable. We will later also see the importance of being inside the right temperature zone, in order to be able to form life, as we know it.

3 Candidates for Life

3.1 What is the Habitable Zone?

The habitable zone is the volume of space around a star where the heat from the star is such that water can exist in liquid form. The temperature must also allow an atmosphere to exist on the surface of the planets. The atmosphere is needed to protect life from, for example, the stellar wind and from high-energy particles. The distance of the habitable zone from a star can be calculated via the following equation:

$$\frac{H_{\text{star}}}{H_{\text{sun}}} = \frac{L_{\text{star}}}{L_{\text{sun}} * 4 * \pi} \tag{1}$$

where H is the distance in AU of the habitable zone from the star. For our Sun, the habitable zone lies between 0.7 AU and 1.5 AU.

We believe that life requires a liquid environment, and the best liquid that we know of, for this purpose, is water. The reason for this is that cell membranes, which separate living organisms from their surroundings, are destroyed at temperatures exceeding 150 degrees centigrade (see, e.g., Lindqvist 2002). On the other hand, life is predicted to be in a coma if the inner temperature is less than -10 degrees centigrade.

3.2 Planets with Orbits Partially within the Habitable Zone

The planets in Table 1 are all orbiting so that they are crossing into the habitable zone (see Extra-Solar Catalogue 2003).

Table 1: Planets Partially in the Habitable Zone.

planet	*AU*	*planet*	*AU*	*planet*	*AU*
Ups And b	0.829	HD 4208	1.69	HD 13498	0.78
HD 4203	1.09	HD 108874	1.07	HD 12681	0.83
HR 810	0.925	HD 128311	1.06	HD 169830	0.823
HD 40979	0.818	HD 82943	0.73	HD 114783	1.20
HD 147513	1.26	HD 20367	1.25	HD 150706	0.82
HD 177830	1.0	HD 210297	1.097	HD 142	0.98
HD 27442	1.18	HD 92788	0.94	HD 28185	1.0
HD 89744	0.88	HD 82943	1.16	?	?

However, all but four of these have highly eccentric orbits

$$ecc^2 = \frac{mainaxis^2 - minoraxis^2}{mainaxis} \tag{2}$$

Some of the problems this will cause will now be reviewed.

- Large variations in seasonal temperatures, more than 200 degrees centigrade in some cases. For the Earth, we have 80 degrees in northern Siberia and 0.6 degrees in Naru.
- Large variations in seasonal length. Short summers and long winters or vice versa. The difference can be larger than a factor of two.
- Variations in how many energy-rich particles and UV radiation that hit the planets.
- Unstable orbits, more easily affected by other large masses in the vicinity.

Hence, these planets are no good candidates, if we are looking for higher forms of life. We can however expect bacterial life if there is a surface for them to hide under. The chance of life in the atmosphere, for both solid and gas planets, is however considered to be negligible. If the assumption that the found planets, are all gas planets, then no life is believed to exist on the eccentric orbit planets.

3.3 Planets with orbits entirely within the habitable zone

Four of the newly discovered planets have almost circular orbits within the habitable zone. They have an eccentricity similar to the ones of Earth (0.017) and Venus (0.005).

The fact that these four planets orbit entirely within the habitable zones, of their stars, gives the following:

- Stable seasonal lengths
- Stable seasonal temperature
- Stable influx of UV radiation

Table 2: Planets with Orbits Entirely within the Habitable Zone.

planet	*Jupitermasses*	*AU*
HD 4208	0.81	1.69
HD 114783	0.9	1.20
HD 27442	1.43	1.18
HD 28185	5.6	0.88

- Stable orbit, less likely to change

All of this gives a greater possibility of the existence of higher life. The assumption of these planets is that they are all gas planets, so if life does exist on them, it must exist in the atmosphere. Life there can include organisms such as medusae, algae or ray-fish-like organisms. These organisms must breathe in different ways from terrestrial forms of life, since oxygen is unlikely to exist there.

4 A Look at Some Specific Planets and Planet Systems

4.1 "Normal" Planets

We take a closer look at HD 28185 and the planet systems of UPS Andromeda, ϵ Eridani and 55 Cygni. We also look at HD 150706, which is a Jupiter-like planet in a Venus-like orbit. All of these orbit stars, which are similar to our Sun. The following summary of planets is done on the basis of SIMBAD and the Extra-solar Planet Catalogue.

Table 3: Some specific planet systems.

Star	*SolarM*	*typ*	*pc*	*FE/H*	*Planet*	*JupiterM*	*AU*
ϵEri	0.8	K2V	3.2	-0.1	b	0.86	3.3
ϵEri	0.8	K2V	3.2	-0.1	c	0.1	40?
55 Cygni	1.03	G8V	13.4	0.29	b	0.84	0.115
55 Cygni	1.03	G8V	13.4	0.29	c	0.21	0.241
55 Cygni	1.03	G8V	13.4	0.29	d	4.05	5.9
HD 28185	0.99	G5	39.4	0.24	b	5.7	1.03
UPS And	1.3	F7	13.47	0.09	b	0.7	0.18
UPS And	1.3	F7	13.47	0.09	c	2.1	0.829
UPS And	1.3	F7	13.47	0.09	d	4.6	2.53
HD150706	?	?	?	?	b	1.0	0.82

- ϵ Eridani

  The planet ϵ Eridani B has a colder star than our Sun and has an orbital distance twice as the orbit of Mars. Since it receives less energy from its star than Mars, we can safely assume that it will be colder than Mars (see Eq. (1)). This planet will thus be very cold if it does not generate any internal heat. These internal heat sources would then be the only places where life could exist. For the planet ϵ Eridani

C, we can see that its orbit is close to the one of Pluto. It is therefore very likely that it is a frozen gas planet, devoid of life.

- 55 Cygni

  The planets 55 Cygni B and C are more or less boiling since they are very close to their star. They have thus a very hot atmosphere and no known form of life can exist there. 55 Cygni D has an orbit similar to the one of Jupiter, but is four times as massive. It can therefore be assumed that it is somewhat hotter than Jupiter due to the inner contraction heat. Life can only exist at atmospheric levels where temperatures and pressures are just right. The life we might expect on one of these planets would probably be similar to medusas and hydrocarbon-based bacteria.

- HD28185

  This planet is selected to represent a planet in an Earth-like orbit around a star similar to the Sun. This planet has a mass similar to the one of Jupiter and has an orbit entirely within the habitable zone. This is a gas planet with a temperature of about 50 degrees centigrade. Life on such a planet could be similar to life in the Earth's oceans. The only difference would be that the corresponding temperature would have to be somewhat hotter, due to the planet's internal contraction heat. Life could probably adapt itself to living in an atmosphere instead of in water, so we can assume that life could exist on this planet.

- UPS Andromeda

  See Fig. 1 for a comparison of the orbit of this planet with those of the inner planets of our Solar system. Planet b is a planet, with a mass of 70% of the Jupiter mass, orbiting at a distance from its star that is only $\frac{1}{3}$ of the orbital distance of Mercury. We can thus assume that no life exists there.

  Planet c has a mass of 2.1 Jupiter masses and lies close to the orbit of Venus. The high temperature of this planet means however that there is no chance for an atmosphere to be suitable for any existence of life.

  Planet d has an eccentric orbit crossing the habitable zone. In this case, large seasonal variations (short, hot summers and cold, dark winters, which are five times as long as the summers) would force life to hibernate under the surface for most parts of the year. On Earth, we have a lot of life-forms like this, so life could exist on UPS Andromeda D. This life would probably be similar to terrestrial algae.

- HD 150706

  This is a Jupiter in an orbit, similar to Venus', and thus relatively hot. If it had a star colder than ours, it would have a summer temperature below 100 degrees centigrade, indicating a possibility of life in the atmosphere. This life can not be similar to ordinary cell-based Earth-life, since they are destroyed at temperatures of 50 degrees centigrade. If the star is as hot or hotter than the Sun, it would be impossible for life-forms, similar to those known to us, to exist.

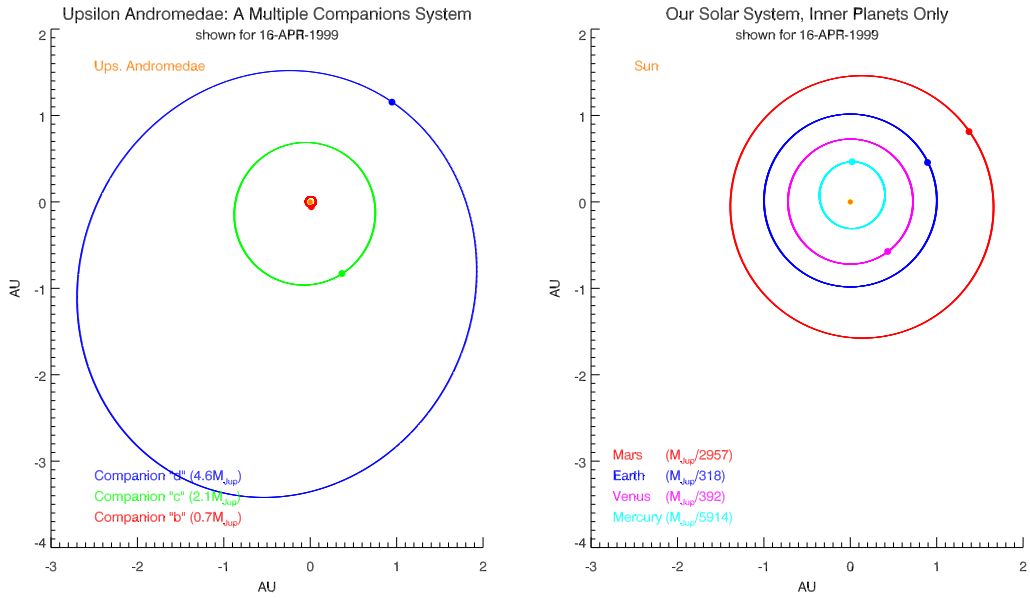

Figure 1: UPS-Andromeda compared to the Solar system.

4.2 Planets around a Dangerous Star - No Candidates for Life

An example of a planet around a dangerous star would be a planet in the system PSR 1257+12, which is a pulsar with five planets.

Table 4: Planet systems around dangerous stars.

planet	*Earthmasses*	*AU*	*Period*
a	0.015	0.19	25.34(d)
b	3.4	0.36	66.54(d)
c	2.8	0.47	98.22(d)
d	?	?	3(y)
e	100	40	170

All of these planets will be hit by high-energy particles and the innermost of them will also boil when the star expands its radius. The inner planets can be assumed to be sterile stones, with no life on the surface or inside. The outmost planet could harbor life, but since it is a gas planet, it seems unlikely.

5 Conclusions

In this paper, we have discussed which forms, life on extra-solar planets could take, if we assume that life will be created through the same processes that took place here on

Earth. On Earth, we think that life began under the surface, in volcano-heated spots. We must keep in mind that no planets have been observed directly and we still do not know anything of the surfaces or compositions of these planets. The knowledge we have is based on Kepler's laws, which gives us an indication of their masses, the planetary discs and also information about the density. With these theories, we can calculate which planets that are most suitable to form life and to live on. From these calculations, we find that there is a very small chance of finding life similar to life on Earth. We can thus stop waiting for UFOs and Martians, since the probability of their existence is far below one.

Acknowledgments

Thanks to Adam Torph at Chalmers for his help with the language in the text. I would also like to thank Christoffer Petersson for aiding me in editing this paper.

References

Bray D., 2001, *Cell Movements*, Garland Publishing

Lindqvist M., 2002, *Course Paper in Astrobiology*, Göteborg University

Voinova M., 2002, Course Paper in Living State Physics, Göteborg University

SIMBAD astronomical database, 2003, `http://simbad.u-strasbg.fr`

Extra-solar Planets catalogue, 2003, `http://cfa-www.harvard.edu/planets`

The Physics of SETI

Adam H. Thorp

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(f98adth@dd.chalmers.se)

*

Abstract

SETI is an astronomical programme with the goal of detecting intelligent life elsewhere in space. This report describes some concepts that are important to SETI, such as the Drake's Equation and Fermi's Paradox, as well as their consequences. Since SETI is mostly run through radioastronomy today, some of the radioastronomical aspects of SETI are also studied.

1 Introduction - What Is SETI?

SETI is an acronym for the Search for Extra-Terrestrial Intelligence. The purpose of this programme is to detect other forms of intelligent life in our universe. Today, the programme is run mostly through radioastronomy. We are trying to detect signals originating from intelligent civilizations in our Milky Way.

SETI is, quite possibly, one of the most important projects in science in general (as well as in radioastronomy in particular) of today. The consequences of a discovery of an extraterrestrial civilization would be very far-reaching indeed. The main obstacle facing SETI today is political short-sightedness: the American government's funding of SETI was cut after fifteen years, since no aliens had been discovered by then. This time-period should be compared with the time it took for intelligent life to develop on Earth - almost five billion years.

SETI is based on three hypotheses and their consequences. These hypotheses are:

- 1st hypothesis: Human intelligence is the result of physical and natural processes that apply throughout the universe.
- 2nd hypothesis: What has happened on Earth could have happened elsewhere.
- 3rd hypothesis: Human intelligence is not the ultimate of what the universe could provide.
- Consequence: There may exist more advanced forms of life in the universe.

*Hot Topics in Astrophysics 2002/2003, Alessandro B. Romeo, Christoffer Petersson, Daniel Persson & Oscar Agertz (Eds.), Chalmers University of Technology and Göteborg University, 2003.

- Observational Test: Undertake SETI.

SETI is not a single project, but rather a collection of independent projects. The most effort is directed towards radio-astronomical monitoring of the frequency band between 1GHz and 10GHz.

2 Historical Background

The first plans for communicating with other planets originated in the early 20th century, when scientists proposed cutting down trees in huge geometric shapes in the Siberian woods to show the inhabitants of Mars and Venus that we humans were intelligent enough to understand geometry and mathematics. The French magazine Le Monde offered a bounty of 100.000 francs to the person who could prove that they had been in contact with an alien civilization. Not Martians, though; they were to close to pose a challenge.

The modern SETI is generally considered to have begun in 1960. That year, Frank Drake undertook the first observation with the explicit purpose of detecting radio signals from another civilization. He performed his observation, which he named project Ozma, at NRAO, the National Radio Astronomy Observatory, in Green Bank, West Virginia of the United States of America. It was a two-week-long observation of two stars, Tau Ceti and Epsilon Eridani. One signal was detected, but it was later shown to be of Terran origin.

Several more observations were undertaken, but they were similarly limited in scope and time. During the seventies, NASA[1] entered the scene. Although NASAs participation in SETI was terminated for political reasons in 1993, it nevertheless changed the face of SETI permanently. Before NASA, SETI had been a concoction of small "backyard" projects, each monitoring only a few stars. NASAs grand projects were of a much larger scale, systematically monitoring hundreds of systems and tens of thousands of channels. The projects of today follow in this tradition.

In 1984, the SETI institute was founded. When NASA left SETI in 1993, the SETI institute took the reins and has been more or less leading the SETI effort since then. The SETI institute is a non-profit, private organization whose two main goals are education and science.

2.1 Frank Drake

Frank Drake is one of the pioneers of SETI, and also one of the greatest icons of the programme. He under took the very first SETI experiment, project Ozma, back in 1960. His most important contribution to SETI, however, is the so-called Drake's Equation, which is discussed in greater detail in Sect. 3. Today, Drake is a member of the Board of Trustees at the SETI institute.

3 Drake's Equation

Frank Drake's (see Sect. 2.1) greatest contribution to SETI is the so-called Drake's Equation, which gives us an estimate of the number of communicating civilizations in

[1] National Aeronatics and Space Administration, the United States space agency

Symbol	Quantity
N	Number of communicating civilizations
R	Rate of stellar formation
f_{planet}	Fractions of stars that have planets
n_{planet}	Average number of planets per star
f_{life}	Fraction of planets that gives rise to life
f_{int}	Fraction of planets with life that gives rise to intelligence
f_{comm}	Fraction of intelligent lifeforms that learn to communicate
L	Average lifespan of a communicating civilization

Table 1: The quantities of Drake's Equation.

Evaluator	Year	R	f_{planet}	n_{planet}	f_{life}	f_{int}	f_{comm}	L	N
Green, Bank	1963	1-10	.5	1-5	1	1	.1	$10^3 - 10^8$	$10^3 - 10^9$
Cameron	1963		1	.3	1	1	.1	10^6	$2{\cdot}10^6$
Sagan	1963	10	1	1	1	.1	.1	10^6	10^6
Shklovsky, Sagan	1966	10	1	1	1	.1	.1	10^6	10^6
Byurakan	1971	10	10	1		.1		10^7	10^6
Oliver	1971	20	.5	1	.2	1	.5		L
Rood, Trefil	1981	.05	.1	.05	.01	.5	.5	10^4	.003

Table 2: Different evaluations of Drake's Equation.

a volume of space. Drake's equation can be seen in Eq. (1). An explanation of the quantities involved can be found in table 1.

$$N = R \cdot f_{planet} \cdot n_{planet} \cdot f_{life} \cdot f_{int} \cdot f_{comm} \cdot L \tag{1}$$

Two things are immediately apparent in Drake's Equation. The first is that no area is specified, and that it thus can be applied to any volume of space. The second is that many of the quantities involved are unknown. Therefore, Drake's Equation is best viewed as a theoretical tool to help us estimate the number of civilizations, not as a formula to give us an exact count.

Several people have made their own estimates of the quantities in Drake's Equation and have achieved vastly differing results. Some example evaluations can be found in Table 2. As is clear from this table, the resulting number of communicating civilization N varies very strongly (twelve orders of magnitude!) depending upon the value of the different quantities in the equation. Until we get some sort of estimate of the different quantities in Drake's Equation, all calculations based on it will amount to nothing more than guesses.

4 Major SETI Projects

SETI is not a single project. Dozens of individual observations have been performed over the years, from Frank Drake's first observation (Sect. 2.1) to the modern seti@home.

4.1 SERENDIP

SERENDIP is an acronym for Search for Extraterrestrial Radio Emissions from Nearby Developed Intelligent Populations. It is run as a "piggyback" project, meaning that it has no dedicated observational equipment, but rather tags along on other observations.

The project has been running continuously since 1979. The first version, SERENDIP I, could monitor 100 channels at a time and was placed at the Hat Creek Observatory in Northern California. The latest version, SERENDIP IV is placed at the Arecibo array in Puerto Rico and is capable of monitoring 168 million channels simultaneously! The array is centered at the hydrogen frequency of 1.420GHz, and since the bandwidth of each individual channel is 0.6Hz, the system is constantly monitoring the sky in a frequency band of 1420±50GHz.

4.2 seti@home

seti@home (pronounced "seti at home") is another important project in the ongoing search for intelligent life elsewhere in our universe. Since there are so many individual channels to monitor (see section 5.4), the limiting factor in SETI is not observation and recording of data, but rather of processing it. The data is received through the Arecibo array, but it is then processed by computer users all over the world. You can download a screen-saver from the project website, which will then receive data packets from the central server, process it and pass it back. There are currently 1.45 million users in the project, and data is being processed at a rate of 41 TeraFLOP[2] per second.

5 Radioastronomy

As noted in section 1, SETI today mostly invests its efforts in radioastronomy, in the band spanning from 1GHz to 10GHz. The reasons for this mostly has to do with physics.

5.1 Why Electromagnetic Signals?

There are several ways, in theory, to transmit a message across interstellar space. The one we are most interested in today is electromagnetic radiation, but other candidates exist as well. However, many of these other candidates have drawbacks that make them less suited for this form of communication.

First, one might use matter. Transmitting a stream of particles across space is certainly possible. However, the laws of relativity limit the maximum velocity of such a transmission to below that of light, meaning a longer time for the message to reach its destination. Electrons and protons would be easily detectable if they were transmitted, but their interaction with the electromagnetic background fields of the Milky Way would scatter the beam and thus the message. Neutrons would not be scattered, but they would still be limited in their maximum velocity by relativity.

As an alternative, a civilization could attempt to use neutrinos as carriers of the message. Their weak interaction with matter would mean that they could pass easily through most obstacles in the Milky Way and they would thus have a long range. This weak interaction would make the message very difficult to detect and receive, however.

[2]FLOP, or Floating Point Operation, is the basic unit of computer processing. 1 TeraFLOP = 1.000.000.000.000 FLOP

5.2 Frequency Limitations

If we assume that electromagnetic radiation is the best candidate for interstellar communication, we still have to decide upon a frequency. At very low frequencies, say below 30MHz (wavelength λ greater than 1m), the signal would be drowned by the background noise of the rest of the universe. Conversely, if an extremely high frequency is used, quantum effects would "increase" the size of the individual photons in the transmission. This would increase the energy costs for transmitting a fixed amount of information.

Water vapour in the Earth's atmosphere imposes further limitations by absorbing high frequency transmissions. Today, SETI astronomers focus their attention in the frequency range of 1GHz to 30GHz (wavelength λ between 1cm and 30cm) in space. On Earth, the upper limit is further reduced to 10GHz (wavelength λ no less than 3cm).

5.3 Bandwidth

The narrower the bandwidth of a signal is, the higher the signal-to-noise ratio will be. However, narrowing the bandwidth in a given frequency band will mean more channels to monitor and hence an increase in the amount of computer power required to monitor the transmission. A bandwidth lower than 0.1Hz will not be practical, since interstellar dispersion of such a signal would make it undetectable.

5.4 What Frequency to Use?

If we assume a bandwidth of no less than 0.1Hz, we still have a staggering 100.000.000.000 frequencies to monitor in the band spanning from 1GHz to 10GHz. It would be far easier if we could monitor a few specific frequencies instead of whole bands. The problem here is one of psychology: what frequency would an alien race use? Today, there are two main candidates, the *hydrogen line* and the *cosmic microwave background frequency*.

Atomic hydrogen is the most common element in the universe. It consists of a single electron orbiting a single proton. The electron has two possible quantum states, called "spin up" and "spin down". When the electron changes its state, a frequency of 1.420GHz, corresponding to a wavelength of 21.106cm is generated. Since atomic hydrogen is the most common element in the universe, that frequency would constitute one possible frequency "standard".

Our universe has been observed to have a *background temperature* of 2.726 K. This temperature T can be calculated as a frequency ν according to Eq. (2).

$$k\mathrm{T} = h\nu \tag{2}$$

In Eq. (2), k is Boltzmann's constant from thermodynamics, and h is Planck's constant from quantum physics. This would be a "universal" frequency if k and h have the same value everywhere in the universe, which they should have according to modern science. If we insert a temperature T of 2.726K into the equation, we get a frequency ν of 56.9GHz. This frequency is very high, which would increase its energy cost as described in Sect. 5.2. It does have one advantage however.

As the universe expands, Hubble's law dictates that the cosmic background temperature decreases. However, the same law dictates that an electromagnetic wave will decrease its frequency as it is stretched along with the universe. That means that equation (2) will always be valid - T and ν will decrease at the same rate.

The hydrogen line and cosmic microwave background line are not the only possible options. An alien might choose to multiply his basic frequency with a constant, such as 2, π, e or something else. If we take all these possibilities into account, however, we find that the number of frequencies to monitor increase very rapidly.

The frequencies previously discussed share a significant disadvantage: they are both very active frequencies in the universe. The hydrogen line is defined as the frequency of the most common element in the universe, and the cosmic microwave background frequency. This means that a message transmitted at either of these frequencies would risk being drowned out by the background hum of the universe itself.

A solution to this would be for an alien to choose a channel in a frequency "hole", a region of the electromagnetic spectrum where there is very little natural activity, such as the "water hole" between 1.42GHz and 1.64GHz. The problem with this approach is that we are then back to monitoring whole regions of the electromagnetic spectrum.

6 Fermi's Paradox

> "If aliens exist, where are they?"

These words were uttered by Enrico Fermi, and they have been named after him. They are not a rhetorical joke, but rather an important insight into one of the biggest problems in SETI today: why do we not see any aliens?

Our galaxy, the Milky Way, is approximately ten billion years old and the Earth itself is close to five billion years old. However, intelligent life, capable of technology, has evolved here only in the last few hundred thousand years, which is an instant when it comes to astronomical timescales. It is certainly possible that equally (and possibly even more) advanced forms of life could have evolved elsewhere and well before us. They could be millions, or even billions, of years ahead of us technologically and if they were as curious, enterprising and expansionistic as we are, they should have colonized the entire Milky Way by now. So why do we not see them?

6.1 Interstellar Spacetravel

While warp travel, hyperspace engines and quantum jump drives only exist inside the minds of science-fiction writers, nothing in modern day physics prevents the construction of vessels traversing the interstellar gulfs between the stars at speeds of a few per cent of the speed of light. We can not build such ships today, but what will tomorrow bring?

If a civilization possessed the technological know-how to build a spacecraft capable of attaining velocities of about ten per cent of the velocity of light, they could spread throughout the galaxy at that speed. Since the Milky Way is approximately 100 000 light years across, one can estimate that the time required to colonize the entire Milky Way would be about one million years. That is a very short time in the history of the universe.

It would not be necessary for a civilization to send manned craft into space. Indeed, a large part of our own exploration of space has been performed by robotic probes. In 1966, the American scientist John von Neumann described a type of self-replicating machine. These machines are nowadays known as "von Neumann machines". They would be very well suited to space exploration. Their implementation would be:

1. Build von Neumann machine

2. Send machine to nearby star

3. Machine explores star system and transmits results back to originating civilization

4. Machine builds replica of itself

5. Machine and replica proceed to new stars

This process would give an exponential growth in the number of machines over time, and in the end we would have a mechanical plague of robotic probes all over our galaxy. The spread of these machines would most likely be even faster than the colonization process. The fact that we have observed no such probes today further strengthens Fermi's Paradox.

6.2 Explaining the Absence of Aliens

If non-human civilizations exist within our galaxy, there are four possible reasons as to why we have not encountered them.

According to the *Flying Saucer theory*, aliens *do* come here to Earth and visit us. All reports of flying saucers and other UFOs[3] are true reports of aliens visiting our world. The problem with this theory is that no UFO reports have been verified.

The *"Prime Directive" theory* is named after the Federation's Prime Directive in the television series Star Trek. The Prime Directive forbids interference in the normal development of a culture. This theory postulates that for some reason, every alien civilization has chosen to avoid (open) contact with us. The Earth is a zoo, or perhaps a historical landmark.

There is a possibility that *no civilization in our galaxy has built spacecraft capable of interstellar travel.* Either the task is beyond their capabilities, or they have chosen not to. As of today, we humans are certainly unable to build such vessels. However, our scientific knowledge does seem to indicate no theoretical obstacles to building such craft (see Sect. 6.1). If *all* alien civilizations for some reason have not to build such vessels, then the question remains: why?

The *eradication* theory is in many ways the saddest. According to it, all civilizations that have developed in our galaxy have shown themselves to be equally self-destructive as we humans are. They have either destroyed themselves, or each other.

7 Future Developments

What will happen to SETI in the future? After the government-sponsored heydays of the early years, the programme is today mostly run through private donations and non-profit organizations. This is unlikely to change, barring the unexpected discovery of a signal.

SETI has greatly profited from the vast increase in computer power over the past years. Since the project mostly consists of sifting through vast amounts of received data in hope of finding a signal, it is very well suited to automation. If computer power continues to increase, it will benefit SETI.

---

[3]Unidentified Flying Object, an acronym used to describe phenomena believed to be alien spacecraft.

7.1 What Will Happen if We Receive a Signal?

This is impossible to tell. There is significant speculation and discussion regarding this but as of today, it remains strictly hypothetical. Today, SETI is mostly a problem of radioastronomy and astrophysics. If we receive a signal, the project will rapidly branch into many other fields, including biological, sociological, psychological, political, theological and military disciplines.

Only time will tell ...

Acknowledgments

I would very much like to acknowledge the help I received from Albert Nummelin for the help I got in selecting the proper literature for this study.

References

Clark S., 2000, *Life on Other Worlds and How to Find It*, Praxis Publishing

Dick S.J., 1996, *The Biological Universe*, Cambridge University Press, Cambridge

Heidmann J., 1995, *Extraterrestrial Intelligence*, Cambridge University Press

Koerner D., LeVay S., 2000, *Here be Dragons*, Oxford University Press, Oxford

Murdin P. (Editor-in-Chief), 2001, *Encyclopedia of Astronomy and Astrophysics*, Nature Publishing Group, Institute of Physics Publishing

Zuckermann B., Hart M.H., 1995, *Extraterrestrials - Where Are They?*, Cambridge University Press, Cambridge

The SETI Institute, `http://www.seti.org`

The Search for Extraterrestrial Intelligence at U.C. Berkely, `http://seti.ssl.berkeley.edu`

seti@home, `http://setiathome.ssl.berkeley.edu`

Student's Workshop

HOT TOPICS In Astrophysics

2002/2003

Warm welcome for students & scientists to the students in astrophysics workshop!

Room F6306, Forskarhuset, 6th floor, Physics

Wednesday 14 May, 13:15-17:15

- Solar Activity and Terrestrial Climate
- Dynamics of the Crab Nebula
- Binary Pulsars
- Gravitational-Wave Astronomy - Theory and Observations
- Is the Fine-Structure Constant Really Constant?

Thursday 15 May, 09:30-11:50

- Planet Formation
- Life on Other Planets?
- The Physics of SETI

Questions? Contact Alessandro: romeo@fy.chalmers.se

STUDENTS' WORKSHOP
"Hot Topics in Astrophysics 2002/2003"

Wednesday 14 May and Thursday 15 May 2003

Room F6306 (at Chalmers, Physics, Forskarhuset, 6th floor)

ORGANIZER: Alessandro Romeo (romeo@fy.chalmers.se)

PROGRAMME

Wednesday 14 May

- **13:15–13:20**
  Welcome to the audience
  — Alessandro Romeo

Session 1 (Chairperson: Katarina Karlsson)

- **13:20–14:05**
  Solar Activity and Terrestrial Climate
  — Tommy Lindfors
- **14:05–14:50**
  Dynamics of the Crab Nebula
  — Oscar Agertz
- **14:50–15:35**
  Binary Pulsars
  — Gautam Narayan

COFFEE BREAK

Session 2 (Chairperson: Eddie Berntsson)

- **15:45–16:30**
  Gravitational-Wave Astronomy
  — Daniel Persson
- **16:30–17:15**
  Is the Fine-Structure Constant Really Constant?
  — Christoffer Petersson

Thursday 15 May

Session 1 (Chairperson: Gautam Narayan)

- **09:30–10:15**
  Planet Formation
  — Katarina Karlsson
- **10:15–11:00**
  Life on Other Planets?
  — Eddie Berntsson
- **11:00–11:45**
  The Physics of SETI
  — Adam Thorp
- **11:45–11:50**
  Thanks to the audience
  — Alessandro Romeo