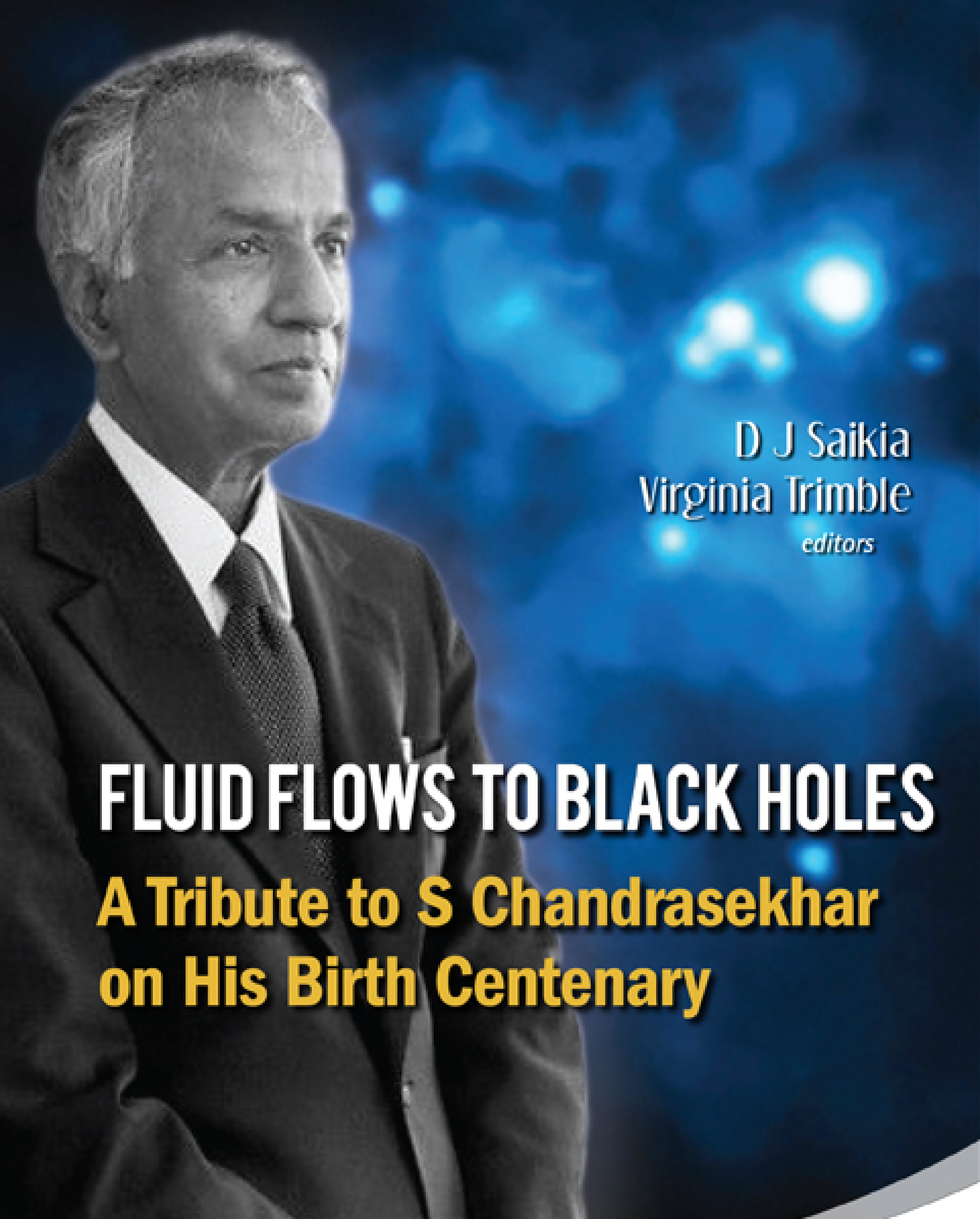


D J Saikia
Virginia Trimble
editors

FLUID FLOWS TO BLACK HOLES

A Tribute to S Chandrasekhar on His Birth Centenary

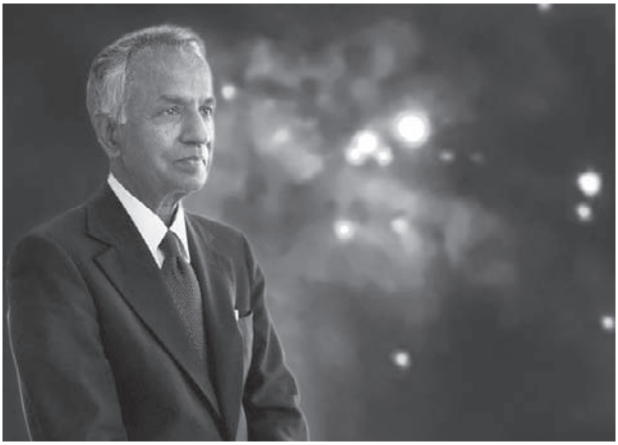
World Scientific

FLUID FLOWS TO BLACK HOLES

A Tribute to S. Chandrasekhar on His Birth Centenary

This page intentionally left blank

FLUID FLOWS TO BLACK HOLES



A Tribute to S. Chandrasekhar on His Birth Centenary

Editors

D. J. Saikia

Tata Institute of Fundamental Research, India

Virginia Trimble

University of California, Irvine, USA &
Las Cumbres Observatory Global Telescope Network, USA

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

FLUID FLOWS TO BLACK HOLES
A Tribute to S Chandrasekhar on His Birth Centenary

ISBN-13 978-981-4374-76-7
ISBN-10 981-4374-76-8

Printed in Singapore.

Preface

Professor Subrahmanyan Chandrasekhar was a legend well within his own lifetime, and continues to be so after his death on August 21, 1995. Thus it is not surprising that many institutions, in India, in the United States, and elsewhere chose to celebrate in various ways the 100th anniversary of his birth on October 19, 1910. One of these celebrations took place at the University of Chicago, where he joined the faculty in January 1937. He remained there for the rest of his life.

The Chandrasekhar Centennial Symposium at the University of Chicago was held from October 15–17, with the International Planning Committee being co-chaired by Professors Robert Wald and Kameshwar C. Wali. It included both scientific and personal reminiscences by students and colleagues who had known him quite well, and reports by other scholars on topics that he had pioneered. A special issue of the *Bulletin of the Astronomical Society of India* (*BASI*), published in March 2011, comprised written versions of some of those talks and additional articles covering topics not represented at the conference by other distinguished astrophysicists. This book contains a dozen articles published in this special issue of BASI, as well as a biographical portrait, his role in 20th century science and several personal reminiscences based on articles published in *Physics Today*, and an account of Chandrasekhar and the legacy of Ramanujan, and Chandrasekhar's impact on Indian astronomy.

One of us (VT) was the PhD student of Chandrasekhar's student Professor Guido Munch, and was lucky to know him at a level that entitled one to address him as Chandra. It has been a privilege for both of us to edit the issue of BASI and this book as a tribute to Professor Chandrasekhar on his birth centenary. We would like to take this opportunity to thank Ms Sandra Rajiva from the Indian Institute of Astrophysics and Ms Lakshmi Narayanan from World Scientific Publishing for their help in editing the special issue of BASI and this book respectively to celebrate Professor Chandrasekhar's birth centenary.

D. J. Saikia and Virginia Trimble
Editors

This page intentionally left blank

Chandrasekhar Centennial Symposium (October 15–17, 2010)

October 19, 2010 marked the 100th anniversary of the birth of Subrahmanyan Chandrasekhar. The symposium was an occasion for those who knew Chandra to commemorate his memory and work, and for those who did not know him to experience firsthand some of the impact he had on 20th century science. The Symposium began with a reception on the 15th evening followed by two full days of talks on October 16 and 17.

Coincidentally, October 15, 2010 was the 100th birthday of Lalitha Chandrasekhar. The opening reception was devoted to the celebration of her birthday. Knowing her fondness for Indian dance and vocal music, a local dance troupe (Natraj Dance Company) and musicians were invited to perform. Although she could not be present in person, her being in good spirits and health was the occasion of joy for many who knew her well. A scroll expressing best wishes for her birthday and the commemoration of Chandra's centennial from Martin Rees, President, the Council and Fellows of the Royal Society of London was presented and read for the occasion by Robert M. Wald. It was later presented in person to Lalitha.

International Planning Committee

Abhay Ashtekar, Naresh Dadhich, Valeria Ferrari, John Friedman, Giuseppe Mussardo, Jayant Narlikar, Roger Penrose, Saul Teukolsky, Robert Wald (co-chair) and Kameshwar C. Wali (co-chair).

Funding for the Symposium was provided by the National Science Foundation and by the Department of Astronomy and Astrophysics, the Enrico Fermi Institute, the Kavli Institute for Cosmological Physics, and the Department of Physics of the University of Chicago.

Speakers at the Symposium

Freeman Dyson: Chandra’s Role in 20th Century Science
Clifford Will: The Unreasonable Effectiveness of the Post-Newtonian Approximation.
Roger Penrose: Mathematical Properties of Black Holes and Colliding Plane Waves
Jayant V. Narlikar: Chandra’s Impact on Indian Astronomy
John Friedaman: Instabilities of Relativistic Stars
Kip S. Thorne: Black Holes
Valeria Ferrari: Gravitational Waves from Perturbed Stars
Martin Rees: Chandra’s Scientific Legacy
James M. Stone: Magnetohydrodynamics in Astrophysical Contexts
Priyamvada Natarajan: The Formation and Growth of Super-Massive Black Holes
Ganesan Srinivasan: Chandra and the Legacy of Ramanujan
Jeremiah P. Ostriker: Galaxy Structure and Formation
Rashid A. Sunyaev: Scattering of Radiation in the Universe: From the CMB and Last Scattering Surface to Clusters of Galaxies and Quasars
Gordon P. Garmire: The Chandra X-ray Telescope

Contents

Chandra: A biographical portrait[*]

Kameshwar C. Wali[†]

Department of Physics, Syracuse University, Syracuse, New York, USA

Abstract. The complexities of three countries — India, England, and the US — helped produce a scientist of rare stature and greatness.

The simple is the seal of the true. And beauty is the splendor of truth.

With the words above, Subrahmanyan Chandrasekhar, popularly known as Chandra, concluded his Nobel Prize lecture on 8 December 1983. Toward the end of his talk, he was describing black holes in the astronomical universe, explaining the simplicity in the underlying physics and the beauty of their mathematical description within the framework of Einstein's theory of relativity. "They are," he said, "the most perfect macroscopic objects there are in the universe."

The Nobel Prize in Physics brought an extremely private and somewhat shy individual into the limelight. For newspaper journalists and broadcast interviewers, neither the simplicity of the physics of the black holes nor the mathematical beauty of their description was of major concern; the pronunciation of Chandra's full name seemed to present them with an astronomical difficulty in and of itself.

The announcement of the prize he shared with William Fowler was greeted with joy and appreciation throughout the scientific world, and he was soon inundated with telephone calls, telegrams, and letters of congratulations and good wishes from his former students, associates, heads of scientific institutions, and governments. Most considered the prize belated and long overdue. But for Chandra, who had been critical of the atmosphere it creates — and of the ways in which some people seemed to go after it — it was to a large extent distorting to science and its true pursuit. He had never considered himself as a possible candidate, since his areas of research, pursued in a single-minded quest for a personal perspective, had not led him into areas that were in the science spotlight.

Lahore and Madras, 1910–25

Chandra was born on 19 October 1910 ("19-10-1910," as Chandra was fond of saying with a rare chuckle) in Lahore, Pakistan (then a part of colonial British India). His father, Chandrasekhara Subrahmanyan Ayyar, was in the government service, the deputy auditor-general of the North Western Railways. Chandra was the first son and the third child in a

[*] Reproduced with permission from *Physics Today*. © 2010 American Institute of Physics.

[†] Kameshwar Wali is the distinguished research professor emeritus.

Figure 1. Subrahmanyan Chandrasekhar, age 6.

family of four sons and six daughters. His mother, Sitalakshmi, was a woman of great talent and intellectual attainment. She married young and received only a few years of elementary schooling, yet she managed to continue her education while bearing 10 children and learned English well enough to adapt Henrik Ibsen's *A Doll's House* and translate a long story by Tolstoy into Tamil. Intensely ambitious for her children, she was to play a pivotal role in Chandra's career.

Being the first son, Chandra inherited the name of his grandfather, Ramanathan Chandrasekhar (referred to as R. C. hereafter). R. C. had been the first in the family to depart from traditional village life and pursue an English education. If, after graduating from high school in 1881, he had continued his college education and completed his BA degree as expected, he probably would have ended up in a high British government post. But he took his Western education seriously: He read English literature and philosophy extensively, studied mathematics and physics, and in general pursued what interested him most rather than what was required of him. This remarkable person,[1] who transformed the lives of his future generations, built a fine home library, which proved to be a very valuable resource to them. Chandra inherited not only his grandfather's name but also his independent streak in the pursuit of knowledge.

Chandra's early education was at home under the tutelage of his parents and private tutors. When he was 11 and the family had permanently settled in Madras (now Chennai), he began his regular schooling at the Hindu High School in the city's Triplicane neighborhood.

Chandra found formal school neither easy nor pleasant. His education at home under private tutors had allowed him the freedom to study what he liked (mainly English and arithmetic). Now he was suddenly required to study history, geography, and general science and was subjected to periodic examinations. It was a disappointing first year, but the promise of the following year's curriculum, which included algebra and geometry, was enough to get him excited. Without waiting for classes to begin, he got the books and studied on his own during the summer vacation. By the time he started his second year, he knew all the geometry and all the algebra the school was going to teach, and in fact more. He kept up his studying during the following three vacations and did extremely well in high school; he became a freshman at Presidency College in Madras when he was only 15 years old.

Those early years of learning were happy years for Chandra. Though the family was growing — a new child every two years or so — his father's income provided a comfortable life. In 1924, Chandra's father built his own house, named Chandra Vilas, in Mylapore, a prestigious suburb of Madras. And because he was in the railway services, he and the family received free railway travel or reduced fares, so they got together more frequently than they could otherwise. The children traveled to all parts of India — a privilege few Indians could afford. Grandfather R. C.'s efforts had paved the way for a new urban life for his children and grandchildren.

Education and urban life could not completely change centuries of tradition, however. Chandra's father, a highly cultivated individual, widely read and traveled, was a traditional father. He was authoritarian and demanded unquestioned obedience. Reserved and undemonstrative, he remained aloof from his children; they in turn could not share their innermost thoughts or feelings with him. Deeper connections were left to Chandra's mother. Sitalakshmi was the vital force of the family, keeping it together, helping the children with their studies, and meeting their needs. Without imposing strict religious discipline, she infused them with the cultural heritage and ideals of Hinduism. Chandra, the eldest son, held a special place in her heart.

Presidency College, 1925–30

Chandra's freshman and sophomore years (1925–27) proceeded smoothly. After he completed his second year with distinction in physics, chemistry, and mathematics, Chandra's next step was to work toward a bachelor's degree.

He wanted to take honors mathematics; he had not only excelled in his mathematics studies, he had long been under the spell of the legendary Srinivasa Ramanujan. Chandra was not quite 10 years old when his mother told him of the death of a famous Indian mathematician named Ramanujan who had gone to England some years earlier, collaborated with some famous English mathematicians, and returned to India only recently with international fame as a great mathematician. Ever since, Ramanujan was a source of inspiration.

Unfortunately, Chandra's father had different ideas. He wanted Chandra to aim for the Indian Civil Service examination to become an ICS officer in His Majesty's government. That was certainly the practical thing for such a brilliant young man to do. From his high-school days, however, Chandra had determined to pursue a career in pure science. He had as an example, his uncle Chandrasekhara Venkata Raman (popularly known as C. V. Raman), who had resigned a high level government post to pursue an academic and research career in physics. Although Chandra wanted to study pure mathematics, as a compromise to his father he opted to study physics and enrolled himself for a BSc honors degree.

The year 1928 proved an extraordinary time for Chandra. First of all, in February and March of that year, Raman, along with Kariamanickam Srinivasa Krishnan, made a fundamental discovery in the molecular scattering of light, later to become known as the Raman effect. Chandra spent the summer months in Calcutta, staying with Raman and working in the laboratory where the discovery was made. He knew enough theoretical physics to participate in the excitement and even explain to the experimentalists the significance of the discovery. He came to know Krishnan very well. Although 12 years apart in age, the two struck up a friendship that lasted through Krishnan's lifetime.

Soon after his return to Madras, Chandra learned from Krishnan that Arnold Sommerfeld was to visit India on a lecture tour in the fall and Madras and Presidency College were on his itinerary. For Chandra that was most exciting news — a rare opportunity to hear the famous man, especially since he had read Sommerfeld's book *Atomic Structure and Spectral Lines* and worked through it on his own. Chandra dreamed of meeting him, impressing him with his knowledge of atomic physics, and discussing plans for his research. Indeed, after the lecture in the science college, Chandra made arrangements to see him in his hotel room the following day. Chandra approached him with the brash confidence of a young undergraduate, but Sommerfeld shocked him by telling him that the quantum theory in the book was out-dated. He told Chandra about recent discoveries — Erwin Schrödinger's wave mechanics and the new quantum mechanics of Werner Heisenberg, Paul Dirac, Wolfgang Pauli, and others. Chandra had also studied classical Maxwell–Boltzmann statistics. Sommerfeld told him that too had undergone a fundamental change in the light of the new quantum mechanics. Seeing a crestfallen young student before him, Sommerfeld offered Chandra the galley proofs of his as yet unpublished paper that contained an account of the new Fermi–Dirac quantum statistics and its application to the electron theory of metals.

Chandra would later characterize that encounter as the "single most important event" in his life. He immediately launched on a serious study of the new developments in atomic theory. Sommerfeld's paper was sufficient for him to learn about Fermi–Dirac statistics and prepare,within a few months, a paper entitled "The Compton Scattering and the New Statistics." Chandra thought it significant enough to merit publication in the *Proceedings of the Royal Society.* But the society required the papers to be communicated by a fellow, a member of the society. As he was browsing through the newly arrived journals in the university library, he had came across Ralph Fowler, a fellow, who had just published his pioneering paper on the theory of white dwarfs based on the new Fermi–Dirac statistics. So Chandra sent his paper to Fowler, who agreed to communicate it and got it published. That chance circumstance was to have a profound influence on Chandra's future scientific career.

Along with his studies, Chandra continued his research, and by the end of his second undergraduate year he had a formidable list of papers to his name. His final year in college was equally eventful. First Heisenberg came through Madras on a lecture tour in October 1929. Krishnan had put Chandra in charge of showing Heisenberg around Madras. A day alone with the famous Heisenberg was an exhilarating experience for young Chandra. In addition, his activities and his prominence in his studies and research had attracted the attention of Lalitha Doraiswamy, a fellow undergraduate who would become his wife. A few months later, in January 1930, he attended the Indian Science Congress, Association

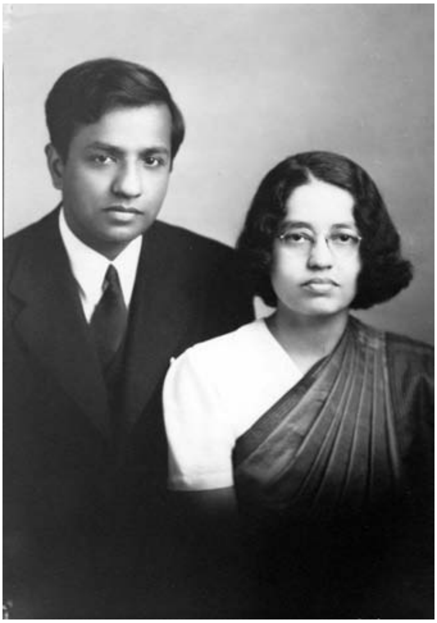


Figure 2. Subrahmanyan Chandrasekhar and his wife, Lalitha, in Williams Bay, Wisconsin, circa 1940.

meeting in Allahabad. He met the celebrated astrophysicist Meghnad Saha and his students and was pleasantly surprised to know his work had become well known. Chandra had the honor of being a dinner guest in the company of some distinguished senior Indian scientists. To top it all, on his return, he was called into the college principal's office. Principal Philip Fyson told him, in strict confidence, that he was going to be offered a Government of India scholarship to continue his studies in England. The scholarship was special, more or less created for him. On 22 May he received official notification that he had been awarded the scholarship and that he could proceed to make the necessary travel arrangements. Chandra decided to go to Cambridge University and study under the guidance of Fowler.

The opportunity to go abroad for advanced studies, ordinarily so difficult a matter both financially and logistically, came to Chandra so unexpectedly and so easily. Nonetheless, he had to face a difficult personal conflict. His mother had been ill since the summer of 1928, just before his encounter with Sommerfeld, and her illness had taken a serious turn. Although she had ups and downs, after two years of every kind of treatment it had become clear she was not going to get well again. If he went to England, he might never see her again.

Tradition and pressure from friends and relatives mounted against leaving his mother in such a condition. But Sitalakshmi herself intervened. Her insistence and persuasion and her solemn desire not to stand in the way of his future prevailed. With her promise of getting well, she persuaded the reluctant Chandra to proceed. Chandra left India on 31 July 1930, leaving behind a loving and caring family, and Lalitha.

Cambridge and Copenhagen, 1930–33

Before leaving India, Chandra had studied Fowler's paper more carefully and further developed the theory of white dwarfs to obtain a more detailed picture of them. On his long voyage from India, as a result of musings and calculations, he found Fowler's theory needed modifications to include special relativistic effects that led to a startling conclusion: *There was an upper limit on the mass of a star that could become a white dwarf in its terminal stage* (see the article by Freeman Dyson, page 13), and that limit could be expressed in terms of fundamental atomic constants.

When Chandra met Fowler for the first time, he handed over two papers, one he had completed in India and the other about the startling discovery he had made on his journey. Fowler was extremely impressed with this young, new student who exhibited so much independence and initiative. After some discussion, Fowler was quite pleased with the first paper, which had extended his own work. However, he was not so sure of the second paper. He offered to send it to Edward Arthur Milne, who Fowler thought was more familiar with the subject. After getting no response from Fowler or Milne for months and seeing no possibility of its publication in *Monthly Notices of the Royal Astronomical Society*, Chandra sent it to the *Astrophysical Journal* on 12 November 1930; it was published the following July.[2]

Among the lectures Chandra attended during his first year were Dirac's lectures on quantum mechanics. He had studied Dirac's book on his own, but he nevertheless attended the lectures faithfully, even though Dirac essentially copied onto the blackboard from his book. Dirac became his official adviser during the second term when Fowler left Cambridge on sabbatical, and Chandra came to know Dirac quite well. "He was very human, extremely cordial to me in a personal way," Chandra recalled. "Even though he was not very much interested in what I was doing, he used to have me for tea in his room in St. John's about once a month. He also came to my rooms for tea and, on some Sundays, used to drive me out to fields outside Cambridge where we used to go for long walks."

Chandra continued to do research on relativistic ionization and on stellar atmospheres and began a correspondence with Milne, who was quite receptive of his work. Milne's encouragement as well as his critical comments were of great help to Chandra during those early days. Within six months they had established a strong rapport, and Milne suggested collaboration and joint publications. Chandra's research efforts were recognized — he was elected to Trinity College's Sheepshanks Exhibition, a special honor bestowed every year to one candidate for proficiency in astrophysics, with an award of £40. He received a congratulatory note from Arthur Eddington with an invitation to meet him on 23 May 1931.

However, on 21 May 1931, Chandra received a devastating telegram:

Mother passed away Thursday 2PM Bear patiently.

Chandra used to write home twice a week to his father and at least once a month to his younger brother Balakrishnan, and also to his mother in Tamil; they were probably the only diversion from his routine. The letters to his father reveal in depth Chandra's life: his work, study; and leisure routines; his worries; his excursions and walks; the scientists around him; financial details (how he spent and saved); and his diet and his health (as indicated by

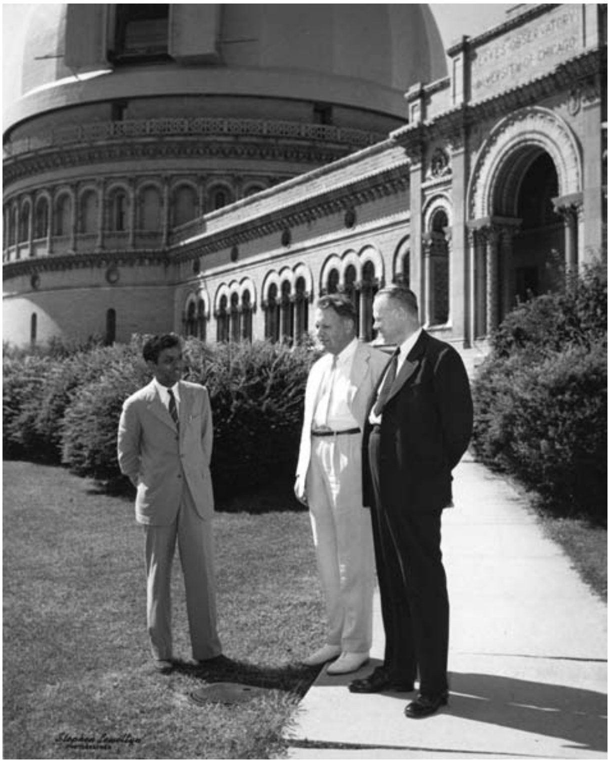


Figure 3. Yerkes Observatory, part of the University of Chicago and located in Williams Bay Wisconsin, was home to Subrahmanyan Chandrasekhar (left) for 27 years.

his weight). His mother's health was constantly on his mind, and in every letter he inquired about it. As her health went through rapid ups and downs, he always hoped she would get better. But that was not to be.

Alone, with no one to share his grief, he went to the riverbank, sat, and wept. Bear patiently, he told himself. He kept his appointment with Eddington two days after the news, received congratulations, and discussed his work, all the while feeling empty inside.

Work was the only panacea for loneliness and grief. He was working on stellar coefficients of absorption with Milne and had plans to spend the summer in Oxford. But after the news of his mother's death, he felt the need for a change from the drudgery of Cambridge and the past 11 months of ceaseless study and research. He thought a few months of diversion on the continent would provide the necessary relief. He spent the summer at the Institute for Theoretical Physics in Göttingen, Germany, where Max Born was working. Although the summer was supposed to be a vacation for Chandra, it became mostly a change of place and a change of study topics. But it helped to broaden his circle of friends on the continent. He returned to Cambridge in early September to begin his second academic year.

As he continued his research and piled up publications, some in collaboration with Milne, a conflict slowly brewed in Chandra's mind. He was in astrophysics by sheer chance — on his own, he had found a problem to work on. He began to have nagging doubts about the value of the work he was doing in astrophysics, as he was not receiving any encouragement in Cambridge. What about his true love — pure mathematics? He had come into physics due to the insistence of his father. Failing to pursue pure mathematics, he felt he would be happier if he could devote himself to pure physics, which he saw as the frontier field in which fundamental discoveries were taking place. The star-studded Cavendish Laboratory was the center of activity.

Dirac was his mentor, but because of a sense of loyalty to Fowler, he hesitated to tell Dirac he wanted to switch fields from astrophysics to theoretical physics. Finally, toward the end of his second year, he revealed to Dirac how unhappy he was in Cambridge and with what he was doing. Dirac, not being in astrophysics, was in no position to convince him otherwise. But he was very nice and understanding, and sympathized with Chandra's situation. He strongly urged Chandra to go to Niels Bohr's Institute for Theoretical Physics in Copenhagen, where he would find a better climate with friendly men who, though younger, were "big men" in physics.

Chandra took Dirac's suggestion and spent his final year in Copenhagen. The atmosphere at the institute was indeed, as Dirac indicated, quite unlike Cambridge. It was extremely friendly and truly international. Chandra found himself in a group of enthusiastic young people, including Max Delbrück, George Placzek, Victor Weisskopf, E. J. Williams, and Léon Rosenfeld. Chandra was particularly drawn to Rosenfeld, from Belgium, whose fiancée was studying astrophysics. There were also frequent visitors, including Oskar Klein and Heisenberg. With new friendships, tea every Sunday at Bohr's house, and walks and bicycle rides in the country, Chandra's life took on a new communal dimension.

He was also happy to be working on a problem in physics that Dirac had suggested: generalizing Fermi–Dirac statistics to more than two particles. Unfortunately, that did not work out. Chandra believed he had solved the problem and wrote a paper titled "On the Statistics of Similar Particles." Bohr and Rosenfeld read the paper and Bohr communicated it for publication to the *Proceedings of the Royal Society*. But Dirac found an error and convinced Chandra that he had not solved the problem Dirac had suggested. The paper had to be withdrawn.

Chandra had hoped to change fields from astrophysics to pure theoretical physics, but his lack of success with Dirac's problem put an end, at least for then, to that idea. Physics was *the* fundamental science, and while Chandra socialized with physicists like Weisskopf, Delbrück, Hans Kopfermann and others who appeared to be at the hub of important discoveries, he was not part of their science.

As December came along, it became clear to Chandra that he had to get his thesis ready to get his degree before the end of his scholarship in August 1933. Back to astrophysics, he set himself to prepare for his thesis a series of papers on distorted polytropes. (A polytrope is gaseous material in equilibrium under its own gravity and in which the pressure and density have a power-law relationship.)

A fellowship and an encounter, 1933–36

The degree became just a formality. Fowler did not find it necessary even to read Chandra's thesis. Chandra felt his future to be bleak, however. His scholarship would end in August and he would be required to return to India as soon as he completed his PhD degree. He was also under pressure from his father to return, but there was no promise of a suitable position that would allow him to continue his research. He was determined to extend his stay in Europe. He would seek support from Cambridge and Copenhagen; if nothing materialized, he had sufficient savings to stay at least six months anywhere in Europe. With little hope, he applied for a fellowship at Trinity College, a wild dream. If it came true, he would have four more years in Cambridge with free rooms in the college, dining privileges at the high table, and an allowance of £300 per year. Fowler was not very optimistic, though — the fellowship was open to candidates from all fields and the competition was formidable.

The dream did come true. The only other Indian who had been elected a Trinity fellow was Ramanujan some 16 years before. Chandra's Cambridge life became more enjoyable. He was no longer as lonely. He felt assured that his work would be appreciated. Astrophysics was going to be his predominant area of research, at least for the next four years. As a Trinity fellow, he could become a fellow of the Royal Astronomical Society on his own merit and did so without much ado. A trip to London to attend RAS meetings every second Friday of the month became a routine in his life and allowed him to make a mark on the tradition-ridden, hierarchical scientific surroundings. The Trinity fellowship also brought an opportunity to visit Russia during the summer of 1934.

The Russian visit renewed Chandra's interest in his own earlier work on the theory of white dwarfs. Neither Fowler nor Milne appreciated the startling discovery he had made. During the intervening years, he had occupied himself with other problems. In Russia, he gave talks about his white dwarf work, and Viktor Ambartsumian, in particular, was quite enthusiastic about his discovery. Ambartsumian suggested Chandra should work out the exact, complete theory devoid of some simplifying assumptions he had made.

During the fall months of 1934, Chandra involved himself in detailed, tedious numerical calculations in order to obtain as exact a theory of the white dwarf as one could construct within the framework of relativistic quantum statistics and the known features of stellar interiors. He accomplished the task by the end of 1934 and submitted two papers to the RAS. At the society's invitation, he presented a brief account of his results at the January 1935 meeting. His findings raised challenging and fundamental questions: What happens to the more massive stars as they continue to collapse? Are there other terminal stages different from white dwarfs? Instead of getting appreciation and recognition for a fundamental discovery, Chandra unexpectedly faced what amounted to a public humiliation. No sooner had he presented his paper than Eddington, who had been his mentor and who had followed his work closely, ridiculed the basic idea of relativistic degeneracy on which Chandra's work was based. Eddington characterized the theory as amounting to reduction ad absurdum behavior of the star, tantamount to stellar buffoonery.[3] Chandra sought the support of eminent physicists, who without exception agreed that his derivations were flawless, but Eddington's authority prevailed among the astronomers as he continued to attack the theory.

Eddington's denunciation was a traumatic experience for Chandra. In the face of such opposition, he decided to gracefully withdraw from the controversy instead of engaging in a dogged fight. He stopped further work on the theory of white dwarfs and went on to research in other areas. As he said[4]

> I foresaw for myself some thirty to forty years of scientific work, and I simply did not think it was productive to constantly harp on something which was done. It was much better for me to change the field of interest and go into something else. If I was right, then it would be known as right. For myself, I was positive that a fact of such clear significance for evolution of the stars would in time be established or disproved. I didn't see a need to stay there, so I just left it.

More than two decades passed before the Chandrasekhar limit became an established fact. It has been hailed as one of the most important discoveries of the last century, since it paved the way to the discovery of the other two presently known terminal stages of stars: neutron stars and black holes.

A voyage to the New World, 1936

During the fall of 1935, Chandra received an offer of a lectureship at Harvard University from Harlow Shapley, director of the Harvard College Observatory. The appointment could begin in December or January and required at least three months' stay. Chandra accepted the offer and had a highly successful and productive first visit to America from 30 December 1935 to 25 March 1936. He attended the American Astronomical Society meeting at Princeton University, gave 10 lectures at the observatory, and at an invitation from its director, Otto Struve, visited the Yerkes Observatory in Williams Bay, Wisconsin. A future for Chandra in America seemed to chart itself without any effort on his part. He received two offers, one from Harvard and the other from Yerkes. At Harvard, he would join the Society of Fellows, and at Yerkes he would have a research associateship. He received the latter offer aboard ship during his return voyage to Cambridge, and it came directly from Robert Hutchins, the president of the University of Chicago, with a prepaid cable for his answer.

Chandra chose to accept the offer from Yerkes, persuaded by Struve's arguments in favor of close cooperation between a theorist like him and observational astronomers. With Gerard Kuiper and Bengt Strömgren (a good friend from Copenhagen) also coming to Yerkes, a formidable group of young theorists and observational astronomers was in the making. Thus Chandra felt that as far as his scientific career was concerned, his immediate future had been virtually settled for him. It was time to think of other matters before setting forth to America. He had been away from home for nearly six years and it was time to return. He planned a short trip of three months. There was also the matter of marriage. Chandra had met Lalitha when both were undergraduate students at Presidency College, and they had developed an "understanding" of a lifelong commitment. Though the intervening six years had raised concerns of their future, once they met again in Madras all the doubts and uncertainties vanished. They were married on 11 September 1936, and after one short month in Cambridge they set forth to America.

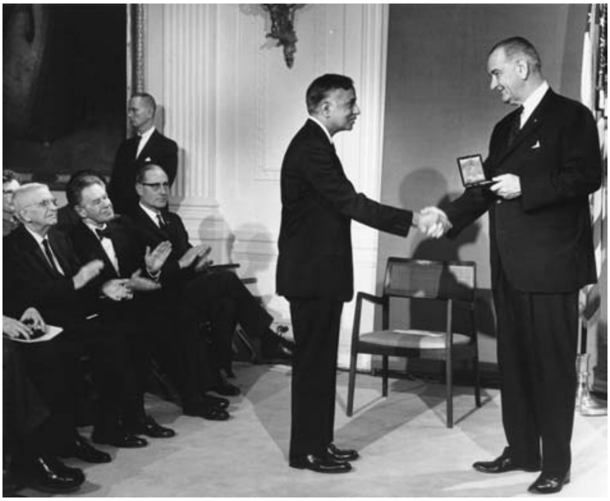


Figure 4. The National Medal of Science being presented to Subrahmanyan Chandrasekhar by President Lyndon B. Johnson in 1967.

Williams Bay and Chicago, 1937–95

Chandra and Lalitha, newly married, arrived in the US in 1937, and Chandra joined the faculty of the University of Chicago at Yerkes Observatory. He immediately took on the task of developing a graduate program in astronomy and astrophysics. It wasn't too long before his reputation as a teacher, his youth, and his enthusiasm for research began to attract students from all parts of the world. As a teacher and a lecturer, Chandra was a grand master who brought elegance and scholarship that literally charmed his listeners and kept them spellbound. He was also the sole editor of the *Astrophysical Journal* during the years 1952–71. He played a decisive role in transforming the journal which had been essentially the private property of the University of Chicago, into the national journal of the American Astronomical Society and one of the foremost astrophysics journals in the world.

Chandra and Lalitha lived in Williams Bay for the next 27 years. In 1964 they moved to the Hyde Park neighborhood of Chicago, near the university. Elected a fellow of the Royal Society of London in 1944, Chandra was named the Morton D. Hull Distinguished Service Professor at the University of Chicago in 1946 and remained at the university until his death in 1995.

The judgment of posterity

Chandra often told his life story as follows:

> I left India and went to England in 1930. I returned to India in 1936 and married a girl who had been waiting for six years, came to Chicago, and lived happily thereafter.

It may be so. But the Chandra one knows is the product of the complexities of three widely different countries: India, the land of his birth with its ancient culture and traditions, which undoubtedly influenced his early childhood and youth; England, the land of colonial masters, where his scientific research mushroomed and matured; and finally America, his adopted homeland, where he continued his research and became one of the foremost scientists of the 20th century.

Although there have been many scientists whose discoveries had perhaps greater impact and whose names have become more illustrious, in my opinion Chandra stands alone for his single-minded pursuit of his science and his devotion to the life of the mind. His extraordinary success in his scientific work was marked by an extraordinary effort, an intensity, a fervor for completeness, elegance, and above all else a personal, aesthetic perspective that extended beyond his well-known scientific papers and monographs. For example, when he was chosen for the University of Chicago's 1975 Ryerson Lecture, Chandra said that his preparations for his talk "Shakespeare, Newton, and Beethoven, or Patterns of Creativity,"

> consisted in reading several biographies of Shakespeare, his sonnets (in A. L. Rowse's editions) very carefully, and listening with the text (together with Ruth and Norman Lebovitz) to all the great tragedies (in their Marlowe editions); reading several biographies of Beethoven (particularly Turner's and Sullivan's); and similarly reading several biographies of Newton; besides, the lives of Rutherford, Faraday, Michelson, Moseley, Maxwell, Einstein, Rayleigh, Abel; and books and essays by Hadamard, Poincaré, and Hardy and the works of Keats and Shelley and most particularly Shelley's *A Defense of Poetry* and King-Hele's biography of Shelley.[5]

Chandra often quoted a letter from Milne:

> Posterity, in time, will give us all our true measure and assign to each of us our due and humble place. He really succeeds who perseveres according to his lights, unaffected by fortune, good or bad. And it is well to remember there is no correlation between the judgment of posterity and the judgment of contemporaries.

This first centennial celebration of Chandra's birth may or may not be the moment to determine the true measure of posterity. However, the Chandrasekhar Centennial Symposium, held at the University of Chicago in October, and this special issue of PHYSICS TODAY mark the beginning of that posterity's judgment to bestow on him his due place as a scientist of rare stature and greatness.

References

1. For more details, see C. S. Ayyar, "Family History" (1946), Subrahmanyan Chandrasekhar Papers, box 6, folder 4, Special Collections Research Center, University of Chicago Library.
2. S. Chandrasekhar, *Astrophys. J.* **74**, 81 (1931).
3. For details, see, for instance, K. C. Wali, *Chandra: A Biography of S. Chandrasekhar*, U. Chicago Press, Chicago (1991), p. 124.
4. Ref. 3, p. 146.
5. K. C. Wali, ed., *A Scientific Autobiography: S. Chandrasekhar*, World Scientific, Hackensack, NJ (in press).

Chandrasekhar’s role in 20th-century science*

Freeman Dyson†

Institute for Advanced Study in Princeton, New Jersey, USA

Abstract. Once the astrophysics community had come to grips with a calculation performed by a 19-year-old student sailing off to graduate school, the heavens could never again be seen as a perfect and tranquil dominion.

In 1946 Subrahmanyan Chandrasekhar gave a talk at the University of Chicago entitled “The Scientist.”[1] He was then 35 years old, less than halfway through his life and less than a third of the way through his career as a scientist, but already he was reflecting deeply on the meaning and purpose of his work. His talk was one of a series of public lectures organized by Robert Hutchins, then the chancellor of the university. The list of speakers is impressive, and included Frank Lloyd Wright, Arnold Schoenberg, and Marc Chagall. That list proves two things. It shows that Hutchins was an impresario with remarkable powers of persuasion, and that he already recognized Chandra as a world-class artist whose medium happened to be theories of the universe rather than music or paint. I say “Chandra” because that is the name his friends used for him when he was alive.

Basic science and derived science

Chandra began his talk with a description of two kinds of scientific inquiry. “I want to draw your attention to one broad division of the physical sciences which has to be kept in mind, the division into a basic science and a derived science. Basic science seeks to analyze the ultimate constitution of matter and the basic concepts of space and time. Derived science, on the other hand, is concerned with the rational ordering of the multi-farious aspects of natural phenomena in terms of the basic concepts.”

As examples of basic science, Chandra mentioned the discovery of the atomic nucleus by Ernest Rutherford and the discovery of the neutron by James Chadwick. Each of those discoveries was made by a simple experiment that revealed the existence of a basic building block of the universe. Rutherford discovered the nucleus by shooting alpha particles at a thin gold foil and observing that some of the particles bounced back. Chadwick discovered the neutron by shooting alpha particles at a beryllium target and observing that the resulting radiation collided with other nuclei in the way expected for a massive neutral twin of the proton. As an example of derived science, Chandra mentioned the discovery by Edmond Halley in 1705 that the comet now bearing his name had appeared periodically in the sky at

[*] Reproduced with permission from *Physics Today*. © 2010 American Institute of Physics.

[†] Freeman Dyson is a retired professor.

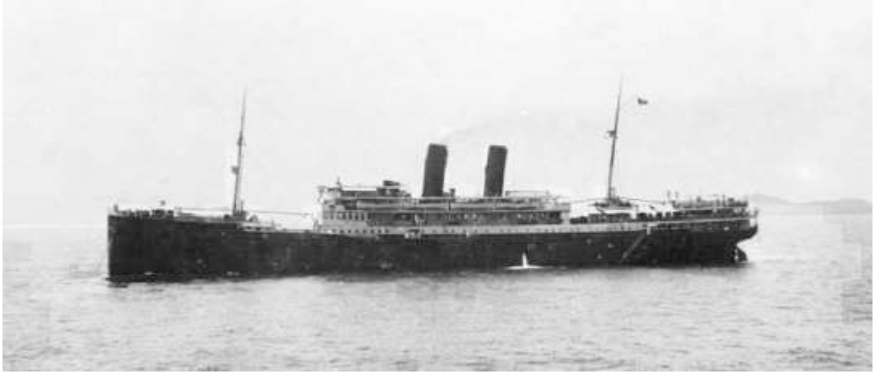


Figure 1. The SS *Pilsna*, a member of the Lloyd Triestino fleet, sailed from India to Europe in the early 20th century. In 1930 Subrahmanyan Chandrasekhar sailed on the ship on his way to study with Ralph Fowler at Cambridge University. En route, he refined an earlier calculation of Fowler's; the so-called Chandrasekhar limit implied by the new calculation was to have profound consequences. (©Penelope Fowler. Courtesy of Historical Photographs of China, University of Bristol.)

least four times in recorded history and that its elliptical orbit was described by Newton's law of gravitation. He also noted the discovery by William Herschel in 1803 that the orbits of binary stars are governed by the same law of gravitation operating beyond our solar system. The observations of Halley and Herschel did not reveal new building blocks, but they vastly extended the range of phenomena that the basic science of Newton could explain.

Chandra also described the particular examples of basic and derived science that played the decisive role in his own intellectual development. In 1926, when Chandra was 15 years old but already a physics student at Presidency College in Madras (now Chennai), India, Enrico Fermi and Paul Dirac independently discovered the basic concepts of Fermi–Dirac statistics: If a bunch of electrons is distributed over a number of quantum states, each quantum state can be occupied by at most one electron, and the probability that a state is occupied is a simple function of the temperature. Those basic properties of electrons were a cornerstone of the newborn science of quantum mechanics. They paved the way to the solution of one of the famous unsolved problems of condensed-matter physics, explaining why the specific heats of solid materials decrease with temperature and go rapidly to zero as the temperature goes to zero.

Two years later, in 1928, the famous German professor Arnold Sommerfeld, one of the chief architects of quantum mechanics, visited Presidency College. Chandra was well prepared. He had read and understood Sommerfeld's classic textbook, *Atomic Structure and Spectral Lines*. He boldly introduced himself to Sommerfeld, who took the time to tell him about the latest work of Fermi and Dirac. Sommerfeld gave the young Chandra the galley proofs of his paper on the electron theory of metals, a yet-to-be-published article that gave the decisive confirmation of Fermi–Dirac statistics. Sommerfeld's paper was a masterpiece of derived science, showing how the basic concepts of Fermi and Dirac could

Figure 2. Ralph Fowler, shown here in a 1931 photograph, wrote the seminal paper explaining the properties of white dwarf stars that inspired Subrahmanyan Chandrasekhar's revolutionary calculation. Fowler and other important English astrophysicists did not accept the validity of the new work. (Photo courtesy of the AlP Emilio Segrè Visual Archives, V. Ya. Frenkel collection.)

explain in detail why metals exist and how they behave. The Indian undergraduate was one of the first people in the world to read it.

Two years after his meeting with Sommerfeld, at the ripe old age of 19, Chandra sailed on the steamship *Pilsna* to enroll as a graduate student at Cambridge University. He was to work there with Ralph Fowler, who had used Fermi–Dirac statistics to explain the properties of white dwarf stars — stars that have exhausted their supply of nuclear energy by burning hydrogen to make helium or carbon and oxygen. White dwarfs collapse gravitationally to a density many thousands of times greater than normal matter, and then slowly cool down by radiating away their residual heat. Fowler's triumph of derived science included a calculation of the relation between the density and mass of a white dwarf, and his result agreed well with the scanty observations available at that time. With the examples of Sommerfeld and Fowler to encourage him, Chandra was sailing to England with the intention of making his own contribution to derived science.

A sea change

Aboard the *Pilsna*, Chandra quickly found a way to move forward. The calculations of Sommerfeld and Fowler had assumed that the electrons were nonrelativistic particles obeying the laws of Newtonian mechanics. That assumption was certainly valid for Sommerfeld. Electrons in metals at normal densities have speeds that are very small compared with the speed of light. But for Fowler, the assumption of Newtonian mechanics was not so safe. Electrons in the central regions of white dwarf stars might be moving fast enough to make relativistic effects important. So Chandra spent his free time on the ship repeating Fowler's calculation of the behavior of a white dwarf star, but with the electrons obeying the laws of Einstein's special relativity instead of the laws of Newton. Fowler had calculated that for a given chemical composition, the density of a white dwarf would be proportional to the square of its mass. That made sense from an intuitive point of view. The more massive the star, the stronger the force of gravity and the more tightly the star would be squeezed together. The more massive stars would be smaller and fainter, which explained the fact that no white dwarfs much more massive than the Sun had been seen.

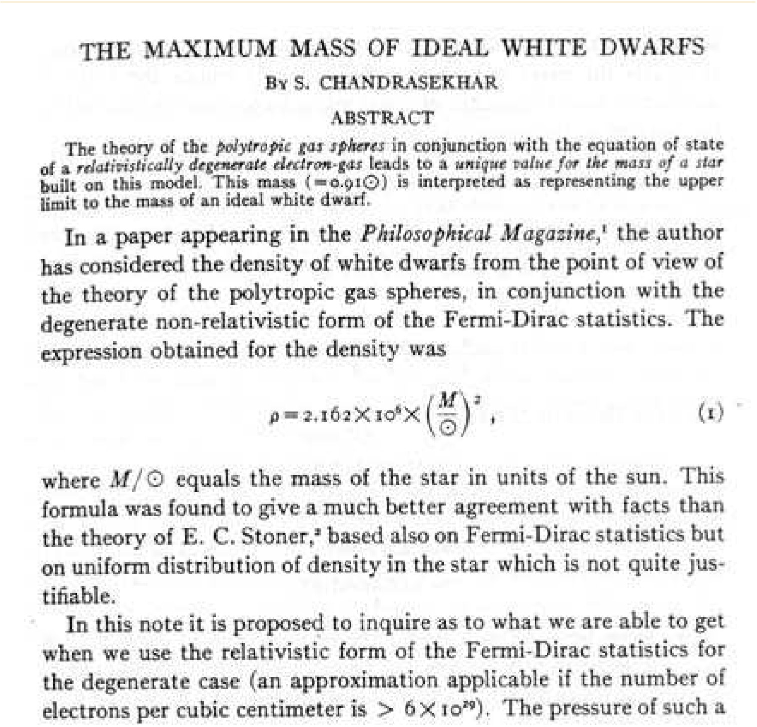
To his amazement, Chandra found that the change from Newton to Einstein has a drastic effect on the behavior of white dwarf stars. It makes the matter in the stars more compressible, so that the density becomes greater for a star of given mass. The density does not merely increase faster as the mass increases, it tends to infinity as the mass reaches a finite value, the Chandrasekhar limit. Provided its mass is below the limit, physicists can model a white dwarf star with relativistic electrons and obtain a unique mass-density relation; there are no models for white dwarfs with mass greater than the Chandrasekhar limit. The limiting mass depends on the chemical composition of the star. For stars that have burned up all their hydrogen, it is about 1.5 times the mass of the Sun.

Chandra finished his calculation before he reached England and never had any doubt that his conclusion was correct. When he arrived in Cambridge and showed his results to Fowler, Fowler was friendly but unconvinced and unwilling to sponsor Chandra's paper for publication by the Royal Society in London. Chandra did not wait for Fowler's approval but sent a brief version of the paper to the *Astrophysical Journal* in the US.[2] The journal sent it for refereeing to Carl Eckart, a famous geophysicist who did not know much about astronomy. Eckart recommended that it be accepted, and it was published a year later, Chandra had a coolhead. He had no wish to engage in public polemics with the British dignitaries who failed to understand his argument. He published his work quietly in a reputable astronomical journal and then waited patiently for the next generation of astronomers to recognize its importance. Meanwhile, he would remain on friendly terms with Fowler and the rest of the British academic establishment, and he would find other problems of derived science that his mastery of mathematics and physics would allow him to solve.

The decline and fall of Aristotle

Astronomers had good reason in 1930 to react with skepticism to Chandra's statements. The implications of his discovery of a limiting mass were totally baffling. All over the sky, we see an abundance of stars cheerfully shining with masses greater than the limit. Chandra's calculation says that when those stars burn up their nuclear fuel, there will exist no equilibrium states into which they can cool down. What then, can a massive star do when it runs out of fuel? Chandra had no answer to that question, and neither did anyone else when he raised it in 1930.

The answer was discovered in 1939 by J. Robert Oppenheimer and his student Hartland Snyder. They published their solution in a paper, "On Continued Gravitational Contraction."[3] In my opinion, it was Oppenheimer's most important contribution to science. Like



THE MAXIMUM MASS OF IDEAL WHITE DWARFS

By S. CHANDRASEKHAR

ABSTRACT

The theory of the *polytropic gas spheres* in conjunction with the equation of state of a *relativistically degenerate electron-gas* leads to a *unique value for the mass of a star* built on this model. This mass ($=0.91\odot$) is interpreted as representing the upper limit to the mass of an ideal white dwarf.

In a paper appearing in the *Philosophical Magazine*,[1] the author has considered the density of white dwarfs from the point of view of the theory of the polytropic gas spheres, in conjunction with the degenerate non-relativistic form of the Fermi-Dirac statistics. The expression obtained for the density was

$$\rho = 2.162\times 10^6 \times \left(\frac{M}{\odot}\right)^2 , \qquad (1)$$

where $M/\odot$ equals the mass of the star in units of the sun. This formula was found to give a much better agreement with facts than the theory of E. C. Stoner,[2] based also on Fermi-Dirac statistics but on uniform distribution of density in the star which is not quite justifiable.

In this note it is proposed to inquire as to what we are able to get when we use the relativistic form of the Fermi-Dirac statistics for the degenerate case (an approximation applicable if the number of electrons per cubic centimeter is $> 6\times 10^{29}$). The pressure of such a

Figure 3. Subrahmanyan Chandrusekhar’s discovery of a limiting mass for an ideal white dwarf appeared in a two-page paper published in 1931. The limiting value of 0.9 solar mass is different from the modern value, which is 1.5 solar masses. The difference results from Chandra’s using an obsolete estimate of the chemical composition of the star.

Chandra’s contribution nine years earlier, it was a masterpiece of derived science, taking some of Einstein’s basic equations and showing that they give rise to startling and unexpected consequences in the real world of astronomy. The difference between Chandra and Oppenheimer was that Chandra started with the 1905 theory of special relativity, whereas Oppenheimer started with Einstein’s 1915 theory of general relativity. In 1939 Oppenheimer was one of the few physicists who took general relativity seriously. At that time it was an unfashionable subject, of interest mainly to philosophers and mathematicians. Oppenheimer knew how to use it as a working tool to answer questions about real objects in the sky.

Oppenheimer and Snyder accepted Chandra’s conclusion that there exists no static equilibrium state for a cold star with mass larger than the Chandrasekhar limit. Therefore, the fate of a massive star at the end of its life must be dynamic. They worked out the solution to the equations of general relativity for a massive star collapsing under its own weight and discovered that the star is in a state of permanent free fall — that is, the star continues forever to fall inward toward its center. General relativity allows that paradoxical behavior because the time measured by an observer outside the star runs faster than the time measured by an observer inside the star. The time measured on the outside goes all the way from now to the end of the universe, while the time measured on the inside runs only for a few days. During the gravitational collapse, the inside observer sees the star falling freely at high speed, while the outside observer sees it quickly slowing down. The state of permanent free fall is, so far as we know, the actual state of every massive object that has run out of fuel. We know that such objects are abundant in the universe. We call them black holes.

With several decades of hindsight, we can see that Chandra's discovery of a limiting mass and the Oppenheimer–Snyder discovery of permanent free fall were major turning points in the history of science. Those discoveries marked the end of the Aristotelian vision that had dominated astronomy for 2000 years: the heavens as the realm of peace and perfection, contrasted with Earth as the realm of strife and change.

Chandra and Oppenheimer demonstrated that Aristotle was wrong. In a universe dominated by gravitation, no peaceful equilibrium is possible. During the 1930s, between the theoretical insights of Chandra and Oppenheimer, Fritz Zwicky's systematic observations of supernova explosions confirmed that we live in a violent universe.[4] In the same decade, Zwicky discovered the dark matter whose gravitation dominates the dynamics of large-scale structures. After 1939, astronomers slowly and reluctantly abandoned the Aristotelian universe as more evidence accumulated of violent events in the heavens. Radio and x-ray telescopes revealed a universe full of shock waves and high-temperature plasmas, with outbursts of extreme violence associated in one way or another with black holes.

Every child learning science in school and every viewer watching popular scientific documentary programs on television now knows that we live in a violent universe. The "violent universe" has become a part of the prevailing culture. We know that an asteroid collided with Earth 65 million years ago and caused the extinction of the dinosaurs. We know that every heavy atom of silver or gold was cooked in the core of a massive star before being thrown out into space by a supernova explosion. We know that life survived on our planet for billions of years because we are living in a quiet corner of a quiet galaxy, far removed from the explosive violence that we see all around using more turbulent parts of the universe. Astronomy has changed its character totally during the past 100 years. A century ago the main theme of astronomy was to explore a quiet and unchanging landscape. Today the main theme is to observe and explain the celestial fireworks that are the evidence of violent change. That radical transformation in our picture of the universe began on the good ship *Pilsna* when the 19-year-old Chandra discovered that there can be no stable equilibrium state for a massive star.

New ideas confront the old order

It has always seemed strange to me that the work of the three main pioneers of the violent universe — Chandra, Oppenheimer, and Zwicky — received so little recognition and acclaim at the time when it was done. Those discoveries were neglected, in part, because all three pioneers came from outside the astronomical profession. The professional astronomers of the 1930s were conservative in their view of the universe and in their social organization. They saw the universe as a peaceful domain that they knew how to explore with the standard tools of their trade. They were not inclined to take seriously the claims of interlopers with new ideas and new tools. It was easy for the astronomers to ignore the

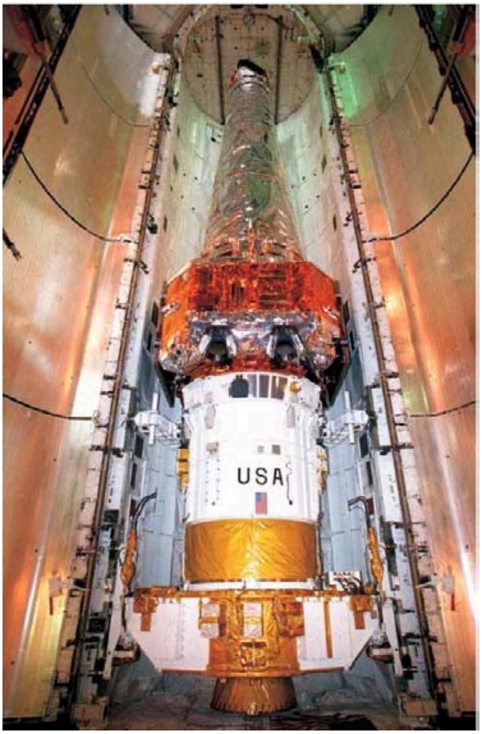


Figure 4. The Chandra X-ray Observatory is one of several telescopes casting an eye on the violent universe. *Chandra* is seen here loaded in the *Columbia* space shuttle a few days before its 23 July 1999 launch. (Courtesy of NASA.)

outsiders because the new discoveries did not fit into the accepted ways of thinking and the discoverers did not fit into the established astronomical community.

In addition to those general considerations, which applied to all three of the scientists, individual circumstances contributed to the neglect of their work. For Chandra, the special circumstances were the personalities of Arthur Eddington and Edward Arthur Milne, who were the leading astronomers in England when Chandra arrived from India. Eddington and Milne had their own theories of stellar structure in which they firmly believed; both of those were inconsistent with Chandra's calculation of a limiting mass. The two astronomers promptly decided that Chandra's calculation was wrong and never accepted the physical facts on which it was based.

Zwicky confronted an even worse situation at Caltech, where the astronomy department was dominated by Edwin Hubble and Walter Baade. Zwicky belonged to the physics department and had no official credentials as an astronomer. Hubble and Baade believed that Zwicky was crazy, and he believed that they were stupid. Both beliefs had some basis in fact. Zwicky had beaten the astronomers at their own game of observing the heavens, using a wide-field camera that could cover the sky 100 times faster than could other telescope cameras existing at that time. Zwicky then made an enemy of Baade by accusing him of being a Nazi. As a result of that and other incidents, Zwicky's discoveries were largely ignored for the next 20 years.

The neglect of Oppenheimer's greatest contribution to science was mostly due to an accident of history. His paper with Snyder, establishing in four pages the physical reality of black holes, was published in the *Physics Review* on 1 September 1939, the same day Adolf Hitler sent his armies into Poland and began World War II. In addition to the distraction created by Hitler, the same issue of the *Physics Review* contained the monumental paper by Niels Bohr and John Wheeler on the theory of nuclear fission — a work that spelled out for all who could read between the lines, the possibilities of nuclear power and nuclear weapons.[5] It is not surprising that the understanding of black holes was pushed aside by the more urgent excitements of war and nuclear energy.

Each of the three pioneers, after a brief period of revolutionary discovery and a short publication, lost interest in fighting for the revolution. Chandra enjoyed seven peaceful years in Europe before moving to America, mostly working without revolutionary implications, on the theory of normal stars. Zwicky, after finishing the sky survey that revealed dark matter and several types of supernovae, became involved in military problems as World War II was beginning; ultimate he became an expert in rocketry. Oppenheimer, after discovering the most important astronomical consequence of general relativity, turned his attention to mundane nuclear explosions and became the director of the Los Alamos laboratory.

When I tried in later years to start a conversation with Oppenheimer about the importance of black holes in the evolution of the universe, he was as unwilling to talk about them as he was to talk about his work at Los Alamos. Oppenheimer suffered from an extreme form of the prejudice prevalent among theoretical physicists, overvaluing pure science and undervaluing derived science. For Oppenheimer, the only activity worthy of the talents of a first-rate scientist was the search for new laws of nature. The study of the consequences of old laws was an activity for graduate students or third-rate hacks. He had no desire in later years to return to the study of black holes, the area in which he had made his most important contribution to science. Indeed, Oppenheimer might have continued to make important contributions in the 1950s, when black holes were an unfashionable subject, but he preferred to follow the latest fashion. Oppenheimer and Zwicky did not, like Chandra, live long enough to see their revolutionary ideas adopted by a younger generation and absorbed into the main stream of astronomy.

From stellar structure to Shakespeare

Chandra would spend 5–10 years on each field that he wished to study in depth. He would take a year to master the subject, a few more years to publish a series of journal articles demolishing the problems that he could solve, and then a few more years writing a definitive book that surveyed the subject as he left it for his successors. Once the book was finished, he left that field alone and looked for the next topic to study.

That pattern was repeated eight times and recorded in the dates and titles of Chandra's books. *An Introduction to the Study of Stellar Structure* (University of Chicago Press, 1939) summarizes his work on the internal structure of white dwarfs and other types of stars. *Principles of Stellar Dynamics* (University of Chicago Press, 1942) describes his highly original work on the statistical theory of stellar motions in clusters and in galaxies. *Radiative Transfer* (Clarendon Press, 1950) gives the first accurate theory of radiation transport in stellar atmospheres. *Hydrodynamic and Hydromagnetic Stability* (Clarendon Press, 1961) provides a foundation for the theory of all kinds of astronomical objects — including stars, accretion disks, and galaxies — that may become unstable as a result of differential rotation. *Ellipsoidal Figures of Equilibrium* (Yale University Press, 1969) solves an old problem by finding all the possible equilibrium configurations of an incompressible liquid mass rotating in its own gravitational field. The problem had been studied by the great mathematicians of the 19th century — Carl Jacobi, Richard Dedekind, Peter Lejeune Dirichlet, and Bernhard Riemann — who were unable to determine which of the various configurations were stable. In the introduction to his book, Chandra remarks,

> These questions were to remain unanswered for more than a hundred years. The reason for this total neglect must in part be attributed to a spectacular discovery by Poincaré, which channeled all subsequent investigations along directions which appeared rich with possibilities; but the long quest it entailed turned out in the end to be after a chimera.

After the ellipsoidal figures opus came a gap of 15 years before the appearance of the next book, *The Mathematical Theory of Black Holes* (Clarendon Press, 1983). Those 15 years were the time during which Chandra worked hardest and most intensively on the subject closest to his heart: the precise mathematical description of black holes and their interactions with surrounding fields and particles. His book on black holes was his farewell to technical research, just as *The Tempest* was William Shakespeare's farewell to writing plays. After the book was published, Chandra lectured and wrote about nontechnical themes, about the works of Shakespeare and Beethoven and Shelley, and about the relationship between art and science. A collection of his lectures for the general public was published in 1987 with the title *Truth and Beauty*.[1]

During the years of his retirement, he spent much of his time working his way through Newton's *Principia*. Chandra reconstructed every proposition and every demonstration, translating the geometrical arguments of Newton into the algebraic language familiar to modem scientists. The results of his historical research were published shortly before his death in his last book, *Newton's "Principia" for the Common Reader* (Clarendon Press, 1995). To explain why he wrote the book, he said, "I am convinced that one's knowledge of the Physical Sciences is incomplete without a study of the *Principia* in the same way that one's knowledge of Literature is incomplete without a knowledge of Shakespeare."[6]

Chandra's work on black holes was the most dramatic example of his commitment to derived science as a tool for understanding nature. Our basic understanding of the nature of space and time rests on two foundations: first, the equations of general relativity discovered by Einstein, and second, the black hole solutions of those equations discovered by Karl Schwarzschild and Roy Kerr and explored in depth by Chandra. To write down the basic equations is a big step toward understanding, but it is not enough. To reach a real understanding of space and time, it is necessary to construct solutions of the equations and to explore all their unexpected consequences. Chandra never said that he understood more about space and time than Einstein, but he did. So long as Einstein did not accept the existence of black holes, his understanding of space and time was far from complete.

When I was a student at Cambridge, I studied with Chandra's friend Godfrey Hardy, a pure mathematician who shared Chandra's views about British imperialism and Indian politics. When I came, Hardy was old and he spent most of his time writing books. With the arrogance of youth, I asked Hardy why he wasted his time writing books instead of doing research. Hardy replied, "Young men should prove theorems. Old men should write books." That was good advice that I have never forgotten. Chandra followed it too. I do not know whether he learned it from Hardy.

This article is based on a talk I gave for the Chandrasekhar Centennial Symposium at the University of Chicago on 16 October 2010.

References

1. S. Chandrasekhar, *Truth and Beauty: Aesthetics and Motivations in Science*, U. Chicago Press, Chicago (1987).
2. S. Chandrasekhar, *Astrophys. J.* **74**, 81 (1931).
3. J. R Oppenheimer, H. Snyder, *Phys. Rev.* **56**, 455 (1939).
4. See, for example, F. Zwicky, *Morphological Astronomy*, Springer, Berlin (1957), Secs. 8 and 9.
5. N. Bohr, J. A. Wheeler, *Phys. Rev.* **56**, 426 (1939).
6. S. Chandrasekhar, *Curr. Sci.* **67**, 495 (1994).
7. Ref. 2, reprinted in K. C. Wali, *A Quest for Perspectives: Selected Works of S. Chandrasekhar, with Commentary*, vol. 1, Imperial College Press, London (2001), p. 13.

Chandrasekhar and the legacy of Ramanujan

G. Srinivasan[*]

Raman Research Institute (retired), Bangalore, India

Received October 11, 2011

1. Introduction

Srinivasa Ramanujan is so far outside the circumference of my comprehension that I am naturally apprehensive to speak on this topic, particularly when there are authorities in the audience, like Professor Dyson who has *'played in Ramanujan's garden'* for over six decades. My only credential is that most of the things I will say, I heard personally from Chandra and Professor Richard Askey. May I, therefore, begin on the same note on which Chandra ended one of his most memorable lectures?

> *First, my fear; then my curtsy; last my speech.*
> *My fear, is your displeasure,*
> *My curtsy, my duty, and my speech, to beg your pardon.*
>
> Henry IV

I came to this department in 1965. My first encounter with Chandra was similar to what others had experienced. During one of the pre-colloquium coffee, when the students mingled with the greats of the department, I went up to Chandra and said *"I am a new student here"*. He said *"Yes, I know. You are taking my course on Statistical Mechanics"*. I thought I had broken the ice, and felt that I should push it along. I said *"I would like to meet you some time"*. He said *"well....?"* I had been warned about such a brush off, and was prepared for it. I said *"Nothing in particular, you know. I would like to chat with you some time"*. He said *"I shall let you know when I am free"*. Nothing happened for three months! And then, one day, he called me to his office around 5 o'clock; it was a Thursday, and the Colloquium had been cancelled because the speaker was stranded due to bad weather. When I entered his office, he was attending to something. So I stood there, absorbing everything I could see. Three or four framed photographs on the walls attracted my attention. I recognized two of them. He looked up at me and said *"Do you recognize them?"* He closed his PARKER pen, laid it carefully on the table, and asked me to sit down. He didn't talk about the photographs then; instead he asked me *"The OLD VIC is in town. Are you going to any of the plays?"* I said *"I have been to* **'Measure for Measure'**

[*] email: gsrini1942@gmail.com

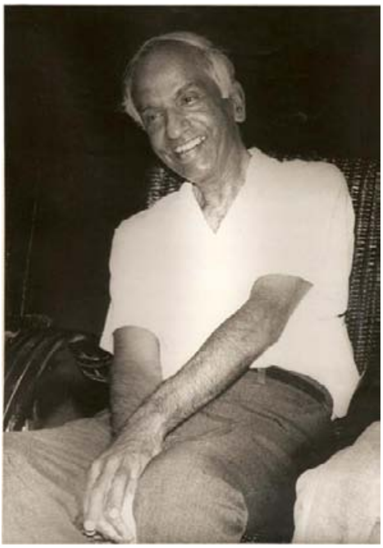


Figure 1. Subrahmanyan Chandrasekhar.

and **'Romeo and Juliet'**.*" "Well, then,"* he asked *"can you tell me when Juliet matured as a person?"* The iceberg had melted. We talked for two hours, **rather he did**! About *Alice, Virginia Wolf, Mozart ...* That was the beginning of a friendship that was to last till he passed away.

Two days later, it was a Saturday, he told me to meet him at the front steps of the Museum of Science and Industry on the lake shore. He took me to the Mathematics section, where among other things there was a gallery of portraits of all the great mathematicians. **And there was Ramanujan's photograph.** Chandra beamed and said *"If Ramanujan was Hardy's discovery,* **that** *is my discovery."* And then he told me about his discovery of 1936. I recall it as if it was yesterday. But since I cannot recall every word that was said forty five years ago – like Chandra could! – I shall read from a speech Chandra gave at the Royal Society in May 1994:

"Hardy was to give a series of 12 lectures on subjects suggested by Ramanujan's life and work at the **Harvard Tercentenary Conference of Arts and Sciences** in autumn of 1936. In the spring of that year, Hardy told me that the only photograph of Ramanujan available at that time was one of him in cap and gown, *'which makes him look ridiculous.'* And he asked me whether I would try to secure, on my next visit to India, a better photograph which he might include with the published version of his lectures. It happened that I was in India that same year from July to October. I knew that Mrs. Ramanujan was living somewhere in South India, and I tried to find where, at first without success. On the day before my departure for England in October 1936, I traced Mrs. Ramanujan to a house in Triplicane, Madras. I went to her house and found her living under extremely modest conditions. I asked her if she had any photograph of Ramanujan which I might give to Hardy. She told me that the only one she had was the one in the passport which he had secured in London early in 1919. I asked her for the passport and found that the

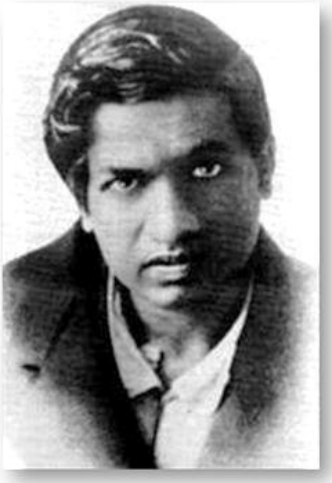


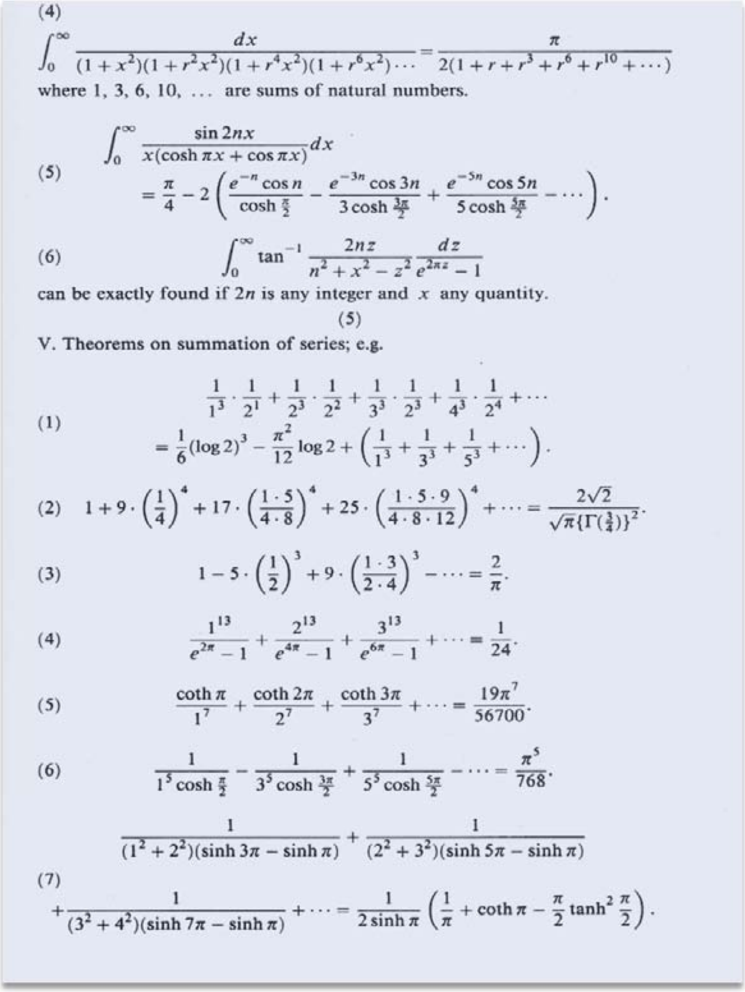
Figure 2. Srinivasa Ramanujan.

Figure 3. G.H. Hardy.

photograph was sufficiently good (even after 17 years) that one could make a good negative and copies. It is this photograph that appears in Hardy's book.... It is of interest to recall Hardy's reaction to the photograph: **'He looks rather ill** (and no doubt he was very ill): **but he looks all over the genius that he was.'**

Who was this person about whom Chandra was so excited? It is possible that there may be one or two young persons in the audience who may not have encountered the name of Srinivasa Ramanujan. Certainly in India, the way science is taught now is completely devoid of any sense of history. Let me, therefore stick my neck out, and make a few remarks.

Toward the end of January 1913, G.H. Hardy at Trinity College in Cambridge received a letter dated 16 January 1913. It began and ended as follows:



(4)

$$\int_0^\infty \frac{dx}{(1+x^2)(1+r^2x^2)(1+r^4x^2)(1+r^6x^2)\cdots} = \frac{\pi}{2(1+r+r^3+r^6+r^{10}+\cdots)}$$

where 1, 3, 6, 10, ... are sums of natural numbers.

(5)

$$\int_0^\infty \frac{\sin 2nx}{x(\cosh \pi x + \cos \pi x)} dx = \frac{\pi}{4} - 2\left(\frac{e^{-n}\cos n}{\cosh \frac{\pi}{2}} - \frac{e^{-3n}\cos 3n}{3\cosh \frac{3\pi}{2}} + \frac{e^{-5n}\cos 5n}{5\cosh \frac{5\pi}{2}} - \cdots\right).$$

(6)

$$\int_0^\infty \tan^{-1}\frac{2nz}{n^2+x^2-z^2}\frac{dz}{e^{2\pi z}-1}$$

can be exactly found if $2n$ is any integer and x any quantity.

(5)

V. Theorems on summation of series; e.g.

(1)

$$\frac{1}{1^3}\cdot\frac{1}{2^1} + \frac{1}{2^3}\cdot\frac{1}{2^2} + \frac{1}{3^3}\cdot\frac{1}{2^3} + \frac{1}{4^3}\cdot\frac{1}{2^4} + \cdots = \frac{1}{6}(\log 2)^3 - \frac{\pi^2}{12}\log 2 + \left(\frac{1}{1^3} + \frac{1}{3^3} + \frac{1}{5^3} + \cdots\right).$$

(2)

$$1 + 9\cdot\left(\frac{1}{4}\right)^4 + 17\cdot\left(\frac{1\cdot 5}{4\cdot 8}\right)^4 + 25\cdot\left(\frac{1\cdot 5\cdot 9}{4\cdot 8\cdot 12}\right)^4 + \cdots = \frac{2\sqrt{2}}{\sqrt{\pi}\{\Gamma(\frac{3}{4})\}^2}.$$

(3)

$$1 - 5\cdot\left(\frac{1}{2}\right)^3 + 9\cdot\left(\frac{1\cdot 3}{2\cdot 4}\right)^3 - \cdots = \frac{2}{\pi}.$$

(4)

$$\frac{1^{13}}{e^{2\pi}-1} + \frac{2^{13}}{e^{4\pi}-1} + \frac{3^{13}}{e^{6\pi}-1} + \cdots = \frac{1}{24}.$$

(5)

$$\frac{\coth \pi}{1^7} + \frac{\coth 2\pi}{2^7} + \frac{\coth 3\pi}{3^7} + \cdots = \frac{19\pi^7}{56700}.$$

(6)

$$\frac{1}{1^5\cosh\frac{\pi}{2}} - \frac{1}{3^5\cosh\frac{3\pi}{2}} + \frac{1}{5^5\cosh\frac{5\pi}{2}} - \cdots = \frac{\pi^5}{768}.$$

(7)

$$\frac{1}{(1^2+2^2)(\sinh 3\pi - \sinh \pi)} + \frac{1}{(2^2+3^2)(\sinh 5\pi - \sinh \pi)} + \frac{1}{(3^2+4^2)(\sinh 7\pi - \sinh \pi)} + \cdots = \frac{1}{2\sinh \pi}\left(\frac{1}{\pi} + \coth \pi - \frac{\pi}{2}\tanh^2\frac{\pi}{2}\right).$$

Figure 4. One of the many sheets in which Ramanujan had written down his theorems. These sheets were attached to the letter addressed to Hardy. None of the theorems had any proofs!

> *Dear Sir,*
> *I beg to introduce myself to you as a clerk in the Accounts Department of the Port Trust Office at Madras on a salary of only 20 pounds per annum. I am now 23 years of age. I have no University education but have undergone the ordinary school course. After leaving school I have been employing the spare time at my disposal to work at mathematics..... I would request you to go through the enclosed papers. Being poor, if you are convinced that there is anything of value I would like to have my theorems published..... requesting to be excused for the trouble I give you,*
>
> *I remain, Dear Sir, Yours truly,*
> *S. Ramanujan*

There were about twelve sets of theorems; **none of them had any 'proofs'**. As E.H. Neville has remarked *"No one who was in the mathematical circles in Cambridge at that time can forget the sensation caused by this letter."*

At first, Hardy thought the letter was a hoax. He sought the help of his close collaborator Littlewood. After pouring over Ramanujan's theorems, Hardy said:

> *"They defeated me completely. I had never seen anything in the least like them before. A single look at them is enough to show that they could only be written down by a mathematician of the highest class. They must be true because, if they were not true, no one would have had the imagination to invent them."*

Hardy decided that Ramanujan must be brought to England. As C. P. Snow has written *"Once Hardy was determined, no human agency could have stopped Ramanujan, but they needed a certain amount of help from a* **superhuman** *one."* I shall return to this 'divine intervention' a little later.

Ramanujan arrived in Cambridge in April 1914. The next five years saw an incredible burst of creativity from Ramanujan. In 1918, he was elected a **Fellow of the Royal Society** and a **Fellow of Trinity College**. Soon he became extremely ill and had to return to India early in 1919, and died on January 12, 1920. Hardy wrote thus in his biographical notice for Ramanujan:

> *"In his insight into algebraical formulae, transformation of infinite series, and so forth, that was most amazing. On this side most certainly I have never met his equal, and I can compare him only with Euler or Jacobi."*

Fifteen years later, in his Harvard Lectures, Hardy reassessed what he had written:

> *"I do not think now that this extremely strong language is extravagant. It is possible that great days of formulae are finished, and that Ramanujan ought to have been born 100 years ago; but he was by far the greatest formalist of his time. "*

Eric Neville, a mathematician of distinction and Fellow of Trinity, and who played a crucial role in Ramanujan coming to Cambridge, said:

> *"Srinivasa Ramanujan was a mathematician so great that his name transcends jealousies, the one superlatively great mathematician whom India has produced in the last thousand years".*

2. Ramanujan Institute

Thirty years after Ramanujan's death a mathematical institute named after him was established in Madras. Chandra played a very important role in this. Sometime after the World War ended, one of his former classmates Sir Alagappa Chettiar wrote to him inquiring if it might be useful for him to found a mathematical institute in memory of Ramanujan. Sir Alagappa Chettiar was a very successful entrepreneur and a philanthropist. Chandra enthusiastically supported the idea, and the institute was inaugurated on January 26, 1950,

Figure 5. Temple at Kumbakonam.

the day India became a Republic! T. Vijayaraghavan, one of the most talented among Hardy's former students, was appointed as the first Director. Chandra continued to take keen interest in the development of the Institute. Unfortunately, Vijayaraghavan died in 1955 at a comparatively young age. C.T. Rajagopal, a noted analyst, was appointed as the next Director *on the recommendation of Andre Weil and Chandrasekhar*.

Soon funds started drying up. And when Alagappa Chettiar died in 1957, the fate of the institute hung in balance. Rajagopal wrote to Chandra saying that *'the institute will cease to exist from the first of next month'*. Chandrasekhar came to its rescue. He wrote to the Prime Minister Jawaharlal Nehru, explaining the origin of the institute and the seriousness of its condition. Nehru replied promptly and, as Chandra has written, ¢the reply was refreshing¢. Nehru wrote:

> *"Even if you had not put in your strong recommendation in favour of the Ramanujan Institute of Mathematics, I would not have liked anything to happen which put an end to it. Now that you have also written to me on this subject, I shall keep in touch with this matter and I think I can assure you that the institute will be carried on."*

And it was! In 1967, the Ramanujan Institute was merged with the Department of Mathematics of the University of Madras at the suggestion of the University Grants Commission, and renamed as the Ramanujan Institute for Advanced Study in Mathematics. Although the funding improved, it is still extremely modest compared to what other research institutes in India get from the government. It is a pity that this institute did not attain the same distinction as, for example, the School of Mathematics at the Tata Institute in Bombay.

3. The Goddess of Namakkal

And now for the divine intervention! When the opportunity arose for Ramanujan to go to Cambridge, he was initially hesitant. As Eric Neville described it *"Ramanujan declined,* *reluctant to lose caste, and deferring to the opposition on the part of his parents, whose objections were religious."* In 1914, Neville visited Madras and won Ramanujan's trust completely. Ramanujan expressed his delight to go to Cambridge, and his parents' opposition had been withdrawn. According to the folklore, his mother had a vivid dream in which Ramanujan was surrounded by Europeans, and she heard the **Goddess Namagiri** commanding her to no longer stand between her son and the fulfilment of his life's purpose.

Chandra was always annoyed at the suggestion of some of Ramanujan's contemporaries that he was a deeply religious man. He was very sensitive to this. On more than one occasion Chandra had remarked:

> *"Quite generally, it may be stated that.... there was very little correlation between 'observance' and 'belief'. In particular, I can vouch from my own personal experience that some of the 'observances' that one followed were largely for the purpose of not offending the sensibilities of one's parents, relations, and friends".*

In a BBC radio programme, Chandra mentioned how his mother tied a sacred thread around his wrist on the eve of his voyage to England, and told him not to remove it.

Chandra corresponded with Hardy on the subject of 'Ramanujan and God', and was inclined to accept Hardy's views expressed in a letter dated February 19, 1936:

> *".... And my own view is that, at the bottom and to a first approximation, R. was (intellectually) as sound an infidel as Bertrand Russell or Littlewood....*
> *One thing I am sure. R. was not in the least the 'inspired idiot' that some people seem to have thought him. On the contrary, he was (except for a period when his mental equilibrium was definitely upset by illness) a very shrewd and sensible person: very individual, of course, and with a reasonable allowance of the minor eccentricities of genius, but fundamentally normal and sane."*

The subtle difference between 'observance' and 'belief' came up many times during my conversations with Chandra. When he returned from India in 1968 after delivering the Nehru Memorial Lecture and the Ramanujan Memorial Lecture, he talked to me at length about his visit. Among other things, he mentioned his private meeting with the Prime Minister Mrs. Indira Gandhi, during which she said to him *"You do not know how difficult it is to administer this country. Yesterday, you talked to us about astronomy. Suppose someone had told me that you went to the Ganges this morning and had a bath to prevent the snake Raghu from swallowing the Sun, what can I make of you? That is the kind of persons I have to deal with."*

4. The Lost Notebook

Was Ramanujan born a hundred years too late, as hinted by Hardy? Contemporary mathematicians have firmly rejected this notion. What has led to this reappraisal of Ramanujan?

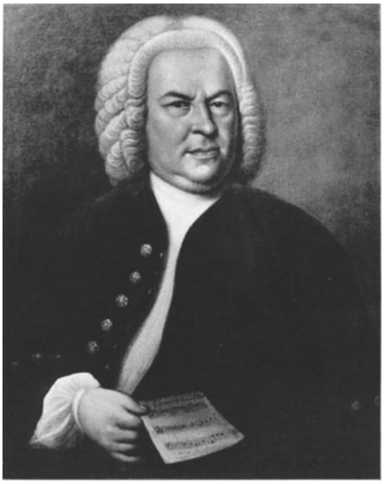


Figure 6. Bruce Berndt has compared Ramanujan with J.S. Bach.

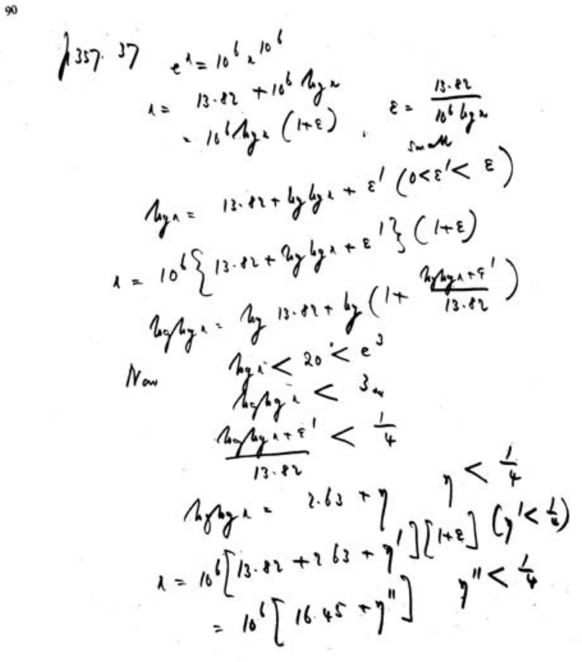
Bruce Berndt, one of the foremost authorities on Ramanujan, has compared him with Johann Sebastian Bach. Bach died in 1750, and was largely unknown. The turnaround came with Felix Mendelssohn's performance of the St Mathew Passion on 11 March, 1829. For Ramanujan, the roughly analogous event was the discovery by George Andrews in 1976 of the Lost Notebook. Since some of you may not be aware of this story, permit me to narrate it.

When G.N. Watson died in 1965, the Royal Society asked J.M. Whittaker to write Watson's Biographical Memoir. For that purpose, Whittaker contacted Mrs. Watson and asked if he could examine the papers that Watson might have left in his study. Whittaker recalled:

> *"... papers covered the floor of a fair sized room to a depth of about a foot, all jumbled together, and were to be incinerated in a few days. One could only make lucky dips and, as Watson never threw away anything, the result might be a sheet of mathematics, but more probably a receipted bill or a draft of his income tax return for 1923. By extraordinary stroke of luck one of my dips brought up the Ramanujan material."*

This 'material', consisting of some 87 loose sheets, was part of a batch of papers the Registrar at the University of Madras had sent to Hardy in 1923. These sheets contained the work Ramanujan had done during the last year of his life, literally in his death bed. Whittaker passed on his lucky dip to Robert Rankin, Watson's successor in Birmingham. Rankin, in turn, deposited them with Trinity College. There it lay in the Trinity Archives till George Andrews discovered and rescued them in 1976.

These loose sheets contain remarkable results pertaining to **"mock theta function"**, the subject of Ramanujan's last letter to Hardy three months before he died. Watson worked on Ramanujan's notes for many years before World War II. Indeed, 'mock theta functions' was the topic of his famous Presidential Address to the Mathematical Society in 1935.



90

p357. 37 $e^x = 10^6 x^{10^6}$

$x = 13.82 + 10^6 \log x$

$= 10^6 \log x \, (1+\varepsilon)$, $\varepsilon = \frac{13.82}{10^6 \log x}$ small

$\log x = 13.82 + \log\log x + \varepsilon' \quad (0<\varepsilon'<\varepsilon)$

$x = 10^6\{13.82 + \log\log x + \varepsilon'\}(1+\varepsilon)$

$\log\log x = \log 13.82 + \log\left(1 + \frac{\log\log x + \varepsilon'}{13.82}\right)$

Now $\log x < 20 < e^3$

$\log\log x < 3$

$\frac{\log\log x + \varepsilon'}{13.82} < \frac{1}{4}$

$\log\log x = 2.63 + \eta \qquad \eta < \frac{1}{4}$

$x = 10^6[13.82 + 2.63 + \eta'][1+\varepsilon] \quad (\eta' < \frac{1}{4})$

$= 10^6[16.45 + \eta''] \qquad \eta'' < \frac{1}{4}$

Figure 7. A random page from the ‘Lost Notebook of Ramanujan’.

Concluding that address, Watson said:

> *“Such a formula gives me a thrill which is indistinguishable from the thrill which I feel when I enter the Sagrestia Nuova of the Capelle Medicee and see before me the austere beauty of the four statues representing “Day¢, “Night”, “Evening”, and “Dawn” which Michelangelo has set over the tombs of Giuliano de’ Medici and Lorenzo de’ Medici.”*

Reminiscing at the Royal Society about the discovery of the Lost Notebook, Chandra remarked,

Andrews told me, at a later time, that when he presented a paper on the ‘Lost Notebook’ at a meeting of the American Mathematical Society, Dr. Olga Taussky-Todd, who was chairing the session, said:

> *“The discovery of the Lost Notebook is as sensational a discovery for the mathematicians as a complete draft of a tenth symphony of Beethoven would have been to the musicians.”*

Let me next quote Professor Richard Askey:

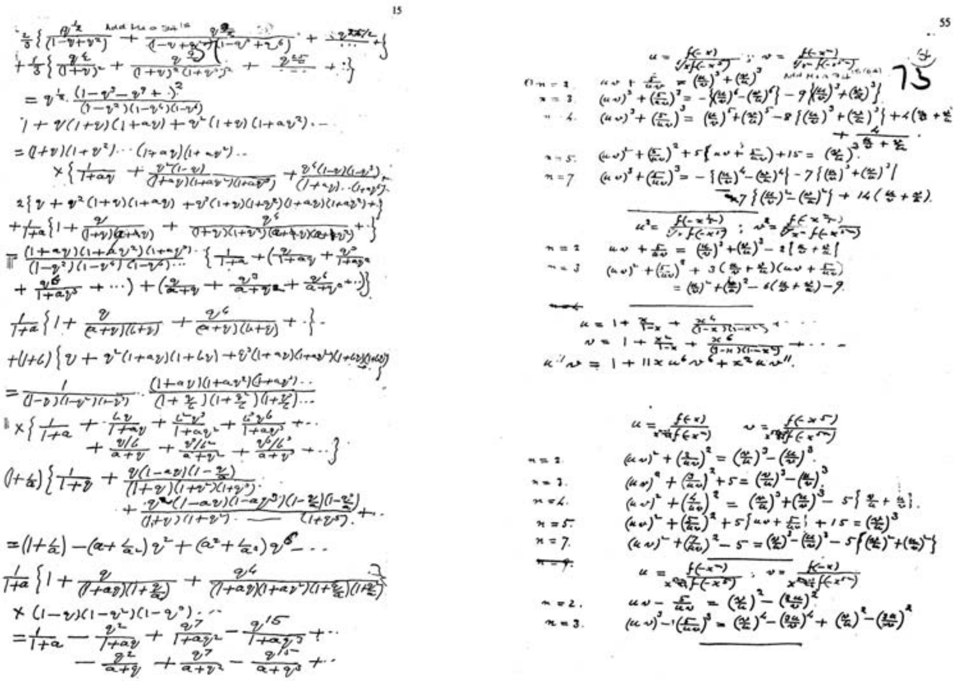


Figure 8. Two of the 87 loose sheets discovered by Whittaker.

> *"Try to imagine the quality of Ramanujan's mind, one which drove him to work unceasingly while deathly ill, and one great enough to grow deeper while his body became weaker. I stand in awe of his accomplishments; understanding is beyond me. We would admire any mathematician whose life's work was half of what Ramanujan found in the last year of his life while he was dying."*

The discovery of the Lost Notebook, and the subsequent deep work done by Andrews, attracted a lot of attention. **The Hindu**, one of the leading newspapers in India, published an interview with Andrews. It followed it up with an interview of Mrs. Ramanujan. She was 80 years old at that time. In that interview, Mrs. Ramanujan lamented the fact that a statue of Ramanujan had never been made, although one had been promised. When Richard Askey heard of this, his reaction was simple: if Ramanujan's widow wanted a bust of her husband she should have it! Askey identified Paul Granlund, a distinguished sculptor at Gustavus College, Saint Peter, Minnesota. The passport photograph of Ramanujan, discovered by Chandra, was all that Granlund had! Initially four busts were made. One of them was for Mrs. Ramanujan. Askey acquired one. Chandra and Lalitha acquired two. One of these two was to be gifted by Chandra and Lalitha to the **Indian Academy of Sciences** on the occasion of its **Golden Jubilee in 1984**; Chandra was a Foundation Fellow of the Academy, founded by his uncle Sir C.V. Raman. Chandra arranged for the Indian Embassy in Washington to send me two busts; the one to be presented to Mrs. Ramanujan, and the

$$\int_0^\infty e^{-3\pi x^2}\frac{\sinh \pi x}{\sinh 3\pi x}dx = \frac{1}{e^{\frac{2}{3}\pi}\sqrt{3}}\sum_{n=0}^{\infty} \frac{e^{-2n(n+1)\pi}}{(1+e^{-\pi})^2(1+e^{-3\pi})^2....(1+e^{-(2n+1)\pi})^2}$$

Figure 9. Ramanujan's magical formulae inspired G.N. Watson to compare them to the great sculptures by Michelangelo at the Medici Chapel in Florence.

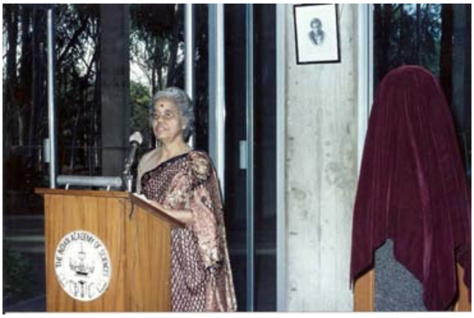


Figure 10. Lalitha Chandrasekhar unveiling the bust of Ramanujan at the Raman Research Institute in Bangalore.

one to be presented by them to the Academy. Chandra and Lalitha came to Bangalore in October 1984 for the Golden Jubilee Meeting. Unfortunately, the meeting had to be postponed due to the assassination of Indira Gandhi. They came again in February of 1985 for the rescheduled Golden Jubilee Meeting.

It was then that they formally gifted the Academy with one of the busts. It is located at the entrance of the lovely library at the Raman Research Institute, which shares the campus with the Indian Academy of Sciences.

In the meantime, I had arranged for S. Ramaseshan, the President of the Academy, to go to Madras and present the other bust to Mrs. Ramanujan.

5. Ramanujan as an inspiration

Most of you knew Chandra well. You must be aware that Ramanujan made a tremendous impact on Chandra's life. Chandra first heard of Ramanujan when he was barely ten years old. His mother told him that a famous Indian mathematician had died the previous day.

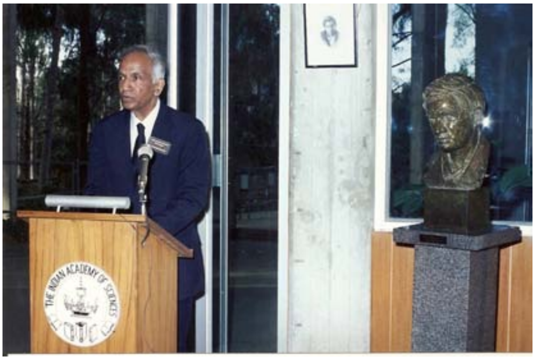


Figure 11. Chandra addressing the gathering after the unveiling.

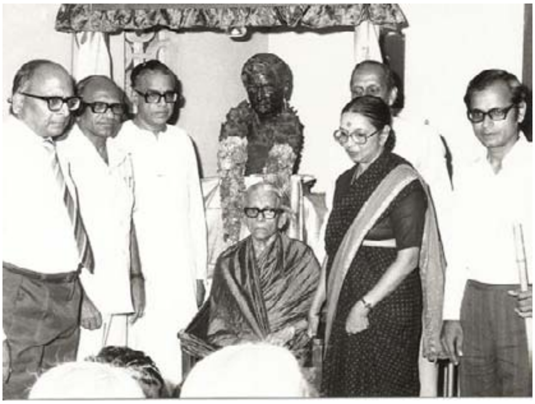


Figure 12. Ramanujan's widow with the bust of her husband.

Sixty seven years later, Chandra reminisced about the influence Ramanujan had on his generation:

> *"I can still recall the gladness I felt at the assurance that one brought up under circumstances similar to my own, could have achieved what I could not grasp...."*
> *"The fact that Ramanujan's early life was spent in a scientifically sterile atmosphere, that his life in India was not without hardship, that under circumstances that appeared to most Indians as nothing short of miraculous, he had gone to Cambridge, supported by eminent mathematicians, and had returned to India with every assurance that he would be considered, in time, as one of the most original mathematicians of the century – these fact were enough – more than enough – for aspiring young Indian students to break their bonds of intellectual confinement and perhaps soar the way that Ramanujan had."*
> *".. The Indian scientific community were exceptionally fortunate in having before them the example of Ramanujan. It is hopeless to try to emulate him.* **But he was there even as the Everest is there.***"*

Figure 13. Mount Everest.

Figure 14. A view of the Kinchinjunga.

We have gathered here to remember another Indian who soared high. Chandra felt that *Ramanujan represents* **so extreme a fluctuation from the norm** *that his being born an Indian must be considered to a large extent accidental.* In his attitude to the pursuit of science, his achievements and his scholarship, Chandra, too, was an extreme fluctuation. What did Chandra think of himself? The following concluding sentences of his lecture at the Golden Jubilee meeting of the Indian Academy of Sciences might hold a clue:

> *"The pursuit of science has often been compared to the scaling of mountains, high and not so high. But who amongst us can hope, even in imagination, to scale the Everest and reach its summit when the sky is blue and the air is still, and in the stillness of the air survey the entire Himalayan range in the dazzling white of the snow stretching to infinity? None of us can hope for a comparable vision of nature and of the universe around us. But there is nothing mean or lowly in standing in the valley below and awaiting the sun to rise over Kinchinjunga."*

When Chandra died, the Indian Academy of Sciences requested me to take the initiative to make a bust. The first thing that came to my mind was the **Ramanujan-Chandrasekhar connection.**

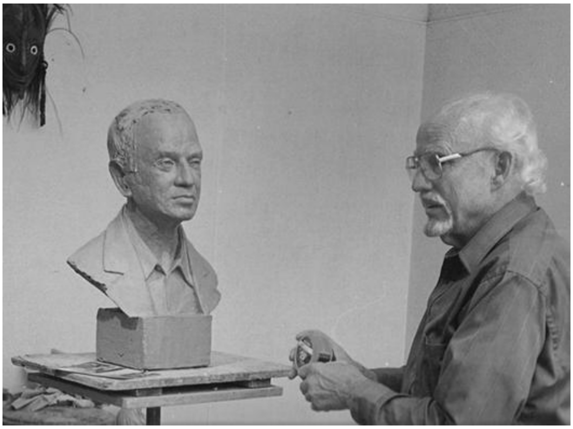


Figure 15. Paul Granlund sculpting the bust of Chandra.

With Richard Askey's help, I approached Paul Granlund and asked if he would undertake to make a bust of Chandra. I was delighted when he not only agreed, but was most enthusiastic. *But unlike Mrs. Ramanujan, Lalitha Chandrasekhar was not at all keen that a bust should be made. I had to work very hard before she finally consented.*

Therefore, it was a matter of great personal satisfaction to me that Lalitha came to Bangalore to unveil the bust.

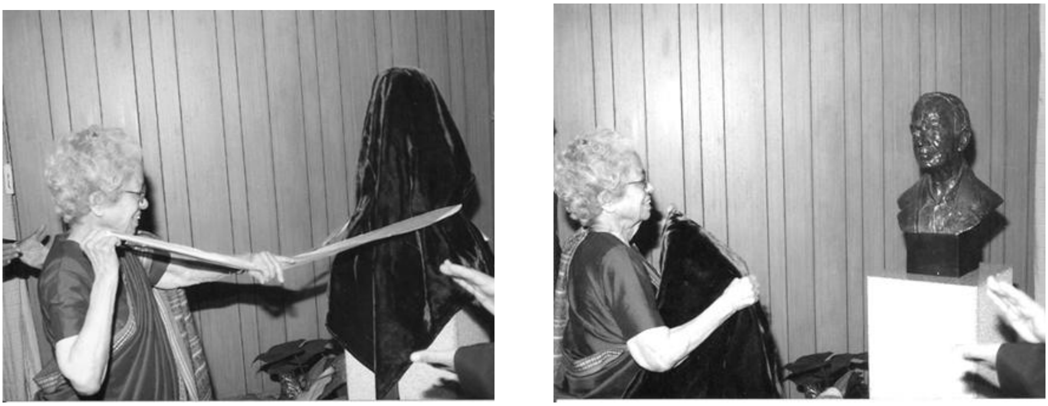


Figure 16. Lalitha Chandrasekhar unveiling the bust of Chandra at the Raman Research Institute.

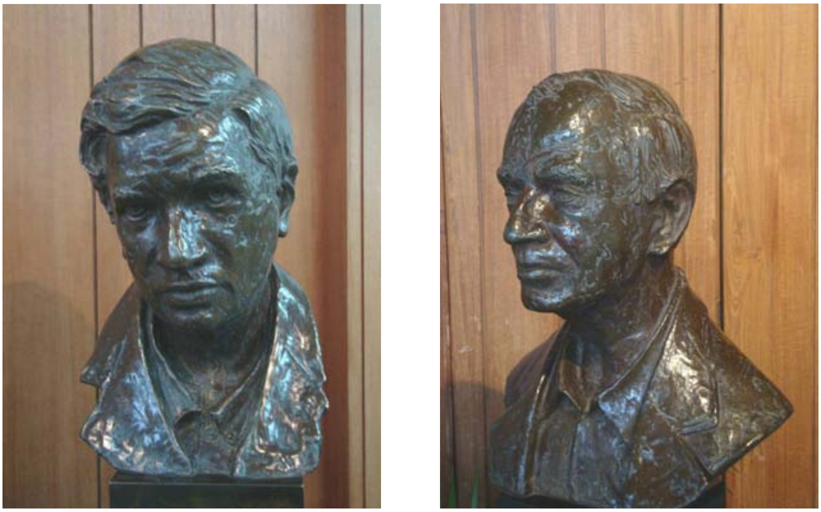


Figure 17. Paul Granlund's busts of Srinivasa Ramanujan (left) and Subrahmanyan Chandrasekhar (right).

Today, Chandra's bust stands adjacent to Ramanujan's bust at the Library of the Raman Institute in Bangalore.

This page intentionally left blank

Chandra's influence on Indian astronomy

Jayant V. Narlikar*

Inter-University Centre for Astronomy and Astrophysics, Pune 411 007, India

Received September 29, 2011

Abstract. The extraordinary achievements of Subrahmanyan Chandrasekhar (Chandra) have guided and inspired many younger astrophysicists. The brief survey seeks to highlight a few specific cases in India where, through his writings, lectures and discussions, Chandra made a lasting impact. It will be argued that although at a general, somewhat superficial level, Chandra is a light beacon to be followed, very few Indian astrophysicists reached a level where they could engage Chandra in a scientific discussion on a topic that interested him.

1. Introduction

A centenary symposium provides an admirable opportunity to review the impact of the concerned scientist on his recognized field of research. Subrahmanyan Chandrasekhar ('Chandra' henceforth) had a very successful career in theoretical astrophysics, topped with many awards and distinctions including the 1983 Physics Nobel Prize, which he shared with William A. Fowler. He died on August 21, 1995 at the age of 85 years. So this centenary symposium gives us the opportunity of evaluating Chandra's impact in his native land of India, some 15 years after his passing away.

To begin such an evaluation I can do no better than reproduce extracts from Chandra's own assessment of his work (Odelberg 1984):

"...After the early preparatory years, my scientific work has followed a certain pattern motivated, principally, by a quest after perspectives. In practise, this quest has consisted in my choosing (after some trials and tribulations) a certain area which appears amenable to cultivation and compatible with my taste, abilities, and temperament. And when after some years of study, I feel that I have accumulated a sufficient body of knowledge and achieved a view of my own, I have the urge to present my point of view, ab initio, in a coherent account with order, form and structure.

There have been seven such periods in my life: stellar structure, including the theory of white dwarfs (1929-1939); stellar dynamics, including the theory of Brownian motion (1938-1943); the theory of radiative transfer, including the theory of stellar atmospheres

[*] email: jvn@iucaa.ernet.in

and the quantum theory of the negative ion of hydrogen and the theory of planetary atmospheres, including the theory of the illumination and the polarization of the sunlit sky (1943-1950); hydrodynamic and hydromagnetic stability, including the theory of the Rayleigh-Bénard convection (1952-1961); the equilibrium and the stability of ellipsoidal figures of equilibrium, partly in collaboration with Norman R. Lebovitz (1961-1968); the general theory of relativity and relativistic astrophysics (1962-1971); and the mathematical theory of black holes (1974-1983). The monographs which resulted from these several periods are:

1. An Introduction to the Study of Stellar Structure (1939, University of Chicago Press; reprinted by Dover Publications, Inc., 1967).
2a. Principles of Stellar Dynamics (1943, University of Chicago Press; reprinted by Dover Publications, Inc., 1960).
2b. Stochastic Problems in Physics and Astronomy, Reviews of Modern Physics, **15**, *1-89 (1943) reprinted in Selected Papers on Noise and Stochastic Processes by Nelson Wax, Dover Publications, Inc., 1954.*
3. Radiative Transfer (1950, Clarendon Press, Oxford; reprinted by Dover Publications, Inc., 1960).
4. Hydrodynamic and Hydromagnetic Stability (1961, Clarendon Press, Oxford; reprinted by Dover Publications, Inc., 1981).
5. Ellipsoidal Figures of Equilibrium (1968, Yale University Press).
6. The Mathematical Theory of Black Holes (1983, Clarendon Press, Oxford)."

This pattern is, I believe, unique in the sense that I know of no other scientist who had systematically compartmentalized his interests so precisely that he never revisited any of the earlier fields of interest. We will encounter in Section 6 an example of this trait. But it follows that given the diversity of interests in the above list, two scientists following the lead given by Chandra may not share a common interest.

Before coming to the topic of my presentation, I should briefly outline the evolution of Indian astronomy and astrophysics over the period of Chandra's work. The theoretical work was mostly done in physics departments, except for general relativity, which was mostly done in mathematics departments. Amongst the former, during the 1930s and 1940s, Delhi and Allahabad universities stood out, with D.S. Kothari in Delhi and M.N. Saha in Allahabad. Amongst the latter, Calcutta University and Banaras Hindu University were prominent in hosting schools of general relativity.

This pre-eminence of universities declined, however, in the post-independence, post-1947 era. The emphasis on research dwindled, and instead found another outlet in the so called autonomous research institutes. These, including the CSIR laboratories and the prestigious Tata Institute of Fundamental Research (TIFR), proliferated in the post-independence era. Their impact on the growth of astronomy and astrophysics (A&A) was predictable and has continued till today. The major Indian research in A&A today comes from the TIFR, the Raman Research Institute (RRI), the Indian Institute of Astrophysics (IIA), the Physical Research Laboratory (PRL), the Institute of Mathematical Sciences (IMSc), the Harish-Chandra Research Institute (HRI) and the Aryabhatta Research Institute of Observational Sciences (ARIES). The heavy tilt away from the universities was to some extent counterbalanced by the creation of the Inter-University Centre for Astronomy and Astrophysics (IUCAA), an institution within the university sector, providing guidance and facilities to university academics working in A&A, besides very successfully conducting its own research.

With this background we may now try to assess Chandra's impact.

2. The 1930s and 1940s

In his biography of Chandra, Kameshwar Wali (1991) has described how Chandra as a student in 1930 attended a get-together in the house of Meghnad Saha in Allahabad, following a major scientific meeting in the city. That the talk at Saha's house was related to issues in A&A sets a contrast to a modern meeting of a similar kind where the talk would be largely on local or national politics. As it turned out a few years later Saha himself took part in national politics, took up positions in important committees, became an M.P. and so on.

Nevertheless Chandra's work on white dwarfs did have impact on a young Indian trained by Saha at Allahabad who left for Cambridge for his Ph.D. in astrophysics. Daulat Singh Kothari got his degree in 1933 and after his return to India he set up the Physics Department at the University of Delhi. Significantly, the new department was named as Department of Physics and Astrophysics. Kothari, following Chandra, was interested in astronomical objects made of dense matter. Normally one considers the ionized state of matter as arising at high enough temperature. Kothari showed that under high pressure also, ionization could be achieved (Kothari 1938).

Sir A.S. Eddington wrote: *"I mentioned that we only gradually came to realize that ionization could be produced by high pressure as well as high temperature. I think the first man to state this explicitly was D.S. Kothari. Stimulated by some work of HN Russall, Kothari has made what I think is an extremely interesting application."* Further commenting on Kothari's work, Arnold Sommerfeld wrote :*"During the times of Galileo and Kepler the planets were at the foucs of astronomical interest but in view of the developments of the last few decades the interest has shifted to stellar physics and spiral nebula. It is noteworthy that the Indian D.S. Kothari has developed an audacious relationship between the old fashioned planets and the now discovered newest heavenly bodies, the white dwarfs"*.[1]

Indeed by keeping temperatures low and pressures high, Kothari could simulate conditions inside a planet through pressure ionization. One important conclusion he arrived at was that a "cold body" cannot have radius exceeding that of planet Jupiter (Kothari 1938; Auluck 1939).

Kothari's tenure at the University of Delhi, unfortunately came to an end when, in 1948, he became scientific advisor to the Defence Minister. Like other scientists of repute such as Saha, Bhabha, Bhatnagar, etc., Kothari too left the academic world in favour of an important government appointment. As after achieving independence, the nation needed brains of demonstrated ability to come forward to create the national infrastructure, many intellectuals were so lost by the academia.

[1] http://www.vigyanprasar.gov.in/scientists/DKothari.htm (Subodh Mahanti: Daulat Singh Kothari, The Architect of Defence Science in India).

Nevertheless, not *all* brains were lost this way. Some continued to work in the universities, and later, in the autonomous research institutes. As mentioned earlier, the emergence of a school of general relativity and gravitation (GRG) came up at the Banaras Hindu University (BHU). Its mentor was my father Vishnu Vasudeva Narlikar (VVN) who had worked in Cambridge under the guidance of A.S. Eddington. VVN was more or less contemporary of Chandra at Cambridge and the two continued to exchange correspondence in the subsequent years.

In the 1940s, Prahlad Chunilal Vaidya, a student of VVN came up with an important exact solution in general relativity. Whereas the classic Schwarzschild solution of 1916 describes the gravitational field of a spherically symmetric, non-radiatinig mass distribution, the Vaidya solution describes the gravitational field of a radiating mass (Vaidya 1943).

At that time VVN sensed that GRG might play a more decisive role in astrophysics than expected hitherto. Nevertheless, to get an expert's assessment he wrote to Chandra to ask if relativists like him should engage themselves in research in relativistic astrophysics. Chandra replied, expressing his view that he did not expect general relativity to be crucial in any part of astrophysics. As a result of this negative assessment VVN as well as Vaidya stayed away from further work in relativistic astrophysics. One can see the logic behind Chandra's assessment through this simple calculation. The dimensionless quantity

$$\alpha = \frac{2GM}{c^2R} \tag{1}$$

describes the gravitational effect of mass M, radius R on the ambient spacetime geometry. For a significant impact of general relativity α needs to approach 1. For white dwarf stars this ratio is $\alpha \sim 10^{-5}$. As believed by most astronomers including Chandra, in the 1940s (and even a decade later) more dense objects with higher α lay in the realm of speculation.

We may look upon α in (1) as made up of two quantities: the mass (M) and the average density (ρ) of the stellar size object. Then the condition for general relativity to be important is that

$$\alpha = \left(\frac{32\pi}{3}\right)^{1/3} \frac{GM^{2/3}\rho^{1/3}}{c^2} \sim 1. \tag{2}$$

As mentioned earlier, for white dwarfs, $\alpha \sim 10^{-5} - 10^{-4} \ll 1$. However, in the 1960s two important discoveries led to the raising of α. Neutron stars with densities approaching 10^{15} times that of water became known. These raised α to within the range $(10^{-2} - 10^{-1})$. The second discovery was that of quasars which were suspected to have masses as high as 10^9 solar masses. This possibility also raised α to a value close to unity at comparatively modest densities, thus making the system of significance to general relativity. We may recall that in his historic controversy with Chandra, Eddington himself had expressed his lack of belief in the existence of what are today known as 'black holes' (Eddington 1935).

These discoveries showed that two decades earlier Chandra also had erred in grossly under-estimating the impact of general relativity on astrophysics. In 1963 a symposium on the new subject 'Relativistic Astrophysics' had been held in Dallas, Texas and it launched several investigations in general relativity of relevance to astrophysics. Chandra did not attend the meeting since, because of racial problems still existing in the South he avoided travelling to the southern states. Ironically, however, he later made significant contributions to this interdisciplinary field.

3. Alumni of Osmania University

In a personal recollection to this author, VVN had mentioned the offer sent through him to Chandra by Professor S. Radhakrishnan, Vice-Chancellor of BHU in the 1940s. The offer was an invitation to head an observatory, and the associated astronomical research, at a handsome salary. The telescope itself was promised by the industrialist Birla family. Chandra declined the offer with thanks. His reservations, as expressed to VVN were largely to do with the lack of an assurance that the autonomy promised by Dr Radhakrishnan would continue after he left the position of the V.C. This was a valid fear, since India has seen several instances of individuals after attaining important positions like that of a V.C., disowning the promises made by their predecessors. Chandra was also afraid of the disruption in his own research brought upon by the administrative and infrastructural issues of running an observatory.

Nevertheless in his capacity as a distinguished visitor, Chandra played a constructive role in India. In my correspondence with Professor Saleh Mohammed Alladin, who retired from the Astronomy Department of Osmania University, Hyderabad, Dr Alladin recalled attending a graduate course on general relativity given by Chandra in 1959 in the Physics Department of the University of Chicago. While stating that Chandrasekhar's lectures were lucid and provided a good background of the subject, Alladin mentions *'Professor Chandrasekhar used to emphasize that mathematical work should not only be correct but should also be elegantly expressed'*. Those of us who have read Chandra's book reassessing Newton's Principia, will have seen echos of this sentiment there too.

Alladin recalls that he was due to be interviewed for a post at Osmania University in 1964 on the same day that Chandra was to visit the University. Between lunch and tea arranged by the University in honour of Chandra, Alladin's interview for the post had been scheduled. However, no interview took place and when the tea party began, the Vice-Chancellor called Alladin and asked him to join the party. He was introduced to the experts invited to interview him. However, at the end of the party the V.C. congratulated him for his selection, without an interview! Alladin feels that Chandra may have been consulted and spoken in his favour.

This episode tells of the flexibility still existing in the university system which enabled a Vice-Chancellor to make an appointment based on recommendation but no interview. Flexibility can be used both ways: to corrupt a system or to invigorate it. Alladin's work on merging galaxies, which he started as a graduate student of D. Nelson Limber — himself a former student of Chandra — later fructified in a lot of very interesting work on dynamics of galaxies at Osmania University (Vardya 1994). In today's highly regulated system the Vice-Chancellor is powerless in terms of what he can do to recruit highly qualified staff. The present decline of Osmania University is a classic example of this situation, and presents a sad contrast to the situation described by Alladin in 1964.

While visiting Hyderabad in 1962, Alladin recalls, Chandra also helped another astronomer from Osmania University, K.D. Abhyankar, in selecting the site for the proposed

Rangapur Observatory. Abhyankar's own work on radiative transfer was considerably influenced by the discussions he had with Chandra during this visit (Abhyankar 1990).

4. Indian graduate students

Chandra had two graduate students from India who later returned home and continued their research there. They were (the late) S.K. Trehan and Bimla Buti, both working in the area of plasma oscillations, two stream instability, etc. Trehan worked in the Applied Mathematics Department of Punjab University, Chandigarh whereas Buti joined the Physical Research Laboratory, Ahmedabad. At the time of writing Trehan is no more.

Buti mentions that at the time of her tenure as graduate student at Chicago University the research publications belonging to the Ph.D. thesis had to be single authored, written by the student, only. So all her thesis publications are under her name only (Buti 1962, 1963a,b).

As the only surviving member from India who worked under Chandra's guidance, I asked Bimla her impressions of Chandra as a human being. She wrote:

"I was impressed and influenced, directly or indirectly, by some of his following habits and actions:

He was a man of very simple habits.

He himself was an extremely disciplined person and expected discipline around him e.g., from all the students in his class. But he was never harsh.

Without fail, he would visit the library and glance through the latest journals.

He was extremely hard working and thorough not only in scientific work but in all respects. However, he would find time for some other activites like gardening, musical concerts, reading classic novels.

While preparing his manuscripts, he was very particular about the English grammar and even punctuations. He would tell his students to follow this pattern.

He had a terrific memory. At a social gathering, he would narrate stories about his pleasant interactions (scientific and social) with other great scientists like Einstein etc. He would keep everyone busy, for hours, with anecdotes.

I personally found him a very friendly and affectionate person."

5. From white dwarfs to neutron stars and the Sun

Coming now to a later era, 1960-70, my former colleague at the Tata Institute of Fundamental Research (TIFR), Kumar Chitre has provided useful inputs. His own work was influenced by two of Chandra's interests : stellar structure and hydrodynamic and hydromagnetic stability.

In the 1960s, the problem of determining the limiting mass of a neutron star, like the Chandrasekhar mass limit for white dwarfs posed a challenge to theoreticians. The answer lies in finding the correct equation of state for matter with density approaching $\sim 10^{15}\mathrm{g\,cm^{-3}}$, the density expected to be present within the core of a neutron star. S.M. Chitre at TIFR and V. Canuto at CCNY went through the exercise with the solution that the limiting mass is around $2M_\odot$ (Canuto & Chitre 1973). A somewhat different line leading to the same answer was followed by Pandharipande in the University of Illinois at Urbana.

Chitre was also guided by Chandra's work on hydrodynamic and hydromagnetic stability to look at the stability of solar models. The spectrum of eigenfrequencies derived numerically by perturbing the solar model could then be compared with the accurately observed accoustic mode oscillation frequencies. This technique helps in arriving at an accurate model of the Sun with a confident prediction of solar neutrino flux emerging from it. The value so obtained agrees very well with that obtained from the standard model of the Sun. So one was led to the conclusion that the observed deficit of solar neutrino flux had to come from neutrino physics, e.g., from neutrino oscillations (Antia & Chitre 1997, 1998).

Both these works demonstrate practical applications of ideas Chandra had propagated decades ago.

6. Antonov instability

Chandra's name and achievement have become textbook material. Even secondary school children in India will have encountered his life story sometime during their school studies. The younger generation therefore views him more in awe as a scientist who had scaled high peaks of excellence than really understanding what exactly he did achieve. His academic interaction with a young working scientist has thus been somewhat rare. The episode described by T. Padmanabhan is therefore of some interest. I quote him almost verbatim in what follows.

"My main academic encounter with Professor S. Chandrasekhar was related to the question of Antonov instability.

During my postdoctoral years (1986-87) at Institute of Astronomy, Cambridge, Donald Lynden-Bell got me interested in the study of statistical mechanics of gravitating systems and, in particular, in Antonov instability, first described by Antonov in 1962. His original derivation was quite complicated (Antonov 1962) and I was trying to understand its physical origin from a simpler point of view from the structure of the equations describing an isothermal sphere.

Chandrasekhar, in his work in 1939, has discussed how the equations of stellar structure (including the ones describing the isothermal sphere) can be reduced to a first order differential equation by using two variables u, v. The solutions to isothermal sphere equation in these variables is represented as a spiral in the uv plane. I realized the key dimensionless parameter $q = [RE/GM^2]$ - where R is the radius, M is the mass and E is the energy - which describes the Antonov instability can also be expressed in terms of these variables. In fact, I found that q = constant curves are straight lines in the uv plane! Any solution with a fixed value of q is given by the intersection of two curves (one spiral and one straight line) in the uv plane. The condition that these lines have to intersect immediately leads to the condition for Antonov instability. This is shown in figure 4.2 page 316 of my review. (See Padmanabhan 1990).

I was surprised that Chandrasekhar, in his ref., did not bother to plot lines of constant $[RE/GM^2]$ *in a corresponding figure. If he had done that he would have discovered Antonov instability nearly 25 years before Antonov!*

I was happy to get this simpler derivation but wanted to know whether Chandra had some thoughts on this matter. I wrote him a letter around 1990 when I was working on my review and asked him about it. He sent me a polite reply saying that as the matter refers to something which he did nearly 50 years back he cannot quite recollect what his thoughts were when he performed this analysis. Later on, when I met him at IUCAA, I had attempted to discuss this problem with him. He told me that he likes to work in an area for sometime and then move on but almost never re-visits that topic. In fact, he was not very keen to know the details of my derivation. It was an interesting approach to physics in the sense that he almost showed certain amount of reluctance to discuss/revisit the problem he had pioneered, once he had moved on."

This is the aspect of Chandra that I had hinted at earlier in this presentation: that once he made a transition from one major topic to another and written a monograph on the work just completed, the topic became a 'closed book' for him.

7. Concluding remarks

In a moving account sent as a letter to me, Ramnath Cowsik has narrated his various encounters with Chandra. It becomes interesting in the present context because it throws light on Chandra's attitude to science and other intellectual pursuits as expressed before Indian students. I mention a few instances next.

At a radio interview in Mumbai Cowsik asked him a question, that he said, had been prompted by many students : "How does a student prepare for a career in Physics? Should he first have a serious study of theoretical physics or should he learn experimental techniques?.." Chandra answered this question in his inimitable way: "Different students depending upon their temperament and preparation approach physics in their own unique ways. Each of these is as valid as any other. But what is important is they dedicate themselves to academic life. It does not matter through which gate that one enters a garden. Once you are in, you may wander enjoying a bloom here or a bough there".

The dedication ceremony of my centre IUCAA (Inter-University Centre for Astronomy and Astrophysics) was highlighted by Chandra's talk entitled *The Series Paintings of Claude Monet and the Landscape of General Relativity*. That was in 1992 December. Cowsik recalls a seminar Chandra gave at the Indian Institute of Astrophysics, entitled "On the oscillations of a star as a problem in the scattering of gravitational waves". In both lectures he showed how to map one problem into another and thereby from the second obtain an elegant solution to the first problem.

Finally, I end with an account of my first meeting with Chandra. In 1960 when I was in the final stages of completing the Mathematical Tripos Examination at Cambridge, my father wrote to Chandra to explore the possibility of my becoming his research student. Chandra replied that he was in the process of changing his field of research and in this transitional phase he did not intend taking a new research scholar. So I missed the chance of being Chandra's pupil. But I continued at Cambridge as student of Fred Hoyle.

In 1962 I attended the International Conference on General Relativity and Gravitation held near Warsaw. On the first morning I took a stroll before breakfast in the vast well laid gardens of the Polish country house where I was staying. There I met a senior gentleman in a dark suit, evidently from the Indian subcontinent. He smiled and introduced himself. I reciprocated, although my face may have shown some surprise as to what a mathematical astrophysicist was doing at a relativity conference. For Chandra vounteered the information: "I have decided to work in the field of general relativity. What better place than an international conference, to get to know the areas where intellectual challenges exist? So I have come as a student to learn."

At early fifties, as I estimated Chandra's age to be, most scientists begin to taper down their research. Here was someone entering a new field with the enthusiasm of a twenty-odd year old. It is this attitude that I, as a twenty-four year old, felt the need to copy.

References

Abhyankar K.D., 1990, Bull. Astr. Soc. India, 18, 109

Antia H.M., Chitre S.M., 1997, MNRAS, 289, L1

Antia H.M., Chitre S.M., 1998, MNRAS, 339, 239

Antonov V.A., 1962, Vestn. Leningrad Gos. Univ., 7, 135 [English translation in Dynamics of Globular Clusters, IAU Symp. 113, eds. J. Goodman and P. Hut (Reidel, Dordrecht), 1985.]

Auluck F.C., 1939, MNRAS, 99, 239

Buti B., 1962, Phys. Fluids, 5, 1

Buti B., 1963a, Phys. Fluids, 6, 89

Buti B., 1963b, Phys. Fluids, 6, 100

Canuto V., Chitre S.M., 1973, Phys. Rev. Lett., 30, 999

Chandrasekhar S., 1939, An Introduction to the Study of Stellar Structure, University of Chicago Press, Chicago

Chandrasekhar S., 1943a, Principles of Stellar Dynamics, University of Chicago Press, Chicago

Chandrasekhar S., 1943b, Reviews of Modern Physics, 15, 1

Chandrasekhar S., 1950, Radiative Transfer, Clarendon Press, Oxford

Chandrasekhar S., 1961, Hydrodynamic and Hydromagnetic Stability, Clarendon Press, Oxford

Chandrasekhar S., 1968, Ellipsoidal Figures of Equilibrium, Yale University Press, Yale

Chandrasekhar S., 1983, The Mathematical Theory of Black Holes, Clarendon Press, Oxford

Eddington A.S., 1935, The Observatory, 58, 259

Kothari D.S., 1938, Proc. Roy. Soc., A., 165, 486

Odelberg W., ed, Les Prix Nobel, The Nobel Prizes 1983, The Nobel Foundation, Stockholm

Padmanabhan T., 1990, Statistical Mechanics of Gravitating Systems, Phys. Rep., 188, 285

Vaidya P.C., 1943, Current Science, 12, 183

Vardya M.S., 1994, Def. Sci. Journal, 44, 207

Wali K.C., 1991, Chandra: a biography of S. Chandrasekhar, University of Chicago Press

This page intentionally left blank

Chandrasekhar and the history of astronomy

Virginia Trimble[*]

Department of Physics and Astronomy, University of California, Irvine CA 92697, USA, and
Las Cumbres Observatory Global Telescope Network, Goleta, California, USA

Received August 12, 2011

Abstract. Chandrasekhar's own books, papers, and oral history interviews make clear that he was generally more interested in the present and future of astrophysics than in its past. Nevertheless, late in his life and after his death, historians of science have somewhat entangled him in two supposedly controversial issues, one concerning precursors of his mass limit for degenerate stars and the other his relationship with Eddington. Neither story is an entirely happy one.

1. The fate of famous scientists

Biographers write biographies and historians write books and papers, and no one, living or dead, has much defense against them. Among scientists I've known, Richard Feynman, Fred Hoyle, and Carl Sagan have each been the subject of at least three. Thus no one should be surprised that there are biographies and encyclopedia articles about Chandra (Wali 1991, 2008) and Eddington (Douglas 1956; Stanley 2007, 2008), and indeed even a biography of Eddington by Chandrasekhar (1983). Not that you had any doubts before, but you cannot come away from any of these without realizing that each made both extraordinarily many and extraordinarily important contributions to 20th century astrophysics. But be thou chaste as ice, as pure as snow, thou shalt not escape calumny (Hamlet, Act III, Sc. 1, to save you looking it up).

2. Degenerate stars, or, who discovered the Chandrasekhar limit?

I first encountered this issue more than 30 years ago (Trimble 1979) when I reviewed a semi-popular book by I.S. Shklovskii (1978) called 'Stars: Their Birth, Life, and Death' and claimed to have learned from it that the Chandrasekhar limit was really discovered by Yakov (variously Jacov) Frenkel in 1928. Very soon after that issue of Sky & Telescope hit the newsstands, there arrived a manilla envelope from the University of Chicago, in which Chandra had enclosed copies of his 1931 papers and a hand-written note pointing out that

[*] email: vtrimble@uci.edu

these were the first papers to use explicitly an equation of state with pressure proportional to (density)$^{4/3}$, although this is implicit in the Frenkel paper. Indeed so he had said on page 409 of his stellar structure book (Chandrasekhar 1939).

First I apologized and then I sat down to read the Frenkel paper with a German born friend who was a professional interpreter. She read the short words and I the long ones. She quickly concluded that German was not Frenkel's first language and I that he had not discovered the Chandresekhar limit. Rather, he was somehow addressing degenerate baryons, though the neutron had yet to be found (Chadwick 1932) and protons never get a chance to be degenerate. And there, I supposed, the issue would rest.

It did not and has been raised again by Nauenberg (2008) and again by Nauenberg (2011) and Blackman (2011) in connection with the Chandrasekhar centenary, with a rebuttal from Wali (2011). There is no disagreement about what the several relevant papers say, but only about what Chandra knew and when he knew it, and how credit should be divided, and eponyms awarded. All agree that Fowler (1926) was first to derive an equation of state, $P = K_1(\rho)^{5/3}$ for completely degenerate matter, neglecting any possible effects of special or general relativity. With this EoS, you can build configurations of any total mass, and Chandra had read the Fowler paper before leaving India. Next, Edmund C. Stoner (1929) and Wilhelm Anderson (1929) looked for deviations from Fermi-Dirac degeneracy implied by large densities and so by occupation of momentum states close to $E = p^2/2m = m_e c^2$. Stoner reported an upper limit to densities and Anderson a modification of the EoS in the direction toward $P = K_2(\rho)^{4/3}$ which we now associate with completely relativistic degeneracy. His tables and formulae imply an upper limit to degenerate masses, but explicit calculation of this limit was left to Stoner (1930). Chandra had not had access to those papers before leaving India.

Both approximated white dwarf stars by uniform density spheres. This is not actually a foolish thing to do, and still has pedagogical value (Hansen, Kawaler & Trimble 2004, p. 16). Israel (1987) affirms that $P = K_2(\rho)^{4/3}$ is and should be called the Stoner-Anderson equation of state. Chandrasekhar (1931a,b) famously, perhaps even notoriously did his critical calculation on board ship in 1930, and Wali (2011) has concluded that he was not aware of either Stoner's or Anderson's work at the time. His work was therefore independent, but, more to the point, he adopted Eddington's (1926) polytropes for his models which could, therefore, be in hydrostatic equilibrium, which constant density stars cannot, and real ones must be. A very similar limiting mass was derived by Landau (1932, but paper submitted February 1931), but he is not mentioned by either Nauenberg (2011) or Blackman (2011).

Did Chandra give adequate credit to his predecessors? Simply reading his 1939 book one would think so, though his student, Guido Munch, said very much later that the Stellar Atmospheres book credits him only with drawing figures and not for the couple of chapters he wrote. In any case, Chandrasekhar carried on work largely on stellar structure, especially degenerate stars until 1935. And then significant portions of the roof fell in.

3. Chandrasekhar and Eddington

In the interim, Eddington had apparently been concentrating on his 'fundamental theory' (Israel 1987), and so, although he was in regular touch with Chandra and his ongoing work, he had perhaps not immediately thought what the consequences would be (Wali 1982). And when he did, he put in a paper on 'Relativistic Degeneracy' to be read at the January, 1935 meeting of the Royal Astronomical Society immediately after Chandra's presentation of extensive numerical analysis indicating that the fate of massive stars must be something other than gradually cooling white dwarfs. The next part of the story can be read by anybody with access to old journals, because Observatory, then as now, published more or less verbatim accounts of the RAS meetings, and January 1935 appears in Volume 58, page 37ff. Eddington announced that there is no such thing as relativistic degeneracy and that Dr. Chandrasekhar had rubbed in his result to a reductio ad absurdum, leaving among most hearers the impression that Chandrasekhar had made a simple mistake in his calculations.

Chandra was given no opportunity to respond at that meeting, nor was he later in the year when Eddington spoke at the Paris IAU on the non-existence of relativistic degeneracy. Eddington's toes remained dug in for a number of years thereafter, despite support for the $(\rho)^{4/3}$ equation of state from physicists (Wali 1982) and some observational confirmation (Nauenberg 2011). The questions we might reasonably ask are:

1. Was Eddington's behaviour outside of tolerances? Not, one must conclude, by Eddington's standards. This was, after all, the person who said he did not think it necessary to read a paper by Professor Milne, because it would be absurd for him (Eddington) to pretend that he (Milne) has the remotest chance of being right (Wali 1982). On other occasions, he said things equally harsh about James Jeans and others (Stanley 2008), and his remark at another meeting about generation of subatomic energy in stars, "If the honorable gentleman does not think the center of the sun is hot enough, then let him go and find a hotter place", has joined the body of folklore we share with students.

2. Why did the Eddington toes remain so firmly buried? Stanley (2008) has pretty firmly ruled out the unpleasant suggestion that racial prejudice entered into it. Israel (1987) worked carefully through Eddington's scientific output from the time of his 1923 paper 'The Mathematical Theory of Relativity', and concluded that Eddington had gradually become so wedded to his own definition of the stress tensor and its imbedding in his 'fundamental theory' that he simply couldn't conceive of the Stoner-Anderson equation of state describing anything in the real world.

3. What were the consequences, especially for Chandra? Remember the title of his 1983 book, 'Eddington: The Most Distinguished Astrophysicist of His Time'. If you are thinking that nobody writes a book just to say nasty things about someone else, you need to think again. But, more to the point, Chandra need not have written an Eddington biography at all. Nor would he have needed to have given the obituary and centenary talks cited by Wali (1982). Was Chandrasekhar's acceptance of a position in the United States partly a reaction to Eddington's attitude and the expectation that it might interfere with a successful career in the UK? Perhaps. But more firmly (Wali 2008), the Eddington controversy entered into Chandra's decision to write up his work on stellar structure in 1939 and move on to stellar dynamics. This set the pattern for much of the rest of his career, during which, as virtually everybody has noticed, he worked intensely on a topic until he felt he had learned what he had set out to learn, wrote it up, and moved on to something else, rarely looking back. Thus, just possibly, Eddington's behaviour helped nudge his younger colleague in the most productive possible direction (Dyson 2010).

4. In their own words (and deeds)

What Eddington said to Cecilia H. Payne when she expressed a strong desire to become an astronomer was, “I see no insuperable obstacle”, but he advised her to go to the United States, which she did, finding the obstacles from 1925 to 1956 (when she was finally appointed to a professorship at Harvard) high, but indeed not insuperable.

What Chandra said to me when I was passing through Chicago in early May 1968, en route to a brief postdoc at Cambridge University was, “You must give a colloquium”, which I did, to a nearly-filled room, despite the detail that it was Saturday and that the topic was ‘Motions and Structure of the Filamentary Envelope of the Crab Nebula’. Some of the interesting things he said to other colleagues are included in five reminiscences in the December, 2010 issue of Physics Today.

What of whimsy? Eddington had his cycling number, x , the largest number such that he had cycled at least that many miles on at least that many days. It had reached 75 when he wrote to Chandra in 1938, and the concept will be recognized as the ancestor of the Hirsch index, h, having to do with citations of papers. My own piece of Chandra whimsy was his response to my question about why he had never been on one of the decadal survey committees used, in the US, to set equipment and other astrophysical priorities, starting in 1962. He responded immediately, “No one ever asked me”, and, after a moment’s cogitation, continued with a verse or two of an English folk rhyme ending with a pompous young man saying to a farmer’s daughter, “Then I cannot wed you my fair young maid. Nobody asked you sir, she said.” Not surprisingly, he gave “said” the Yorkshire pronunciation required to sustain the rhyme.

A widely reproduced Eddington quote came from his 1935 RAS talk in opposition to relativistic degeneracy: The star has to go on radiating and radiating and contracting and contracting until, I suppose, it gets down to a few km radius, when gravity becomes strong enough to hold in the radiation, and the star can at last find peace. This is surely as good a prediction of black holes or at least horizons, as is to be found in the 18th century writings of John Michell and Pierre Simon de Laplace.

Somewhat less well known is the last line of Chandrasekhar (1932) submitted during his brief stay in Copenhagen: “Given a container containing electrons and atomic nuclei (total charge zero), what happens if we go on compressing the material indefinitely?” Equally clearly, this is a prologue to neutron stars. James Chadwick announced neutrons in February, 1932 and the paper was submitted on September 28th, but in fact neutron stars had to wait another year for Baade & Zwicky (1933), about whom there are also many stories, but they must wait for another book.

References

Anderson W., 1929, Zs. f. Phys., 54, 433
Anderson W., 1930, Tartu Pupi, 29
Baade W., Zwicky F., 1933, Proc. USNAS, 20, 254
Blackman B.G., 2011, Physics Today, July, p. 8
Chadwick J., 1932, Nature, 129, 312
Chandrasekhar S., 1931a, ApJ, 74, 81
Chandrasekhar S., 1931b, MNRAS, 91, 446

Chandrasekhar S., 1932, Zs. f. Ap 5, 34
Chandrasekhar S., 1939, An Introduction to the Study of Stellar Structure, Univ. of Chicago Press
Chandrasekhar S., 1950, Radiative Transfer, Oxford Univ. Press
Chandrasekhar S., 1983, Eddington: The Most Distinguished Astrophysicist of His Time, Cambridge Univ. Press
Douglas A.V., 1956, The Life of Arthur Stanley Eddington, London Nelson
Dyson F.J., 2010, Physics Today, December, p. 44
Eddington A.S., 1926, The Internal Constitution of the Stars, Chapter 4
Fowler R.H., 1926, MNRAS 83, 114
Frenkel J., 1928, Zs. f. Phys., 47, 819
Hansen C.J., Kawaler S.D., Trimble V., 2004, Stellar Interiors: Physical Principles, Structure and Evolution, 2nd Edition, Springer-Verlag: New York, p. 16
Israel W., 1987, in S.W. Hawking & W. Israel, Eds. 300 Years of Gravitation, Cambridge Univ. Press, p. 199
Koertege N., 2008, Editor, New Dictionary of Scientific Biography, Charles Scribner & Sons
Landau L., 1932, Phys. Z. Sowjetunion, 1, 285
Nauenberg M., 2008, J. Hist. Astron., 39, 297
Nauenberg M., 2011, Physics Today, July, p. 8
Shklovskii I.S., 1978, Stars: Their Birth, Life, and Death, Freeman, San Francisco
Stanley M., 2007, Practical Mystic, Univ. of Chicago Press
Stanley M., 2008, in Koertege 2008, Vol, 2, p. 338
Stoner E.C., 1929, Phil., Mag. 7, 63
Stoner E.C., 1930, Phil., Mag. 9, 944
Trimble V., 1979, Sky & Telescope, 57, 279
Wali K.C., 1982, Physics Today, October, p. 33
Wali K.C., 1991, Chandra: a biography of S. Chandrasekhar, Univ. of Chicago Press
Wali K.C., 2008, in Koertege 2008, Vol 2, p. 87
Wali K.C., 2011, Physics Today, July, p. 9

This page intentionally left blank

Compact stars and the evolution of binary systems

E. P. J. van den Heuvel

Astronomical Institute, “Anton Pannekoek”, University of Amsterdam, The Netherlands

Received January 31, 2011; accepted February 27, 2011

Abstract. The Chandrasekhar limit is of key importance for the evolution of white dwarfs in binary systems and for the formation of neutron stars and black holes in binaries. Mass transfer can drive a white dwarf in a binary over the Chandrasekhar limit, which may lead to a Type Ia supernova (in case of a CO white dwarf) or an Accretion-Induced Collapse (AIC, in the case of an O-Ne-Mg white dwarf; and possibly also in some CO white dwarfs) which produces a neutron star. The *direct* formation of neutron stars or black holes out of degenerate stellar cores that exceed the Chandrasekhar limit, occurs in binaries with components that started out with masses ≥ 8 M$_\odot$.

This paper first discusses possible models for Type Ia supernovae, and then focusses on the formation of neutron stars in binary systems, by direct core collapse and by the AIC of O-Ne-Mg white dwarfs in binaries. Observational evidence is reviewed for the existence of two different direct neutron-star formation mechanisms in binaries: (i) by electron-capture collapse of the degenerate O-Ne-Mg core in stars with initial masses in the range of 8 to about 12 M$_\odot$, and (ii) by iron-core collapse in stars with inital masses above this range. Observations of neutron stars in binaries are consistent with a picture in which neutron stars produced by e-capture collapse have relatively low masses, $\sim$1.25 M$_\odot$, and received hardly any velocity kick at birth, whereas neutron stars produced by iron-core collapses are more massive and received large velocity kicks at birth. Many of the globular cluster neutron stars and also some of the neutron stars in low-mass binaries in the Galactic disk are likely to have been produced by AIC of O-Ne-Mg white dwarfs in binaries. AIC is expected to produce normal strongly magnetized neutron stars, which in binaries can evolve into millisecond pulsars through the usual recycling scenario.

Keywords : stars: binaries: general – stars: evolution – stars: white dwarfs – stars: neutron – stars: pulsars: general – stars: supernovae: general

1. Introduction

The Chandrasekhar limit is of key importance for the evolution of white dwarfs in binary systems which receive mass from a companion star. This is now the favoured model for the origin of Type Ia supernovae, which are known to be excellent ‘standard candles’ (e.g.

Riess *et al.* 1998; Phillips *et al.* 1999). Their use in cosmology has led to the discovery of dark energy (Schmidt *et al.* 1998; Perlmutter *et al.* 1999).

It was realized long ago (Hoyle & Fowler 1960) that Carbon-Oxygen (CO) white dwarfs, which are the final products of stars less massive than about 8 $M_{\odot}$, still contain a large amount of nuclear fuel, and that the ignition of carbon under degenerate conditions in the interior of a white dwarf will lead to runaway nuclear fusion, converting most of the star into ^{56}Ni, thereby causing the star to blow up in a gigantic explosion with energy equivalent to that of a supernova. Since hydrogen and helium are absent in the spectra of Type Ia supernovae, and these spectra are dominated by products expected from explosive carbon burning, these supernovae fit very well with the thermonucelar explosion model of a carbon-oxygen (CO) white dwarf (Hoyle & Fowler 1960; Nomoto 1982a).

In order to trigger carbon ignition the mass of the white dwarf has to grow to the Chandrasekhar limit. The only realistically conceivable way to achieve this, is by the transfer of matter from a companion star in a binary system. This led Whelan & Iben (1973) to suggest that Type Ia supernovae originate from binaries in which a red giant star is transferring mass to a CO white dwarf. This is the so-called ‘Single-Degenerate’ (SD) model for Type Ia supernovae. This model was worked out for example by Nomoto (1982a,b).

Subsequently it was realized by Webbink (1984) and Iben & Tutukov (1984) that wide binaries of intermediate mass (components between $\sim$2 and 8 $M_{\odot}$) may after several stages of mass exchange and common-envelope evolution leave very close binary systems consisting of two CO white dwarfs. When their orbital periods are shorter than about one day these systems will within a Hubble time merge, due to loss of orbital angular momentum by emission of gravitational waves. If the merger product has a mass larger than the Chandrasekhar limit, it may explode as a Type Ia supernova. This is the Double Degenerate (DD) model for Type Ia supernovae (however, see Saio & Nomoto 1985, 2004, for the alternative view that a DD merger produces a neutron star). I will briefly discuss the merits of the SD and DD models in Section 2.

In binaries in which the mass-receiving white dwarf is of the O-Ne-Mg type (these can under certain conditions be produced by stars with initial masses between $\sim$8 and 12 $M_{\odot}$; see Miyaji *et al.* 1980; Podsiadlowski *et al.* 2004), mass transfer to the star until it reaches the Chandrasekhar limit will lead to the collapse of the star due to the capture of degenerate electrons by nuclei of Ne and Mg (e.g. Miyaji *et al.* 1980; Nomoto 1984; Canal, Isern & Labay 1990; Pylyser & Savonije 1988). The outcome of this Accretion-Induced Collapse (AIC) is expected to be a neutron star. (Under very special conditions a CO white dwarf might in some cases collapse to a neutron star, e.g. Canal, Isern & Labay 1990.) This is one way to produce neutron stars in binary systems, which may later evolve into X-ray binaries and binary radio pulsars.

However, the majority of the neutron stars in X-ray binaries and binary radio pulsars are expected to have been produced by the direct core collapse of stars that started their lives with masses $\geq$ 8 $M_{\odot}$. Here there is still a difference between stars which started out with masses between $\sim$8 and 12 $M_{\odot}$, in which a degenerate O-Ne-Mg core forms, which collapses as a result of electron capture (Miyaji *et al.* 1980; Podsiadlowski *et al.* 2004), and stars more massive than about 12 $M_{\odot}$, in which the core passes through all stages of nuclear fusion until a degenerate iron core forms which collapses to a neutron star.

In recent years it has become clear from the study of Be-type X-ray Binaries (Pfahl *et al.* 2002) and of binary radio pulsars (van den Heuvel 2004) that these two types of core collapses most probably produce neutron stars with different properties, i.e. with different masses and kick velocities (Podsiadlowski *et al.* 2004, 2005; van den Heuvel 2004; Dewi, Podsiadlowski & Pols 2005; Schwab, Podsiadlowski & Rappaport 2010).

In Section 3, I discuss in more detail these relatively new findings that appear to confirm that there are two different mechanisms by which neutron stars can form, as was originally suggested by Miyaji *et al.* (1980). I also briefly discuss there the role that AIC may play in the formation of neutron stars in globular clusters and in Low-Mass X-ray Binaries and binary pulsars observed in the Galactic disk.

2. Type Ia supernova scenarios

2.1 The Single Degenerate model and its problems

In the SD model the white dwarf is growing in mass due to the accretion of matter from its non-degenerate companion star. The problem here is that for a wide range of mass-transfer rates, the hydrogen accumulated on the surface of the white dwarf tends to ignite explosively, once the mass of the accreted layer exceeds a threshold value. Such thermonuclear explosions of the accreted hydrogen layer are observed as various types of nova outbursts, and it is quite possible that in many of these explosions much, if not all, of the accreted matter is ejected, such that little or no net-growth of the white dwarf may take place. The critical mass ΔM_c at which nuclear burning ignites decreases with increasing white dwarf mass and increasing accretion rate (e.g. Nomoto 1982a; Prialnik & Kovetz 1995; Townsley & Bildsten 2005).

Figure 1 (from Townsley & Bildsten 2005) depicts the various nuclear burning regimes, of hydrogen-rich matter (70 per cent hydrogen), on the surface of a white dwarf, as a function of accretion rate and white dwarf mass. For a small range of accretion rates, $(1 - 4) \times 10^{-7}$ $M_\odot$/yr, indicated by the hatched band in the figure, the hydrogen burns steadily on the surface of the white dwarf, and for these accretion rates the white dwarf will be able to steadily grow in mass. Below this range of accretion rates, the burning takes place in flashes. The curves in the figure depict the critical masses ΔM_c of the accreted hydrogen layer at the moment at which burning is ignited. The accretion rate onto the WD determines the strength of the outburst. Higher accretion rates lead to less violent outbursts.

For very high accretion rates, above twice the maximum accretion rate for steady nuclear burning, the burning is still steady, but due to the super-Eddington energy generation by this burning, a strong stellar wind develops in which the excess accreted matter is blown away. For these accretion rates, the WD still grows, but less efficiently than in the accretion range for steady burning without a wind, since part of the transferred matter will be lost, and cannot contribute to the growth of the WD.

Binaries in which steady nuclear burning on the surface of the WD takes place were identified with the bright Super Soft X-ray Sources (SSS) discovered by the ROSAT satellite (van den Heuvel *et al.* 1992; see the reviews by Rappaport & Di Stefano 1996 and Kahabka & van den Heuvel 1997, 2006). These typically emit of order 10^{38} ergs/s in the form of very soft X-rays peaking in the energy range 20-100 eV. (It should be remembered

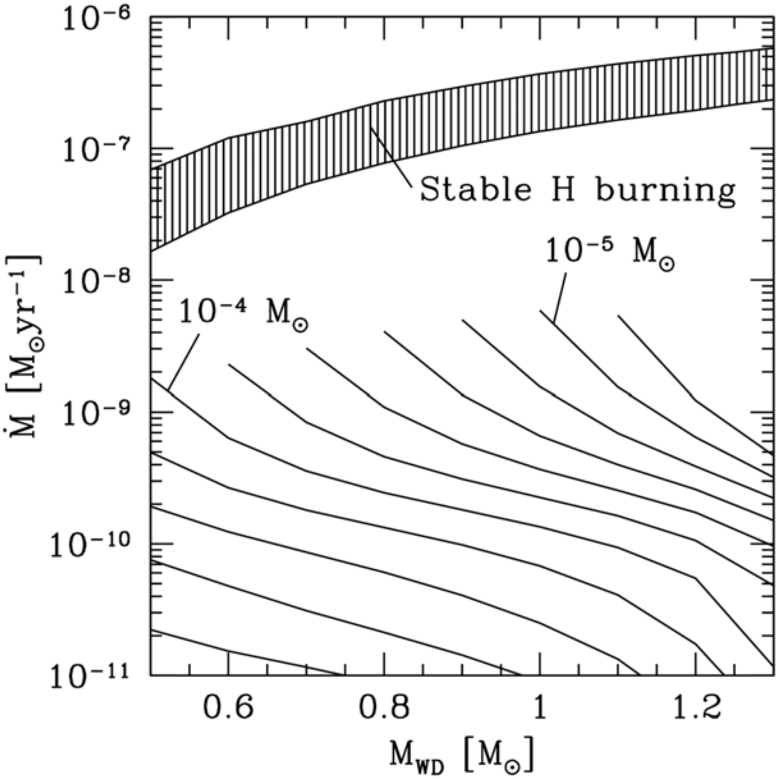


Figure 1. Hydrogen ignition masses $\Delta M_{\rm ign}$ for CO white dwarfs that have reached their equilibrium core temperatures. Contours are equally spaced in ignition masses, labels indicate $\Delta M_{\rm ign}$ in $M_\odot$. The vertically hatched region indicates where steady burning of H is expected (Nomoto 1982a). At higher mass-accretion rates either envelope build-up and expansion into a giant, or the development of a strong wind is expected (after Townsley & Bildsten 2005).

that, contrary to the case of a neutron star, in the case of a WD accretion of matter onto the surface produces far less energy than nuclear burning, so the main energy source of the SSS is nuclear burning, not accretion; see also Kahabka, van den Heuvel & Rappaport 1999.)

Although some symbiotic binaries and old novae also appear as SSS, the main group of SSS is that of the so-called ‘classical’ ones, which are binaries with orbital periods of one day to a few days, in which a donor star in the mass range $1.5-2.5\ M_\odot$ is transferring mass to the white dwarf on a thermal timescale of the donor. This yields typical mass-transfer rates of order $10^{-7}\ M_\odot$ per year. [These binaries are the higher-donor-mass analogues of the Cataclysmic Variables, and formed in the same way, through a phase of Common-Envelope evolution, starting from a wide binary consisting of a red giant with a degenerate core, plus an unevolved companion star of 1.5 to 2.5 $M_\odot$ (e.g. Rappaport & Di Stefano 1996; Kahabka & van den Heuvel 1997).] These classical SSS are an interesting subgroup of potential Type Ia SN progenitor candidates (e.g. Di Stefano 2010).

WDs with mass accretion rates below the range for steady burning, but still quite high, e.g. in the range between $10^{-8.5}$ and 10^{-7} $M_\odot$ per year, have only weak flashes of nuclear burning, such that most of the accreted matter may be retained. These systems, which may appear as various types of novae, may also contribute considerably to the Type Ia SN rate. This is the regime of accretion for observed recurrent novae such as RS Ophiuchi and T Corona Borealis, which are binaries composed of a solar-mass red-giant which fills its Roche lobe plus a massive WD (M$\sim$1.0 to 1.2 $M_\odot$). They erupt every few decades and with an average accretion rate of order $10^{-7.5}$ $M_\odot$/yr, they need $\sim 10^7$ yr to accrete the $\sim$0.3 $M_\odot$ needed to reach the Chandrasekhar limit. So, if these WDs are composed of C and O, they are excellent candidates for producing a Type Ia SN (Bildsten 2010).

Other suggested candidates for the hydrogen SD model are the symbiotic binaries, some of which are also SSS. These are wide binaries consisting of a red giant that does not fill its Roche lobe but has a strong stellar wind, and a WD that is accreting matter from this wind. As wind accretion is relatively inefficient and the lifetimes of the red giants are limited, it is questionable if many of the WDs in symbiotic binaries will ever be able to grow to the Chandrasekhar limit.

An interesting other type of SD model is one in which the donor star is a helium star. Such systems are, like the DD systems, the results of two CE phases. These systems were recognized as potentially interesting Type Ia SN candidate progenitors (e.g. Yungelson 2005 and references therein), which has been confirmed by later simulations (e.g. Wang & Han 2010). Here the accreted helium layer detonates and sends in a shockwave which may ignite carbon close to the centre of the white dwarf, triggering a Type Ia supernova. The mass of the white dwarf may in this case also be below the Chandrasekhar limit (e.g. Yungelson 2005; Bildsten 2010). Along similar lines, several authors have recently suggested that all Type Ia SNe might result from explosions of sub-Chandrasekhar-mass WDs (e.g. see Ruiter, Belczynski & Fryer 2009; Ruiter *et al.* 2010, and references given therein). However, since it remains to be seen whether these can indeed produce ‘standard candle-like’ explosions, as expected from WDs which all explode at the same Chandrasekhar mass, I do not further discuss these models here.

2.2 The relative importance of the SD and the DD scenarios for the Type Ia SN rate in galaxies

In order to examine the relative contributions of the SD and DD models to the Type Ia SN rates in different galaxies, several groups have carried out population synthesis evolution calculations assuming a realistic initial fraction of binary systems (usually between 50 and 100 per cent). Among the SD models one still has to distinguish between SD systems in which the donor is a hydrogen-rich star (SD,H) and one in which it is a helium-burning helium star (SD,He).

In such calculations one starts from a burst of star formation, and follows the evolution of the entire population of this starburst in the course of time, including the evolution of all the types of binary systems. For the starburst one assumes a distribution of the stellar masses (which includes the primary stars of binaries) according to the Initial Mass Function (IMF). The binaries are assigned an orbital semi-major axis and a mass ratio, taken from the observed distributions of these binary parameters (e.g. van den Heuvel 1994). One then has to use a binary evolution code in which the evolution of all types of binaries, with different initial primary star masses, orbital radii (orbits are mostly assumed to be circular) and mass ratios are included.

Here, a number of assumptions have to be made about what happens in certain stages in the evolution of binaries for which the outcome is presently still (very) uncertain. These are particularly the stages in which a binary loses much mass and orbital angular momentum, for example in the very important Common Envelope phases. The outcome of Common Envelope evolution depends critically on what prescription for the energy and angular momentum losses during this phase is assumed, and different authors use here different formalisms, and different values for the 'CE-efficiency' parameter α_{ce} for Common-Envelope evolution, which can give widely different results. This may, after two Common Envelope phases, lead to final orbital dimensions that differ by more than an order of magnitude. It appears that despite using widely different binary evolution and stellar evolution codes, the results obtained by different authors, although numerically quite different, show similar global trends for the predicted SD and DD Type Ia rates as a function of time following the starburst. As an illustration, Figs. 2 and 3 show the results obtained by the Moscow group (Yungelson 2005) and by Claeys *et al.* (2010), respectively. One observes in both cases that the DD model starts with a high rate some $(3-7)\times 10^7$ yr after the starburst, and then decays for a Hubble time following roughly a 1/t relation. The reason for this 1/t behaviour of the DD rate is well understood (e.g. Lipunov, Panchenko & Pruzhinskaya 2011; Moaz 2010).

On the other hand, for the simulations by Yungelson (2005), the (SD,He) model shows a broad peak between 4×10^7 yr and about 10^9 yr, which, during part of this time, can slightly exceed the DD rate. However, in the (SD, He) explosions, Yungelson also included sub-Chandrasekhar mass explosions, by assuming that after 0.15 $\mathrm{M}_{\odot}$ has been accreted also a sub-Chandra white dwarf would, after an edge-lit He-explosion, ignite C-burning close to its centre. Whether this will really occur is uncertain. If one excludes the sub-Chandra cases, the (SD, He) explosions in Yungelson's model terminate at $\sim 2\times 10^8$ yr, and this is also the case in the Claeys *et al.* model (see Fig. 3), and the same is true for the simulation of Wang & Han (2010) . In Yungelson's simulation the (SD,H) rate begins to rise at about 6×10^8 yr and cuts off at 2×10^9 years. On the other hand, in the Claeys *et al.* simulation with α_{ce}=1, the (SD,H) model begins to contribute already at 10^8 yr, and reaches values similar to the DD rate and then, like the DD rate, decays as 1/t for a Hubble time. One thus sees that only the behaviour of the (SD,H) rate is very different between the simulations of Figs. 2 and 3, due to quite different assumptions concerning the behaviour of the accreting white dwarfs for (SD,H) case. Other simulations, such as those by Mennekens *et al.* (2010a,b), Ruiter, Belczynski & Fryer (2009) and Ruiter *et al.* (2010) show a similar behaviour for the DD model, but for the two SD models, differences can be quite considerable.

Observations of the 'delay rates', i.e. the change of the Type Ia SN rate as a function of time, in elliptical galaxies (that have not had star formation for several billions of years), show that these follow a 1/t behavior (Totani *et al.* 2008; Moaz 2010), which appears to suggest that in elliptical galaxies the DD process dominates (although in the simulations Claeys *et al.* this could still be due to the (SD,H) model). For a detailed discussion of the results obtained by different authors, I refer to Nelemans (2010).

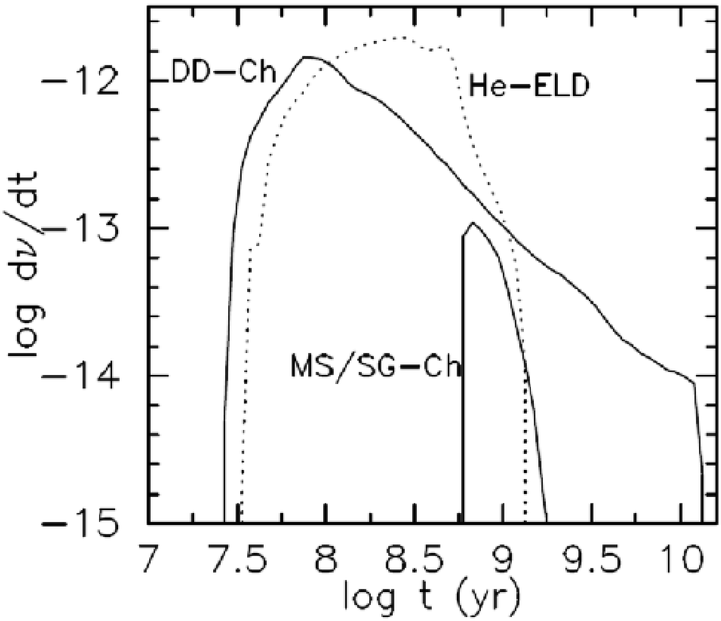


Figure 2. Rates of potential SNIa-scale events after a one year long burst of star formation that produces one solar mass of binary systems, after Yungelson (2005). The Helium-Edge-Lit Detonations include sub-Chandrasekhar-mass events. If these are excluded, the explosions of helium-accretors terminate about 200 million years after the burst of star formation. The MS/SG-Ch systems are the Single-Degenerate H-accretors (SuperSoft X-ray Sources), exploding at Chandrasekhar mass.

One observes from Figs. 2 and 3 that when the starburst is still young (from $(3-4)\times 10^7$ yr on) the (SD,He) and DD processes dominate, and only after $(1-6)\times 10^8$ yr the (SD,H) process begins to kick in, and then may be quite important for at least a few billion years. One also sees in these figures, that the absolute Type Ia rate predicted by these models is very high at young ages. Therefore, all models predict a much higher Type Ia SN rate for star-forming galaxies such as spirals and irregulars, than for elliptical galaxies. This fits very well with the observations, as it is well known that the Type Ia SN rate observed in star-forming galaxies is much higher than that in ellipticals (e.g. Garnavich 2010). The difference between the star-forming galaxies and the ellipticals is, apart form these different SN Ia rates, that the sole type of SNe found in ellipticals is the SN Ia, while in star-forming galaxies one also finds the core-collapse types of SNe: the Types II, Ib and Ic.

The conclusions from these population synthesis calculations are as follows.

(i) Their global predictions are in agreement with the observed trends of the evolution of the Type Ia SN rate with time in galaxies, and particularly,

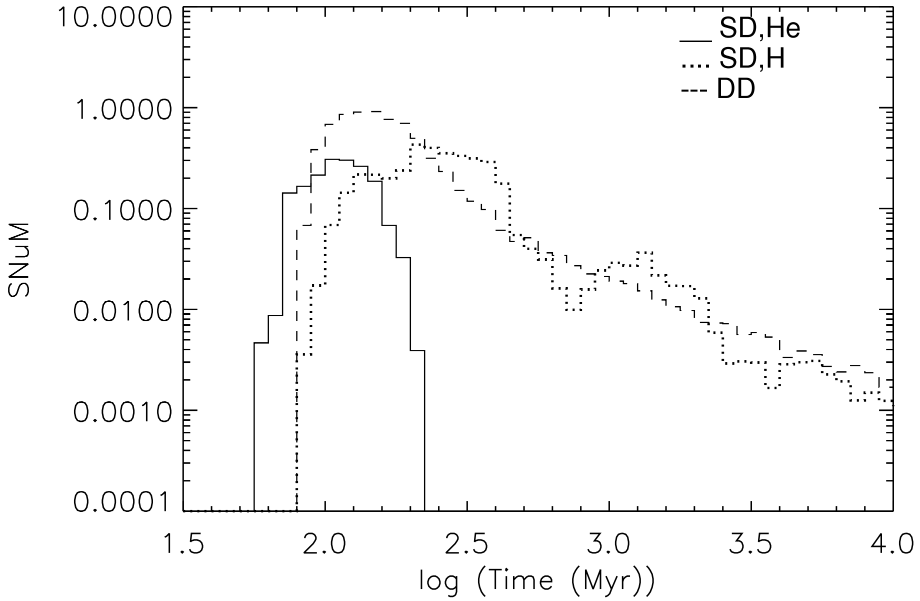


Figure 3. Type Ia SN event rates as a function of time after a single burst of star formation, for the three different progenitor scenarios indicated in the figure, as calculated by Claeys *et al.* (2010), assuming an efficiency parameter for Common-Envelope evolution $\alpha_{\rm ce}$=1.

(ii) the 1/t behaviour of the Type Ia time-delay curve observed in elliptical galaxies fits well with the predictions for the DD process obtained by all authors.

(iii) The much higher observed Type Ia SN rates in star-forming galaxies are well predicted by all models.

(iv) At early times $(3 \times 10^7$ to 2×10^8 yr) the Type Ia SN rate is expected to be dominated by the (SD,He) and the DD processes;

(v) while at middle ages $(0.2 - 2) \times 10^9$ yr the Type Ia SN are expected to be produced by a mix of the (SD,H), (SD,He) and DD processes.

3. Formation processes for neutron stars in binaries: Evidence for two different NS formation mechanisms, yielding different neutron star masses and kick velocities

3.1 Two classes of B-emission X-ray binaries

A very important discovery by Pfahl *et al.* (2002) was that there are two distinct classes of B-emission/neutron star systems (X-ray binaries as well as binary radio pulsars): one with orbits of small eccentricity (<0.25), in which the neutron star received hardly any velocity kick at birth, and a class with substantial orbital eccentricities (0.3 to 0.9) in which the neutron stars must have received a kick velocity of several hundred km/s at birth. A B-emission X-ray binary is a High Mass X-ray Binary (HMXB) consisting of a neutron star plus a rapidly rotating early B-type star, with a mass typically of $\sim$8 to 20 $M_\odot$ and an orbital period between about 15 days and several years. The class with low orbital eccentricities (low birth kicks) is substantial and may comprise the majority of all Be/X-ray binaries (Pfahl *et al.* 2002). This may be partly a 'selection effect', as a considerable fraction of the high-kick systems may have been disrupted at the birth of the neutron star.

3.2 Double neutron stars and their formation history: evidence that low kick velocities are related to low neutron star masses

Double-neutron-star systems tend to have very narrow orbits (see Table 1) and are the later evolutionary products of wide high-mass X-ray binary systems with orbital periods $>$100 days (van den Heuvel & Taam 1984; Bhattacharya & van den Heuvel 1991), which are mostly B-emission X-ray binaries. When the massive star in such a system has expanded to become a red giant, its envelope engulfs the neutron star, causing this star to spiral down into this envelope, reducing its orbital separation by several orders of magnitude. The large energy release due to friction and accretion during this spiral-in process is expected to cause the hydrogen-rich envelope of the giant to be expelled such that a very close binary remains, consisting of the helium core of the giant together with the neutron star. (Depending on the orbital separation at the onset of spiral-in, the helium core itself may already be (somewhat) evolved and possibly already have some C and O in its core.) Due to the large frictional and tidal effects during spiral-in, the orbit of the system is expected to be perfectly circular. The helium star that remains after the spiral-in generates its luminosity by helium burning, which produces C and O, and subsequently by carbon burning, produces Ne and Mg.

If the helium star has a mass in the range 1.6 to $\sim$2.8 $M_\odot$ (corresponding to a main-sequence progenitor in the range of 8 to 12($\pm$1) $M_\odot$; the precise limits of this mass range depend on metallicity and on the assumed model for convective energy transport; Podsiadlowski *et al.* 2004), it will, during carbon burning, develop a degenerate O-Ne-Mg core, surrounded by episodic C- and He-burning shells (Nomoto 1984; Habets 1986). When such a degenerate core develops, the envelope of the helium star begins to expand, causing the onset of mass transfer by Roche-lobe overflow in a binary system (Habets 1986; Dewi & Pols 2003). Roche-lobe overflow leads to the formation of an accretion disk around the neutron star and accretion of matter with angular momentum from this disk will cause the spin frequency of the neutron star to increase. Therefore one expects during the later evolution of these helium stars of relatively low mass, the first-born neutron star in the system to be 'spun up' to a short spin period. This neutron star had already a long history of accretion: first when it was in a wide binary with an early-type (presumably Be) companion; subsequently during the spiral-in phase into the envelope of its companion and now as companion of a Roche-lobe overflowing helium star. Since all binary pulsars which had a history of mass accretion (the so-called 'recycled' pulsars; Radhakrishnan & Srinivasan 1982, 1984) tend to have much weaker magnetic fields than normal single pulsars, it is thought that accretion in some way causes a weakening of the surface dipole magnetic field of neutron stars (Taam & van den Heuvel 1986). Several theories have been put forward to explain this accretion-induced field decay (e.g. Bhattacharya & Srinivasan 1995; Zhang

1998; Cumming, Arras & Zweibel 2004). With a field weakened to about 10^{10} Gauss (as observed in the recycled components of the double neutron stars (see Table 1), and an Eddington-limited accretion rate of helium (4×10^{-8} $M_{\odot}$/yr) a neutron star can be spun-up to a shortest possible spin period of a few tens of milliseconds (Smarr & Blandford 1976; Bhattacharya & van den Heuvel 1991). [If one assumes that spin-up requires Roche-lobe overflow, such spin-up will not take place if the helium star is more massive than about 3.5 $M_{\odot}$, because these stars do not greatly expand during their later evolution and therefore do not go through a sufficiently long-lasting phase of Roche-lobe overflow.]

When the helium star finally explodes as a supernova, the second neutron star in the system is born. This is a newborn neutron star without a history of accretion and is therefore expected to resemble the 'normal' strong-magnetic field single radio pulsars (Srinivasan & van den Heuvel 1982), which have typical surface dipole magnetic fields strengths of $10^{12} - 10^{13}$ Gauss. This theoretical expectation has been confirmed by the discovery of the double pulsar systems PSR J0737$-$3039AB, which consists of a recycled pulsar (star A) with a very rapid spin (P=23 ms) and a weak magnetic field ($7{\times}10^{9}$ G) and a normal strong-magnetic-field ($6{\times}10^{12}$ G) pulsar (star B), with a 'normal' pulse period of 2.8 sec (Burgay *et al.* 2003; Lyne *et al.* 2004; see Table 1). The explosive mass loss in the second supernova has made the orbit eccentric and since the two neutron stars are basically point masses, tidal effects in double neutron star systems will be negligible and there will be no tidal circularization of the orbit. (On timescales of tens of millions of years the orbits may be circularized by a few tens of per cent due to the emission of gravitational waves in the shortest-period system of PSR J0737$-$3039; assuming the observed pulsars to have an age of order half their spin down timescale, this is a negligible effect in all other systems, except in the final stages of spiraling together; e.g. Shapiro & Teukolsky 1983.)

3.3 A correlation between a small velocity kick and a low mass of the neutron star in double neutron star systems

In case of spherically symmetric mass ejection in the supernova explosion there is a simple relation between the orbital eccentricity and the amount of mass ΔM_{SN} ejected in the supernova:

$$e = \frac{\Delta M_{SN}}{M_{ns1} + M_{ns2}} \tag{1}$$

where M_{ns1} and M_{ns2} are the masses of the first- and the second-born neutron stars. I made calculations of the effect of the supernova mass loss plus a kick of 400 km/s, as observed for the young single pulsars (Hobbs *et al.* 2005) on the final orbital eccentricity of a double neutron star. I chose as a representative progenitor system: a binary with a circular orbit and a period of 4.8 hr, consisting of a 2 $M_{\odot}$ helium star plus a 1.38 $M_{\odot}$ neutron star, and assumed the helium star to leave a 1.25 $M_{\odot}$ neutron star, which received a randomly directed kick at birth of 400 km/s. From these calculations it was found that about half of all systems is disrupted by the explosion and that the systems that remain bound have on average an orbital eccentricity $>$0.7.

However, Table 1 shows that five out of the eight double neutron star systems known in the Galactic disk have eccentricities below 0.25. Taking into account that the sudden mass-loss effects of the supernova also induced an orbital eccentricity, this unusually large fraction of low-eccentricity systems strongly suggests that the second-born neutron stars in these systems received at most only a very small velocity kick at their births. This is further confirmed by the fact that, as Dewi, Podsiadlowski & Pols (2005) have shown, the observed correlation between the orbital eccentricity and the spin-period of the recycled neutron stars in the double neutron star systems (Faulkner *et al.* 2005) can be explained only if the second-born neutron stars received hardly any kick velocity at their birth (less than a few tens of km/sec). Indeed, Faulkner *et al.* had already pointed out that this relation could be understood on the basis of equation (1) if no kick had been imparted at birth to the second-born neutron star. It thus appears that, as pointed out by van den Heuvel (2004), these second-born neutron stars belong to the same ‘kick-less’ class as the neutron stars in the low-eccentricity class of Be/X-ray binaries (Pfahl *et al.* 2002).

The same holds for the young strong-magnetic-field pulsar in the eccentric radio-pulsar binary PSR J1145$-$6545 which has a massive white dwarf as a companion (Kaspi *et al.* 2000; Bailes *et al.* 2003; Bhat, Bailes & Verbiest 2008). The orbital eccentricity of 0.172 of this binary shows that the neutron star was the last-born object in the system, since formation of a white dwarf as the second-born object in the system cannot induce an orbital eccentricity (Kaspi *et al.* 2000; Bailes *et al.* 2003, van den Heuvel 2004; Bhat *et al.* 2008). The low value of its eccentricity would be hard to understand if the neutron star received the canonical 400 km/s kick (Hobbs *et al.* 2005) at its birth.

In the eccentric white-dwarf neutron-star system of PSR J1145$-$6545 the mass of the neutron star is known from the measurement of relativistic effects (periastron advance and Shapiro delay) to be 1.27($\pm$ 0.01) $M_\odot$ (Bhat *et al.*2008; number within parentheses indicates the 95% confidence boundary). Also in three of the other low-eccentricity double neutron stars the masses of both stars are accurately known from the measured relativistic effects.

(i) In PSR J0737$-$3039 the second-born neutron star has MB=1.2489($\pm$ 0.0007) and the first-born one has MA= 1.3381($\pm$ 0.0007) $M_\odot$ (Kramer *et al.* 2006).

(ii) In PSR J1756$-$2251 the second-born neutron star has a mass of 1.24($\pm$ 0.02) and the first-born one a mass of 1.32($\pm$ 0.02) $M_\odot$ (Stairs 2008).

(iii) In PSR J1906+0746 we observe the second-born (non-recycled) pulsar, which has a mass of 1.248($\pm$0.018) $M_\odot$, while its (invisible) recycled companion has a mass of 1.365($\pm$ 0.018) $M_\odot$ (Kasian 2008). The observed second-born pulsar here is very young ($\sim 10^5$ yr) and is spinning fast (P=0.144 sec)

In the other double neutron stars the masses of the stars are not yet accurately known, but in the two other low-eccentricity systems the second-born neutron stars must be less massive than 1.30 $M_\odot$ for the following reasons. In all double neutron star systems the relativistic parameter that can be measured most easily is the general relativistic rate of periastron advance, which directly yields the sum of the masses of the two neutron stars (e.g. Stairs 2004). In the systems of PSR J1518+4904 and PSR J1829+2456 the resulting sum of the masses turns out to be 2.62($\pm$ 0.07) $M_\odot$ (Nice, Sayer & Taylor 1996) and 2.53($\pm$0.10) $M_\odot$ (Champion *et al.* 2004), respectively. The individual masses of the neutron stars in these systems are still rather poorly determined, but in both these systems the already crudely determined other relativistic parameters indicate that the second-born neutron star has the lowest mass of the two (Stairs 2004). As in all these systems the sum of the masses is around 2.60 $M_\odot$, the second-born neutron stars in these systems cannot be more massive than 1.30 $M_\odot$.

One thus observes that in the six systems (out of seven, if J1145−6545 is included) with low orbital eccentricities the second-born neutron star has a low mass, close to 1.25 $M_\odot$ and belongs to the low-kick category. This is strong evidence that no (or a low) kick velocity is correlated with a low neutron star mass of around 1.25($\pm$0.05) $M_\odot$ (see also Schwab 2010).

A neutron star mass of $\sim$1.25 $M_\odot$ corresponds to a pre-collapse mass of about 1.44 $M_\odot$ as during the collapse the gravitational binding energy of the neutron star of about 0.20 $M_\odot$ (slightly depending on the assumed equation of state of neutronized matter) is lost in the form of neutrinos. So apparently the cores which collapsed to form these second-born neutron stars had a mass very close to the Chandrasekhar mass.

3.4 Formation mechanisms of neutron stars and possible resulting kicks

There are two basically different ways in which neutron stars are expected to form (Miyaji *et al.* 1980; Canal *et al.* 1990).

(I) In stars which originated in the main-sequence with mass in the range between 8 and about 12($\pm$1) $M_\odot$, and which are in binaries produce helium stars in the mass range 1.6 to 2.8 $M_\odot$. The O-Ne-Mg core which forms during carbon burning becomes degenerate and when its mass approaches the Chandrasekhar mass, and electron captures on Mg and Ne cause the core to collapse to a neutron star. Since these stars did not reach oxygen and silicon burning, the baryonic mass of the neutron star, which forms in this way is expected to be purely determined by the mass of the collapsing degenerate core, which is the Chandrasekhar mass. The gravitational mass of this neutron star is then the Chandrasekhar mass minus the gravitational binding energy of the neutron star, which is about 0.20 $M_\odot$. Thus a neutron star with a mass of about 1.25 $M_\odot$ is expected.

(II) In stars initially more massive than 12($\pm$1) $M_\odot$, the O-Ne-Mg core does not become degenerate and these cores proceed through oxygen and silicon burning to form an iron core. When the mass of this iron core exceeds the Chandrasekhar limit it collapses to form a neutron star. The precise way in which neutrino transport during core bounce and shock formation results in a supernova explosion is not yet fully understood. It appears that first the shock stalls and then several hundreds of milliseconds later is revitalized. Some fall back of matter from the layers surrounding the proto neutron star is expected to occur (e.g. Fryer 2004) such that the neutron star that forms may be substantially more massive than the baryonic mass of the collapsing Fe-core.

The fact that the pre-collapse masses of the low-mass, low-kick neutron stars were very close to the Chandrasekhar limit suggests that these neutron stars are the result of the electron-capture collapse of the degenerate O-Ne-Mg cores of helium stars that originated in the mass range 1.6 to about 2.8 $M_\odot$ (initial main-sequence mass in the range 8 to about 12 $M_\odot$). Can one understand why such neutron stars would not receive a large birth kick whereas those formed by the collapse of an iron core would?

Table 1. Double neutron star binaries in the Galactic disk.

	Spin			Compan.	Pulsar	Sum of		
Pulsar	per.	P_{orb}		Mass	Mass	masses	Bs	
Name	(ms)	(d)	E	$(M_\odot)$	$(M_\odot)$	$(M_\odot)$	$(10^{10}$ G)	Ref
J0737−3039A	22.7	0.10	0.088	1.2489(7)	1.3381(7)	2.5870(3)	0.7	1
J0737−3039B	2770	0.10	0.088	1.3381(7)	1.2489(7)	2.5870(3)	6×10^2	1
J1518+ 4904	40.9	8.63	0.249	$1.05^{+0.45}_{-0.11}$	$1.56^{+0.13}_{-0.45}$	2.62(7)	0.1	2
B1534+12	37.9	0.42	0.274	1.3452(10)	1.3332(10)	2.678(1)	1	3
J1756−2251	28.5	0.32	0.18	1.24 (2)	1.32 (2)	2.56(2)	0.54	4
J1811−1736	104	18.8	0.828	$1.11^{+0.53}_{-0.15}$	$1.62^{+0.22}_{-0.55}$	2.60 (10)	1.3	3
J1829+2456	41.0	1.18	0.139	$1.27^{+0.11}_{-0.07}$	$1.30^{+0.05}_{-0.05}$	2.53(10)	~ 1	5
J1909−3744	144	3.98	0.085	1.365(18)	1.248(18)*	2.613(9)	170	6
B1913+16	59	0.33	0.617	1.3873(3)	1.4408(3)	2.8281(1)	2	4
J1145−6545	394	0.20	0.172	1.01(1)WD	1.27(1)	2.28(1)	10^2	7

References: (1) Kramer *et al.* (2006); (2) Nice *et al.* (1996); (3) Stairs (2004); (4) Stairs (2008); (5) Champion *et al.* (2004); (6) Kasian (2008); (7) Bhat *et al.* (2008).

* The observed pulsar here is the second-born non-recycled strong-magnetic field one.

Burrows & Hayes (1996) have pointed out that the violent large-scale convective motions in the core during O- and Si-burning just prior to the formation of the Fe-core may produce considerable large-scale density inhomogeneities in the mantle of the proto-neutron star. They showed that this may lead to asymmetric neutrino transport and escape, which may easily impart enough momentum to the neutron star to produce a space velocity of 500 km/s. Recent 3-D numerical hydrodynamic core collapse and neutrino transport calculations by Scheck *et al.* (2004) and Arnett & Maekin (2011) confirm this expectation. As no O- and Si-burning occur prior to the e-capture collapse of a degenerate O-Ne-Mg core, neutrino transport in this case may be close to spherically symmetric, leading to no (or a very small) kick velocity imparted during collapse, as was confirmed by detailed numerical hydrodynamical calculations of such stellar cores by Kitaura *et al.* (2006).

3.5 Why are there no low-velocity, young, single radio pulsars?

A detailed statistical study by Hobbs *et al.* (2005) of all available pulsar proper motions showed that the observed velocity distribution of single young (age <3 million years) radio pulsars is excellently represented by one single Maxwellian with a mean 3-D speed of about 400 km/sec, and that there is no evidence for a bimodal velocity distribution (viz.: a separate lower-velocity population of young single pulsars) as had been suggested earlier (e.g. Arzoumanian, Chernoff & Cordes 2002). In terms of the above-described model this would mean that single pulsars are solely the products of iron-core collapse supernovae, whereas neutron star formation by electron-capture collapse would occur only in interacting binaries.

This is indeed precisely what has been suggested by Podsiadlowski *et al.* (2004) on grounds of stellar evolution considerations. These authors argued that if stars in the mass range 8 to about 12 $M_\odot$ are single, they will later in life evolve towards the Asymptotic Giant Branch (AGB), where the convective envelope during ‘dredge-up’ will penetrate the evolved helium core (which on the AGB has a degenerate O-Ne-Mg central core) and during this phase will erode away the outer helium layers of this core down to the degenerate O-Ne-Mg core. This prevents the latter core from further growth towards the Chandrasekhar limit. Their suggestion is therefore that single stars in the mass range 8 to 12($\pm$1) $M_\odot$ do not evolve to core collapse, but after heavy mass loss on the AGB will leave O-Ne-Mg white dwarfs.

Single stars more massive than about 12($\pm$ 1) $M_\odot$ will not produce degenerate O-Ne-Mg cores and will, in this picture, evolve through O- and Si-burning towards an iron-core collapse supernova. According to the above-described model calculations of Scheck *et al.*(2004) and Arnett & Maekin (2011) such a supernova presumably imparts a large kick velocity to the neutron star. Thus, combining these models with that of Podsiadlowski *et al.* (2004), one expects single pulsars to generally have received a large velocity kick at birth.

On the other hand, if the 8 to 12($\pm$1) $M_\odot$ star is in an interacting binary, the star cannot reach the AGB: before that time it already overflows its Roche lobe and loses its hydrogen-rich envelope by mass transfer towards its companion star (and in many cases, partly out of the system). For this reason in these systems a helium star in the mass range 1.6 to 2.8 $M_\odot$ will be left, which will produce a growing degenerate O-Ne-Mg core that evolves towards e-capture collapse. Thus, the e-capture collapse supernovae are, according to the model of Podsiadlowski *et al.* (2004), expected to solely occur in interacting binaries, and these will produce neutron stars of about 1.25 $M_\odot$.

These are then to be identified with the low-kick-velocity (low-mass) neutron stars that we observe in the double neutron star systems in Table 1 (van den Heuvel 2004; Podsiadlowski *et al.* 2005). One therefore would expect the low-kick low-mass neutron stars to solely be formed in interacting binaries, while single stars, or components of wide non-interacting binaries only produce high-kick-velocity neutron stars.

3.6 Consequences for the occurrence of neutron star formation by Accretion-Induced Collapse (AIC)

An important consequence of the above described model for the origins of kicks is that the accretion-induced collapse (AIC) of an O-Ne-Mg white dwarf in a close binary will not induce a sizeable kick velocity to the thus formed neutron star. Since in this process only the binding-energy mass equivalent of a neutron star ($\sim$0.2 $M_\odot$) is explosively lost, the mass-loss-induced runaway velocity of the resulting neutron-star binary is not expected to exceed a few tens of km/s for systems with Cataclysmic-Variable-like dimensions prior to the AIC. Such systems are therefore unlikely to escape from globular clusters and it seems most plausible, in view of the very large white dwarf populations in these clusters, that AIC is the dominant neutron-star forming mechanism in such clusters. This at the same time would explain why in some globular clusters, despite their ages of over ten billion years, still apparently young strongly magnetized radio pulsars are present (Lyne, Manchester & D’Amico 1996; an example is the strong-magentic-field globular cluster binary radio pulsar

PSR B1719−19, with P=1.0 second, B=10^{12} G). Although in globular clusters there is a large population of weakly magnetized millisecond pulsars, produced by accretion-driven binary recycling, the presence of already 3 short-lived strong-magnetic field radio pulsars in globular clusters implies that these have a birthrate that may be higher than that of the $\sim$100 millisecond pulsars known in globular clusters, as the latter ones will live almost eternally (spindown timescales of Gigayears).

Furthermore, the existence of the very wide radio pulsar binaries with circular orbits, such as PSR B0820+02 (orbital period 3.5 years) so far was very puzzling, as at the onset of the mass transfer from the low-mass red-giant progenitor of the white dwarf companion of this pulsar, the orbital period of the system was already about one year (Verbunt & van den Heuvel 1995). A neutron star that received a few hundred km/s kick at its birth could never have remained bound to a low-mass companion star in such a wide system. Formation from a white dwarf by a kick-less AIC in a wide symbiotic-type binary seems a plausible way to solve this problem (although direct formation from the e-capture collapse of a helium core of $\sim$2 $M_\odot$ cannot be excluded).

3.7 Massive neutron stars in binaries: A third type of neutron star?

Since 1975 it has been known that the accreting neutron star in the eclipsing High Mass X-ray Binary Vela X-1 (4U0900-40) has a mass considerably larger than the Chandrasekhar limit. The best modern determination of the mass of this neutron star from the Doppler-measurements of the orbits of both stars in the system is 1.86(±0.15) solar masses (Barziv *et al.* 2001; Quaintrell *et al.* 2003). In view of the very short lifetime of HMXBs the growth in mass by accretion of this neutron star was negligible, hence this neutron star must have been born directly with a large mass.

Models of the final evolution of massive stars by Timmes, Woosley & Weaver (1996) show that at an initial stellar mass of about 19 $M_\odot$ there occurs a jump in the mass of the collapsing iron core from 1.4 to $\sim$1.7 $M_\odot$. Hence, stars more massive than about 19 $M_\odot$, like the progenitor of the neutron star in Vela X-1, are expected to leave behind quite massive neutron stars in the range 1.8 to 2.0 $M_\odot$ (if a fall-back of a few tenths of a solar mass is included) or black holes, in the case of a large fall-back mass.

Interestingely, recently several new massive neutron stars have been discovered in binary radio pulsar systems: PSR J1614−2230 with a mass of 1.97±0.04 $M_\odot$ (Demorest *et al.* 2010) and PSR J1903+0327 with a mass of 1.667±0.021 $M_\odot$ (Freire *et al.* 2011). The latter pulsar is a millisecond one (P=2.15 ms) and its high mass could be the result of a long-lasting accretion phase of a neutron star that started out with a mass $\sim$1.4 $M_\odot$. The same might be true for the 1.97 $M_\odot$ neutron star PSR J1614−2230 (P=3.1 ms). This pulsar is in a relatively wide and circular orbit (P=8.7 days) with a 0.5 $M_\odot$ CO white dwarf companion. Its short pulse period suggests that it has accreted at least $\sim$0.1 $M_\odot$. It could have formed directly with a high mass in core collapse, if its progenitor started out with a mass in excess of 19 $M_\odot$. It is not difficult to make models for such an origin of this system (e.g Lin *et al.* 2010; Tauris, private communication). Alternatively, it might have started out with a mass of $\sim$1.4 $M_\odot$, but then it must have accreted some 0.6 $M_\odot$, which is a very large amount, but not impossible for a millisecond pulsar (e.g. van den Heuvel 1995). In the latter case, however, one wonders why it has not been spun up to a rotation period shorter than one millisecond. For this reason, it seems most likely to me that PSR J1614$-$2230 is the second example of a neutron star that was born already with a large mass, $\geq 1.70\ M_\odot$.

A third example is the compact star in the High Mass X-ray Binary 4U 1700$-$37. This compact star has a mass of $2.44 \pm 0.27\ M_\odot$, and has 0% probability to be $<1.60\ M_\odot$, and only 3.5% probability to be $<2.0\ M_\odot$ (Clark *et al.* 2002). It has never shown regular X-ray pulsations, but its X-ray spectrum is that of an accreting neutron star, which is significantly different from that of an accreting black hole (e.g. Clark *et al.* 2002). Its donor star is a highly luminous O6.5f supergiant star with a mass of $58 \pm 11\ M_\odot$, making it very likely that the progenitor of this compact star had a mass $>19\ M_\odot$. It therefore probably is the most massive neutron star known.

3.8 Discussion; some further consequences of the model

A consequence of the model in which neutron stars formed by electron-capture collapse receive hardly any kicks at birth, whereas those that are formed by Fe-core collapse receive large kicks, is that the formation of bound double neutron stars will be highly biased towards the lower-mass binaries, with components in the mass range 8 to about $12\ M_\odot$. The probability for disruption of such systems will be much lower than for systems in which the helium stars are above about $2.8\ M_\odot$. If the first supernova results from a helium star above $2.8\ M_\odot$, the resulting neutron star will get a large velocity kick, resulting in the formation of a high-eccentricity B-emission X-ray binary, or in the disruption of the system.

If the second helium star (resulting from the Be star in a Be/X-ray binary) has a mass above $2.8\ M_\odot$, it will again produce a high-kick neutron star, such that either the system is disrupted in the second supernova, or a very eccentric system results. As only three out of the eight double neutron stars in the Galactic disk have an orbital eccentricity larger than 0.25, and since some of these eccentricities may also have resulted from the pure mass-loss effects of a helium star with a mass just below $2.8\ M_\odot$, the formation of double neutron stars from systems in which the last-born helium stars are more massive than $2.8\ M_\odot$ seems to rarely occur in nature.

3.9 Summary

There is strong observational evidence, in combination with predictions from stellar evolution theory, for the existence of three classes of neutron stars, with two different formation mechanisms: electron-capture collapse of degenerate O-Ne-Mg cores in stars in binaries with intitial masses between ~ 8 and $\sim 12\ M_\odot$, and iron-core collapse for all stars, single as well as binary, more massive than $\sim 12\ M_\odot$.

Acknowledgments

I thank Gijs Nelemans, Lars Bildsten, Philip Podsiadlowski, Rosanne DiStefano and Thomas Tauris for many stimulating discussions on the physics and formation of compact stars and double neutron stars.

References

Arnett W.D., Maekin C., 2011, Preprint, University of Arizona, Steward Observatory

Arzoumanian Z., Chernoff D.F., Cordes J.M., 2002, ApJ, 568, 289

Bailes M., Ord S.M., Knight H., Hotan, A.W., 2003, ApJ, 595, L49

Barziv O., Kaper L., van Kerkwijk M.H., Telting J.H., van Paradijs J., 2001, A&A, 377, 925

Bhat N.D.R., Bailes M., Verbiest J.P.W., 2008, Phys. Rev.D. 77, 124017

Bhattacharya D., Srinivasan G., 1995, in Lewin W.H.G., van Paradijs J., van den Heuvel E.P.J., eds, The Magnetic Fields of Neutron Stars and Their Evolution, X-ray Binaries, Cambridge Univ. Press, p. 495

Bhattacharya D., van den Heuvel E.P.J., 1991, Physics Reports, 203, 1

Bildsten L., 2010, Lorentz Center Workshop, Observational signatures of Type Ia supernova progenitors, Leiden, The Netherlands, September 20-24, 2010

Burgay M., *et al.*, 2003, Nature, 426, 531

Burrows A., Hayes J., 1996, Phys. Rev. Letters, 76, 352

Canal R., Isern J., Labay J., 1990, ARA&A, 28, 183

Champion D.J., Lorimer D.R., McLaughlin M.A. Cordes J.M., Arzoumanian Z., Weisberg J.M., Taylor, J.H., 2004, MNRAS, 350, L61

Claeys J.S.W., Pols O.R., Vink J., Izzard R.G., 2010, in Kalogera V., van der Sluys M., eds, International Conference on Binaries, AIP Conference Proceedings, New York, 1314, 262

Clark J.S., Goodwin S.P., Crowther P.A., Kaper L., Fairbairn M., Langer N., Brocksopp C., 2002, A&A, 392, 909

Cumming A., Arras P., Zweibel E., 2004, ApJ, 609, 999

Demorest P.B., Pennucci T., Ransom S.M., Roberts M.S.E., Hessels, J.W.T., 2010, Nature, 467, 1081

Dewi J.D.M., Pols O.R., 2003, MNRAS, 344, 629

Dewi J.D.M., Podsiadlowski P., Pols, O.R., 2005, MNRAS, 363, L71

Di Stefano R. 2010, Lorentz Center Workshop, Observational signatures of Type Ia supernova progenitors, Leiden, The Netherlands, September 20-24, 2010

Faulkner A.J., *et al.*, 2005, ApJ., 618, L119

Freire P.C.C, *et al.*, 2011, MNRAS, 412, 2763

Fryer C.L., 2004, in Fryer C.L., ed, Stellar Collapse, Kluwer Acad. Publishers, Dordrecht

Garnavich P., 2010, Lorentz Center Workshop, Observational signatures of Type Ia supernova progenitors, Leiden, The Netherlands, September 20-24, 2010

Habets G.M.H.J., 1986, A&A, 167, 61

Hobbs G., Lorimer D.R., Lyne A.G., Kramer M., 2005, MNRAS, 360, 3, 974

Hoyle F., Fowler W.A., 1960, ApJ, 132, 565

Iben I., Tutukov A.V., 1984, ApJ, Suppl. 54, 335

Kahabka P., van den Heuvel E.P.J., 1997, Annual Rev. Astron. Ap. 35, 69

Kahabka P., van den Heuvel E.P.J. 2006, in Lewin W.H.G., van der Klis M., eds, Compact Stellar X-ray Sources, Cambridge University Press, p. 461

Kahabka P., van den Heuvel E.P.J., Rappaport S.A., 1999, Scientific American, 280, 28

Kasian L., 2008, in Bassa C., Wang Z., Cumming A., Kasspi V.M., eds, 40 years of Pulsars, Millisecond Pulsars, Magnetars and More, Am. Inst. of Phys. Conf. Series, 983, 369

Kaspi V.M., Lyne A.G., Manchester R.N., 2000, ApJ., 543, 1, 321

Kitaura F.S., Janka H.-Th., Muller E., 2006, A&A, 450, 345

Kramer M., *et al.*, 2006, Science, 314, 97

Lin J., Rappaport,lS.A., Podsiadlowski Ph., Nelson L., Paxton B., Todorov, P. 2010, arXiv: 1012.1877v1

Lipunov V.M., Panchenko I.E., Pruzhinskaya M.V. 2011, New Astronomy, 16, 250

Lyne A.G., Manchester R.N., D'Amico, N. 1996, ApJ 460, L41

Lyne A.G., *et al.*, 2004, Science, 303, 5661, 1153

Mennekens N., van Beveren D., de Greve J.-P., De Donder, E., 2010a, A&A, 515, A89

Mennekens N., van Beveren D., de Greve J.-P., De Donder, E., 2010b, in Kalogera V., van der Sluys M., eds, International Conference on Binaries, AIP Conference Proceedings, 1314, 239

Miyaji S., Nomoto K., Yokoi K., Sugimoto D., 1980, PASJ, 32, 303

Moaz D., 2010, in Kalogera V., van der Sluys M., eds, International Conference on Binaries, AIP Conference Proc., 1314, 223

Nelemans G., 2010, Lorentz Center Workshop, Observational signatures of Type Ia supernova progenitors, Leiden, The Netherlands September 20-24, 2010

Nice D.J., Sayer R.W., Taylor J.H., 1996, ApJ., 466, L87

Nomoto K., 1982a, ApJ 253, 798

Nomoto K., 1982b, ApJ 257, 780

Nomoto K., 1984, ApJ, 277, 791

Perlmutter S., *et al.*, 1999, ApJ, 517, 565

Pfahl E., Rappaport S., Podsiadlowski P., Spruit H., 2002, ApJ, 574, 364

Phillips M.M., Lira P., Suntzeff N.B., Schommer R.A., Hamuy M., Maza J., 1999, AJ, 118, 1766

Podsiadlowski P., Langer N., Poelarends A.J.T., Rappaport S., Heger, A., Pfahl, E., 2004, ApJ, 612, 1044

Podsiadlowski P., Dewi, J.D.M., Lasaffre P., Miller J.C., Newton J.G., Stone, J.R., 2005, MNRAS, 361, 1243

Prialnik D., Kovetz A., 1995, ApJ 445, 789

Quaintrell H., Norton, A.J., Ash T.D.C., Roche P., Willems B., Bedding T.R., Baldry I.K., Fender R.P., 2003, A&A, 401, 313

Pylyser E., Savonije G.J., 1988, A&A, 191, 57

Radhakrishnan V., Srinivasan G., 1982, Current Science, 51, 1096

Radhakrishnan V., Srinivasan G., 1984, in Hidayat B., Feast M.W., 1981, eds, Proc 2nd Asian-Pacific Regional Meeting on Astronomy, IAU Bandung Indonesia, p. 423

Rappaport S., Di Stefano R., 1996, IAUS, 165, 415

Riess A.G., *et al.*, 1998, AJ, 116, 1009

Ruiter A.J., Belczynski K., Fryer, C., 2009, ApJ, 699, 2026

Ruiter A.J., Belczynski K, Sim S.A., Hillebrandt W., Fink M., Kromer, M., 2010, in Kalogera V., van der Sluys M., eds, International Conference on Binaries, AIP Conference Proc, 1314, 233

Saio H., Nomoto, K., 1985, A&A, 150, L21

Saio H., Nomoto, K., 2004, ApJ, 615, 444

Scheck L., Plewa T., Janka H.-T., Kifonidis K., Mueller, E., 2004, Phys. Rev. Letters, 92a, 1103

Schmidt B.P., et al, 1998, ApJ, 507, 46

Schwab J., Podsiadlowski Ph., Rappaport S., 2010, ApJ, 719, 722

Shapiro S.L., Teukolsky S.A., 1983, The physics of compact objects, New York , Wiley-Interscience, 663

Smarr L.L., Blandford R.D., 1976, ApJ, 207, 574

Srinivasan G., van den Heuvel E.P.J., 1982, A&A, 108, 143

Stairs I.H., 2004, Science, 304, 547

Stairs I.H., 2008, in C. Bassa, Z. Wang, A. Cumming, V.M. Kasspi, eds, 40 years of Pulsars, Millisecond Pulsars, Magnetars and More, Am. Inst. of Phys. Conf. Series, 983, 424

Taam R.E., van den Heuvel, E.P.J., 1986, ApJ, 305, 235

Timmes F.X, Woosley S.A., Weaver T.A. 1996, ApJ, 457, 834

Totani T., Morokuma T., Oda T., Doi M., Yasuda N., 2008, PASJ, 60, 1327

Townsley D.M., Bildsten L., 2005, ApJ, 628, 395 van den Heuvel E.P.J., 1994, in Shore S.N., Livio M., van den Heuvel E.P.J., eds, Interacting Binaries, Springer, Heidelberg, p. 263

van den Heuvel E.P.J., 1995, JA&A, 16, 255

van den Heuvel E.P.J., 2004, in Schoenfelder V., Lichti G., Winkler C., eds, Proc. 5th INTEGRAL Workshop, ESA SP-552 (Noordwijk, ESA Publ.Div.ESTEC), p. 185

van den Heuvel, E.P.J., Taam, R. E., 1984, Nature, 309, 235

van den Heuvel E.P.J., Bhattacharya D., Nomoto K., Rappaport S.A., 1992, A&A, 262, 97

Verbunt F., Van den Heuvel E.P.J., 1995, in Lewin W.H.G., van Paradijs J., van den Heuvel E.P.J., eds, X-ray Binaries, Cambridge Univ. Press, p. 457

Wang B., Han, Z., 2010, in Kalogera V., van der Sluys M., eds, International Conference on Binaries, AIP Conference Proc., 1314, 244

Webbink R.F., 1984, ApJ, 277, 355

Whelan J., Iben I., 1973, ApJ, 186, 1007

Yungelson L.R., 2005, in Sion E.M., Vennes S., Shipman H.L., eds, White dwarfs: cosmological and galactic probes, ASSL, 332, 163 (arXiv:astro-ph/0409677)

Zhang C.M., 1998, A&A, 330, 195

This page intentionally left blank

Bull. Astr. Soc. India (2011)

Stability of relativistic stars

John L. Friedman[1,*] and Nikolaos Stergioulas[2]

[1]*Department of Physics, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA*
[2]*Department of Physics, Aristotle University of Thessaloniki, Thessaloniki, 56124, Greece*

Received April 4, 2011; accepted April 12, 2011

Abstract. Stable relativistic stars form a two-parameter family, parametrized by mass and angular velocity. Limits on each of these quantities are associated with relativistic instabilities discovered by Chandrasekhar: A radial instability, to gravitational collapse or explosion, marks the upper and lower limits on their mass; and an instability driven by gravitational waves may set an upper limit on their spin. Our summary of relativistic stability theory given here is based on and includes excerpts from the book *Rotating Relativistic Stars*, by the present authors (Friedman & Sterigioulas 2011).

Keywords : stars:general – stars:rotation – stars:neutron – instabilities

1. Introduction

A neutron star in equilibrium is accurately approximated by a stationary self-gravitating perfect fluid.[1] The character of its oscillations and their stability, however, depend on bulk and shear viscosity, on the superfluid nature of its interior, and – for modes near the surface – on the properties of the crust and the strength of its magnetic field.

The stability of a rotating star is governed by the sign of the energy of its perturbations; and the amplitude of an oscillation that is damped or driven by gravitational radiation is governed by the rate at which its energy and angular momentum are radiated. Noether's theorem relates the stationarity and axisymmetry of the equilibrium star to conserved currents constructed from the perturbed metric and fluid variables. Their integrals, the canonical energy and angular momentum on hypersurface can each be written as a functional quadratic in the perturbation, and the conservation laws express their change in terms of the flux of gravitational waves radiated to null infinity.

[*] e-mail: friedman@uwn.edu (JF), niksterg@auth.gr (NS)

[1] Departures from the local isotropy of a perfect fluid are associated with the crust; with magnetic fields that are thought to be confined to flux tubes in the superfluid interior; and with a velocity field whose vorticity is similarly confined to vortex tubes. Departures from perfect fluid equilibrium due to a solid crust are expected to be smaller than one part in $\sim 10^{-3}$, corresponding to the maximum strain that an electromagnetic lattice can support. The vortex tubes are closely spaced; but the velocity field averaged over meter scales is that of a uniformly rotating configuration. Finally, the magnetic field contributes negligibly to the pressure support of the star, even in magnetars with fields of 10^{15} G.

We begin with an action for perturbations of a rotating star from which these conserved quantities are obtained. The action was introduced by Chandrasekhar and his students in the Newtonian approximation (Chandrasekhar 1964; Clement 1964; Lymden Bell & Ostriker 1967), and its generalization to the exact theory was initially due to Chandra, in his pioneering paper on the stability of spherical relativistic stars (Chandrasekhar 1964). Several authors, including Taub, Carter, Chandrasekhar, Friedman, and Schutz (Taub 1954, 1969; Carter 1973; Chandrasekhar & Friedman 1972 a & b, Friedman & Schultz 1975; Friedman 1978) extended it to a Lagrangian formalism for rotating stars in general relativity.

We next review local stability to convection and to differential rotation. A spherical star that is stable against convection is stable to all nonradial perturbations: Only the radial instability to collapse (or explosion) can remain. A turning-point criterion governs stability against collapse and is associated with upper and lower limits on the masses of relativistic stars, the analog for neutron stars of the Chandrasekhar limit. Finally, we consider the additional instabilities of rotating stars. These are nonaxisymmetric instabilities that radiate gravitational waves. They may set an upper limit on the spin of old neutron stars spun up by accretion and on nascent stars that form with rapid enough rotation. Chandrasekhar's was again the pioneering paper, showing that gravitational radiation can drive a nonaxisymmetric instability (Chandrasekhar 1970).

2. Action and canonical energy

One can obtain an action for stellar perturbations by introducing a Lagrangian displacement ξ^α joining each unperturbed fluid trajectory (the unperturbed worldline of a fluid element) to the corresponding trajectory of the perturbed fluid. We denote by p, ϵ, ρ and u^α the fluid's pressure, energy density, rest-mass density and 4-velocity, respectively. A perturbative description can be made precise by introducing a family of (time dependent) solutions

$$\mathcal{Q}(\lambda) = \{g_{\alpha\beta}(\lambda), u^\alpha(\lambda), \rho(\lambda), s(\lambda)\}, \tag{1}$$

and comparing to first order in λ the perturbed variables $Q(\lambda)$ with their equilibrium values $Q(0)$.

Eulerian and Lagrangian changes in the fluid variables are defined by

$$\delta\mathcal{Q} := \frac{d}{d\lambda}Q(\lambda)\bigg|_{\lambda=0}, \qquad \Delta\mathcal{Q} = (\delta + \mathcal{L}_{\boldsymbol{\xi}})\mathcal{Q}, \tag{2}$$

with $\mathcal{L}_{\boldsymbol{\xi}}$ the Lie derivative along ξ^α.

Because oscillations of a neutron star proceed on a dynamical timescale, a timescale faster than that of heat flow, one requires that the Lagrangian change Δs in the entropy per unit rest mass vanishes, and perturbations of u^α, ρ and ϵ are expressed in terms of ξ^α and $h_{\alpha\beta} := \delta g_{\alpha\beta}$ by

$$\Delta u^\alpha = \frac{1}{2}u^\alpha u^\beta u^\gamma \Delta g_{\beta\gamma}, \qquad \Delta\rho = -\frac{1}{2}\rho q^{\alpha\beta}\Delta g_{\alpha\beta}, \qquad \Delta\epsilon = -\frac{1}{2}(\epsilon+p)q^{\alpha\beta}\Delta g_{\alpha\beta}, \tag{3}$$

with $\Delta g_{\alpha\beta} = h_{\alpha\beta} + \nabla_\alpha \xi_\beta + \nabla_\beta \xi_\alpha$. Our restriction to adiabatic perturbations means that the Lagrangian perturbation in the pressure, Δp is given by

$$\frac{\Delta p}{p} = \Gamma \frac{\Delta \rho}{\rho} = -\frac{1}{2}\Gamma\, q^{\alpha\beta}\Delta g_{\alpha\beta}, \tag{4}$$

where the adiabatic index Γ is defined by

$$\Gamma = \frac{\partial \log p(\rho, s)}{\partial \log \rho} = \frac{\epsilon + p}{p}\frac{\partial\, p(\epsilon, s)}{\partial \epsilon}. \tag{5}$$

The perturbed Einstein-Euler equations,

$$\delta(G^{\alpha\beta} - 8\pi\, T^{\alpha\beta}) = 0, \tag{6}$$

are self-adjoint in the weak sense that they are a symmetric system up to a total divergence: For any pairs $(\xi^\alpha, h_{\alpha\beta})$ and $(\widehat{\xi}^\alpha, \widehat{h}_{\alpha\beta})$, the symmetry relation has the form

$$\widehat{\xi}_\beta \delta(\nabla_\gamma T^{\beta\gamma}\sqrt{|g|}) + \frac{1}{16\pi}\widehat{h}_{\beta\gamma}\delta\left[(G^{\beta\gamma} - 8\pi T^{\beta\gamma})\sqrt{|g|}\right] = -2\mathcal{L}(\widehat{\xi}, \widehat{h}; \xi, h) + \nabla_\beta \Theta^\beta, \tag{7}$$

where $\mathcal{L}$ is symmetric under interchange of (ξ, h) and $(\widehat{\xi}, \hat{h})$. A symmetry relation of the form (7) implies that $\mathcal{L}^{(2)}(\xi, h) := \frac{1}{2}\mathcal{L}(\xi, h; \xi, h)$ is a Lagrangian density and

$$I^{(2)} = \int d^4x \mathcal{L}^{(2)} \tag{8}$$

is an action for the perturbed system.

The conserved canonical energy is associated with the timelike Killing vector is the Hamiltonian of the perturbation, expressed in terms of configuration space variables,

$$E_c = \int_S d^3x\, \alpha(\Pi^\alpha \mathcal{L}_{\mathbf{t}}\xi_\alpha + \pi^{\alpha\beta}\mathcal{L}_{\mathbf{t}} h_{\alpha\beta} - \mathcal{L}^{(2)}), \tag{9}$$

where α is the lapse function, and Π^α and $\pi^{\alpha\beta}$ are the momenta conjugate to ξ^α and $h_{\alpha\beta}$,

$$\Pi^\alpha = -n_\gamma \Pi^{\gamma\alpha}, \qquad \pi^{\alpha\beta} = -n_\gamma \pi^{\gamma\alpha\beta}, \tag{10}$$

with

$$\Pi^{\alpha\beta} = \frac{1}{2}\frac{\partial \mathcal{L}(\xi, h; \xi, h)}{\partial \nabla_\alpha \xi_\beta}, \tag{11}$$

$$\pi^{\alpha\beta\gamma} = \frac{1}{2}\frac{\partial \mathcal{L}(\xi, h; \xi, h)}{\partial \nabla_\alpha h_{\beta\gamma}}. \tag{12}$$

The negative signs in Eq. (10) are associated with the choice of a future pointing unit normal and the signature $- + ++$.

The corresponding canonical momentum has the form

$$J_c = \int_S d^3x\, \alpha(\Pi^\alpha \mathcal{L}_{\boldsymbol{\phi}}\xi_\alpha + \pi^{\alpha\beta}\mathcal{L}_{\boldsymbol{\phi}} h_{\alpha\beta}). \tag{13}$$

If one foliates the background spacetime by a family of spacelike but asymptotically null hypersurfaces, the difference $E_2 - E_1$ in E_c from one hypersurface to another to its future is the energy radiated in gravitational waves to future null infinity. Because this energy is positive definite, E_c can only decrease. This suggests that a condition for stability is that E_c be positive for all initial data.

This is, in fact, an appopriate stability criterion, but there is a subtlety, associated with a gauge freedom in choosing a Lagrangian displacement: There is a class of *trivial* displacements, for which the Eulerian changes in all fluid variables vanish. For a one (two) parameter equation of state, these correspond to rearranging fluid elements with the same value of ρ (and s).[2] For a trivial displacement η^α, the same physical perturbation is described by the pairs $h_{\alpha\beta}, \xi^\alpha$ and $h_{\alpha\beta}, \xi^\alpha + \eta^\alpha$, but the canonical energy is not invariant under addition of a trivial displacement, and its sign depends on this kind of gauge freedom. There is, however, a preferred class of *canonical* displacements, the displacements ξ^α that are orthogonal to all trivial displacements, with respect to the symplectic product of two perturbations,

$$W(\widehat{\xi}, \widehat{h}; \xi, h) := \int_\Sigma (\widehat{\Pi}_\alpha \xi^\alpha + \widehat{\pi}^{\alpha\beta} h_{\alpha\beta} - \Pi_\alpha \widehat{\xi}^a - \pi^{\alpha\beta} \widehat{h}_{\alpha\beta}) d^3x. \tag{14}$$

The criterion for stability can then be phrased as follows:

1. If $E < 0$ for some canonical data on Σ, then the configuration is unstable or marginally stable: There exist perturbations on a family of asymptotically null hypersurfaces Σ_u that do not die away in time.

2. If $E > 0$ for all canonical data on Σ, the magnitude of E is bounded in time and only finite energy can be radiated.

The trivial displacements are relabelings of fluid elements with the same baryon density and entropy per baryon. They are Noether-related to conservation of circulation in surfaces of constant entropy per baryon (Calkin 1963; Friedman & Schultz 1978), and canonical displacements are displacements that preserve the circulation of each fluid ring – for which the Lagrangian change in the circulation vanishes.

For perturbations that are not spherical, stable perturbations have positive energy and die away in time; unstable perturbations have negative canonical energy and radiate negative energy to infinity, implying that E becomes increasingly negative. One would like to show that when $E < 0$ a perfect-fluid configuration is strictly unstable, that within the linearized theory the time-evolved data radiates infinite energy and that $|E|$ becomes infinite along a family Σ_u of asymptotically null hypersurfaces. There is no proof of this conjecture, but it is easy to see that if $E < 0$, the time derivatives $\dot{\xi}^\alpha$ and $\dot{h}_{\alpha\beta}$ must remain finitely large. Thus a configuration with $E < 0$ will be strictly unstable unless it admits nonaxisymmetric perturbations that are time dependent but *nonradiative*.

[2] This is not the gauge freedom associated with infinitesimal diffeos of the metric and matter, but a redundancy in the Lagrangian-displacement description of perturbations that is already present in a Newtonian context.

3. Local stability

The criterion for the stability of a spherical star against convection is easy to understand. When a fluid element is displaced upward, if its density decreases more rapidly than the density of the surrounding fluid, then the element will be buoyed upward and the star will be unstable. If, on the other hand, the fluid element expands less than its surroundings it will fall back, and the star will be stable to convection.

As this argument suggests, criteria for convective stability are *local,* involving perturbations restricted to an arbitrarily small region of the star or, for axisymmetric perturbations, to an arbitrarily thin ring. For local perturbations, the change in the gravitational field can be ignored: A perturbation in density of order $\delta\epsilon/\epsilon$ that is restricted to a region of volume $V \ll R^3$ (R the radius of the star) can be regarded as adding or subtracting from the source a mass δm of order $\delta\epsilon V$. Then

$$\frac{\delta m}{M} \sim \frac{V}{R^3}\frac{\delta\epsilon}{\epsilon} \ll \frac{\delta\epsilon}{\epsilon}. \tag{15}$$

The change in the metric is then also smaller than $\delta\epsilon/\epsilon$ by the factor V/R^3, arbitrarily small when the support of the matter perturbation is arbitrarily small. Note that, because the metric perturbation is gauge-dependent, this statement about the smallness of the metric is also gauge-dependent. A more precise way of stating this property of a local perturbation is that a gauge can be chosen in which the metric perturbation is smaller than the density perturbation by a factor of order V/R^3.

Convective instability of spherical relativistic stars was discussed by Thorne (1966) and subsequently, with greater rigor, by Kovetz (1967) and Schutz (1970). An initial heuristic treatment by Bardeen (1970) of convective instability of differentially rotating stars was made more precise and extended to models with heat flow and viscosity by Seguin (1975).

Consider a fluid element displaced radially outward from an initial position with radial coordinate r to $r+\xi$. The displacement vector then has components $\xi^\mu = \delta^\mu_r\xi$. The fluid element expands (or, if displaced inward, contracts), with its pressure adjusting immediately – in sound travel time across the fluid element – to the pressure outside:

$$\Delta p = \boldsymbol{\xi}\cdot\nabla p = \frac{dp}{dr}\xi. \tag{16}$$

Heat diffuses more slowly, and the analysis assumes that the motion is faster than the time for heat to flow into or out of the fluid element: The perturbation is *adiabatic*:

$$\begin{aligned}\Delta\epsilon &= \left(\frac{\partial\epsilon}{\partial p}\right)_s \Delta p \\ &= \left(\frac{\partial\epsilon}{\partial p}\right)_s \frac{dp}{dr}\xi = \Gamma\frac{\epsilon+p}{p}\frac{dp}{dr}\xi,\end{aligned} \tag{17}$$

where $\Gamma := \left(\dfrac{\partial\log p}{\partial\log\rho}\right)_s$ and we have used the adiabatic conditions (3) and (4).

The difference $\Delta_\star\epsilon$ in the density of the surrounding star between r and $r+\xi$ is given by

$$\Delta_\star\epsilon = \xi\frac{d\epsilon}{dr}. \tag{18}$$

The displaced fluid element falls back if $|\Delta\epsilon| < |\Delta_\star\epsilon|$ – if, that is, the fluid element's density decreases more slowly than the star's density:

$$\left(\frac{\partial p}{\partial \epsilon}\right)_s \left|\xi \frac{dp}{dr}\right| < \left|\xi \frac{d\epsilon}{dr}\right|. \tag{19}$$

The star is then stable against convection if the inequality,

$$\left(\frac{dp}{d\epsilon}\right)_\star := \frac{dp/dr}{d\epsilon/dr} < \left(\frac{\partial p}{\partial \epsilon}\right)_s, \tag{20}$$

is satisfied, unstable if the inequality is in the opposite direction.

The convective stability criterion can also be stated in terms of the temperature gradient: If the temperature gradient is superadiabatic – if T decreases faster than an adiabatically displaced fluid element – then the star is unstable against convection.

Within seconds after its formation, a neutron star cools to a temperature below the Fermi energy per nucleon, below 10^{12} K $\sim$ 100 MeV. Its neutrons and protons are then degenerate, with a nearly homentropic equation of state. The star is convectively stable, but its convection modes have low frequencies (of order 100 Hz or smaller). The nonzero frequency arises from the composition gradient in the star, a changing ratio of neutrons to protons. A displaced fluid element does have time to adjust its composition to match that of the background star.

For spherical stars, any perturbation can be written as a superposition of spherical harmonics that are axisymmetric about some axis, and one therefore need only consider stability of axisymmetric perturbations. In fact, Detweiler & Ipser (1973) (generalizing a Newtonian result due to Lebovitz (1965)), show that, apart from local instability to convection, one need only consider radial perturbations: *If a nonrotationg star is stable to radial oscillations and stable against convection, the star is stable.* The Detweiler-Ipser argument shows that the Schwarzschild criterion (20) for stability against convection implies that there are no zero-frequency nonradial modes with polar parity, no time-independent polar-parity solutions to the perturbed Einstein-Euler system. The argument, by continuity of the frequency of outgoing modes, is compelling but not rigorous. It could be made more cleanly and without assumptions about normal modes if one could show directly that the canonical energy was always positive. This may follow from an integral inequality (associated with Eq. (42) of (Detweiler & Ipser 1973)), that is central to the Detweiler-Ipser argument. For a local perturbation – a perturbation for which the metric perturbation is negligible – the criterion for convective instability can easily be written in terms of the canonical energy E_c: For time-independent initial data with $\delta\epsilon = 0, \Delta\epsilon \neq 0$,

$$E_c = \int \frac{1}{\epsilon + p}\left[\left(\frac{\partial p}{\partial \epsilon}\right)_s - \left(\frac{dp}{d\epsilon}\right)_\star\right] \Delta\epsilon^2 \alpha dV, \tag{21}$$

and there are time-independent axisymmetric initial displacements ξ^α for which the canonical energy E_c of a rotating barotropic star is negative if and only if the generalized Schwarzschild criterion is violated.

3.1 Convective instability due to differential rotation: The Solberg criterion

Differentially rotating stars have one additional kind of convective (local) instability. If the angular momentum per unit rest mass, $j = hu_\alpha \phi^\alpha$, decreases outward from the axis of symmetry, the star is unstable to perturbations that change the differential rotation law.

The criterion is easy to understand in a Newtonian context. Consider a ring of fluid in the star's equatorial plane that is displaced outward from r to $r + \xi$, conserving angular momentum and mass. Again the displaced ring immediately adjusts its pressure to that of the surrounding star. If the ring's centripetal acceleration is larger that the net restoring force from gravity and the surrounding pressure gradient, it will continue to move outward. Now in the unperturbed star, the centripetal acceleration is equal to the restoring force. As in the discussion of convective instability, the displaced fluid element encounters the pressure gradient and gravitational field of the uperturbed star at its new position, and the restoring force is the restoring force on a fluid element at $r + \xi$ in the unperturbed star. Thus, if the displaced fluid ring has the same value of v^2/r as the surrounding fluid it will be in equilibrium, and the star will be marginally stable. If a displaced fluid ring has larger v^2/r than its surrounding fluid the star will be unstable.

The difference in acceleration for the background star is $\Delta_\star(v^2/r) = \xi^r \dfrac{d}{dr}(v^2/r)$, and stability then requires

$$\xi^r \frac{d}{dr}\left(\frac{v^2}{r}\right) - \Delta\frac{v^2}{r} > 0, \tag{22}$$

for $\xi^r > 0$.

Because $\Delta j = 0$ and $v(j,r) = j(r)/r$, we have

$$\Delta\frac{v^2}{r} = \Delta\frac{j^2}{r^3} = j^2 \xi^r \frac{d}{dr}\frac{1}{r^3}, \tag{23}$$

while

$$\Delta_\star \frac{v^2}{r} = \xi^r \frac{d}{dr}\frac{j^2}{r^3}, \tag{24}$$

implying

$$\Delta_\star \frac{v^2}{r} - \Delta\frac{v^2}{r} = \xi^r \frac{1}{r^2}\frac{dj^2}{dr}; \tag{25}$$

and the star is stable only if $\dfrac{dj}{dr} > 0$ in the equatorial plane (for $j > 0$), or, equivalently, only if $\partial_\varpi(\varpi^2\Omega) > 0$.

For relativistic stars, the same criterion ordinarily holds, where the specific angular momentum $j = hu_\phi$ is the angular momentum per unit rest mass. Bardeen (1970) gives a heuristic argument for this criterion, and a subsequent comprehensive treatment, including heat flow and viscosity, is due to Seguin (1975). Abramowicz (2004) provides a much quicker and more intuitive derivation for a homentropic star with no dissipation. (The last paragraph was its Newtonian version.)

For a differentially rotating homentropic star with metric

$$ds^2 = -e^{2\nu}dt^2 + e^{2\psi}(d\phi - \omega dt)^2 + e^{2\mu}(dr^2 + r^2 d\theta^2), \tag{26}$$

the angular momentum per unit baryon mass is $\dfrac{\epsilon+p}{\rho}u_\phi = \dfrac{\epsilon+p}{\rho}\dfrac{e^\psi v}{\sqrt{1-v^2}}$, where $v = e^{\psi-\nu}(\Omega-\omega)$ is the fluid velocity measured by a zero-angular-momentum observer. The canonical energy of a local axisymmetric perturbation with $\delta p = 0$ is given by

$$E_c = \int \frac{(\epsilon+p)}{(1-v^2)^2}\left[2v\xi^\alpha\nabla_\alpha(\psi-\nu) - (1+v^2)e^{\psi-\nu}\xi^\alpha\nabla_\alpha\omega\right]\frac{\partial v}{\partial j}\xi^\alpha\nabla_\alpha j\,\sqrt{-g}d^3x(27)$$

implying that there are perturbations for which $E_c < 0$ unless

$$\xi^\alpha\nabla_\alpha j > 0, \quad \text{for } \xi^\alpha \text{ outward-directed,} \tag{28}$$

where outward-directed is defined by

$$\xi^\alpha\left[\nabla_\alpha(\psi-\nu) - \frac{(1+v^2)}{2v}e^{\psi-\nu}\nabla_\alpha\omega\right] > 0. \tag{29}$$

The derivation of the criterion is valid for dust (pressure-free fluid) or for a single particle in the geometry of a rotating star or black hole, where it implies that a circular orbit is stable if and only if j increases outward along the surrounding family of circular equatorial orbits.

This is a simplest example of the turning-point criterion governing axisymmetric stability: A point of marginal stability along a sequence of circular orbits of a particle is a point at which j is an extremum. The turning-point condition can be rephrased in terms of the particle's energy. For a particle of fixed rest mass, the difference in energy of adjacent orbits is related to the difference in its angular momentum by

$$\delta E = \Omega\delta J.$$

Then a point of marginal stability along a sequence of circular orbits of a particle of fixed baryon mass is a point at which its energy is an extremum.

4. Instability to collapse: Turning point criterion

For spherical stars in the Newtonian approximation, instability sets in when the matter becomes relativistic, when the adiabatic index Γ (more precisely, its pressure-weighted average) reaches the value 4/3 characteristic of zero rest mass particles. This quickly follows from the Newtonian form of the canonical energy for radial perturbations of a spherical star: For an initial radial displacement ξ, with $\partial_t\xi = 0$,

$$E_c = \int_0^R dr\left\{\frac{4}{r}p'r^2\xi^2 + \frac{1}{r^2}\Gamma p\left[(r^2\xi)'\right]^2\right\}. \tag{30}$$

Choosing as initial data $\xi = r$ gives

$$E_c = \int_0^R drr^2p\left(\Gamma - \frac{4}{3}\right), \tag{31}$$

implying instability for $\Gamma < 4/3$.

In the stronger gravity of general relativity, even models with the stiffest equation of state must be unstable to collapse for some value of $R/M > 9/8$, the ratio for the most relativistic model of uniform density. By (in effect) computing the relativistic canonical energy,

$$E_c = \int_0^R e^{\lambda+\nu} \left\{ \left[\frac{4}{r} p' - \frac{p'^2}{\epsilon + p} + 8\pi p(\epsilon + p) \right] r^2 \xi^2 + \frac{e^{3\lambda-\nu}}{r^2} \Gamma p \left[(e^{-\nu} r^2 \xi)' \right]^2 \right\}, \quad (32)$$

Chandrasekhar (1964) showed that the stronger gravity of the full theory gives a more stringent condition for stability: A star is unstable if

$$\Gamma < \frac{4}{3} + K\frac{M}{R}, \quad (33)$$

where K is a positive constant of order 1. Because a gas of photons has $\Gamma = 4/3$ and massive stars are radiation-dominated, the instability can be important for stars with $M/R \gg 1$ (Chandrasekhar 1964; Fowler 1966).

Turning point instability

The best-known instability result in general relativity is the statement that instability to collapse is implied by a point of maximum mass and maximum baryon mass, along a sequence of uniformly rotating barotropic models with fixed angular momentum. A formal symmetry in the way baryon mass and angular momentum occur in the first law implies that (as in the case of circular orbits) points of instability are also extrema of angular momentum along sequences of fixed baryon mass.

For dynamical oscillations of neutron stars, the adiabatic index does not coincide with the polytropic index, $\Gamma \neq \dfrac{d \log p(r)/dr}{d \log \rho/dr}$. Chandrasekhar's criterion locates the point of dynamical instability, if one uses the adiabatic index in the canonical energy. The turning point method locates a *secular* instability — an instability whose growth time is long compared to the typical dynamical time of stellar oscillations. For spherical stars, the turning-point instability proceeds on a time scale slow enough to accommodate the nuclear reactions and energy transfer that accompany the change to a nearby equilibrium. For rotating stars, the time scale must also be long enough to accommodate a transfer of angular momentum between fluid rings. That is, the growth rate of the instability is limited by the time required for viscosity to redistribute the star's angular momentum. For neutron stars, this is expected to be short, probably comparable to the spin-up time following a glitch, and certainly short compared to the lifetime of a pulsar or an accreting neutron star. For this reason, it is the secular instability that sets the upper and lower limits on the mass of spherical and uniformly rotating neutron stars.

Note that, if one considers perturbations conforming to the effective equation of state satisfied by the equilibrium star, then Chandrasekhar's canonical energy criterion coincides with the turning-point criterion for spherical stars. The turning point criterion, however, has a longer history. In their 1939 paper, Oppenheimer and Volkoff had already used it to locate the stable part of a sequence of model neutron stars; and Misner & Zapolsky (1964) noticed that, along a sequence of neutron star models, the configuration at which the functional E_c first becomes negative appeared to be the model with maximum mass. In each case, they used models in which the equilibrium configuration and its perturbations are governed by the same one-parameter equation of state. A turning-point method, due initially to Poincaré (1885), then implies that points at which the stability of a mode changes are extrema of the mass (Harrison *et al.* 1965). See Thorne (1967) for a review of the turning point method applied to spherical neutron stars and (Thorne 1978) for later references; a somewhat different treatment is given by Zel'dovich and Novikov (1971). The generalization of the turning point criterion to rapidly rotating stars, due to Friedman, Ipser, and Sorkin (see below) (1988), is based on a general turning-point theorem due to Sorkin (1981, 1982).

One can easily understand why the instability sets in at an extremum of the mass by looking at a radial mode of oscillation of a nonrotating star with an equation of state $p = p(\rho), \epsilon = \epsilon(\rho)$. Along the sequence of spherical equilibria, a radial mode changes from stability to instability when its frequency σ changes from real to imaginary, with $\sigma = 0$ at the point of marginal stability. Now a zero-frequency mode is just a time-independent solution to the linearized Einstein-Euler equations – a perturbation from one equilibrium configuration to a nearby equilibrium with the same baryon number. From the first law of thermodynamics, a perturbation that keeps the star in equilibrium satisfies

$$\delta M = \frac{\mu}{u^t} dN, \tag{34}$$

with μ the chemical potential and N the number of baryons. The relation implies that, for a zero frequency perturbation involving no change in baryon number, the change δM in mass must vanish. This is the requirement that the mass is an extremum along the sequence of equilibria. Models on the *high-density* side of the maximum-mass instability point are unstable: Because the turning point is a star with maximum baryon number as well as maximum mass, there are models on opposite sides of the turning point with the same baryon number. Because μ/u^t is a decreasing function of central density, the model on the high-density side of the turning point has greater mass than the corresponding model with smaller central density.

At the minimum mass, it is the *low-density* side that is unstable: Because the mass is a minimum, the model on the low-density side of the turning point has greater mass than the corresponding model with the same baryon number on the high-density side.

The precise statement of the turning-point criterion is the following result:

Theorem (Friedman, Isper & Sorkin 1988). Consider a continuous sequence of uniformly rotating stellar models based on an equation of state of the form $p = p(\epsilon)$. Let λ be the sequence parameter and denote the derivative $d/d\lambda$ along the sequence by $(\dot{\ })$.
(i) Suppose that the total angular momentum is constant along the sequence and that there is a point λ_0 where $\dot{M} = 0$ and where $\mu > 0$, $(\dot{\mu}\dot{M})^{\cdot} \neq 0$. Then the part of the sequence for which $\dot{\mu}\dot{M} > 0$ is unstable for λ near λ_0.
(ii) Suppose that the total baryon mass M_0 is constant along the sequence and that there is a point λ_0 where $\dot{M} = 0$ and where $\Omega > 0$, $(\dot{\Omega}\dot{M})^{\cdot} \neq 0$. Then the part of the sequence for which $\dot{\Omega}\dot{M} > 0$ is unstable for λ near λ_0.

Friedman, Ipser & Sorkin (1988) point out the symmetry between M_0 and J that implies the maximum-J form of the theorem, and Cook, Shapiro & Teukolsky (1992) first use the theorem in this form.

For rotating stars, the turning point criterion is a sufficient condition for secular instability to collapse. In general, however, collapse can be expected to involve differential rotation, and the turning point identifies only nearby uniformly rotating configurations with lower energy. Rotating stars are therefore likely to be secularly unstable to collapse at densities slightly lower than the turning point density. The onset of secular instability to collapse is at or before the onset of dynamical instability along a sequence of uniformly rotating stars of fixed angular momentum, and recent work by Rezzolla, Katami and Yoshida (2011) appears to show that rapidly rotating stars can also be dynamically unstable to collapse just prior to the turning point.

Searches to determine the line of turning points have covered the set of models with sequences of constant rest mass M_0, extremizing J on each one, or vice versa. This is a computationally expensive procedure, and a more efficient way is summarized in the following corollary due to Jocelyn Read (Read *et al.* 2009):

Regard M_0 and J as functions on the two-dimensional space of equilibria. Turning points are the points at which ∇M_0 and ∇J are parallel. An equivalent statement of this criterion is that the wedge product of the gradients vanishes: $dM_0 \wedge dJ = 0$; or, with the space of equilibria embedded in a 3-dimensional space, $\nabla M_0 \times \nabla J = 0$. In particular, with the space of equilibria parametrized by the central energy density ϵ_c and axis ratio $\mathfrak{r} = r_p/r_e$, the turning points satisfy

$$\frac{\partial(M_0, J)}{\partial(\epsilon_c, \mathfrak{r})} \equiv \frac{\partial M_0}{\partial \epsilon_c}\frac{\partial J}{\partial \mathfrak{r}} - \frac{\partial J}{\partial \epsilon_c}\frac{\partial M_0}{+\partial \mathfrak{r}} = 0. \tag{35}$$

5. Nonaxisymmetric instabilities

Rapidly rotating stars and drops of water are unstable to a bar mode that leads to fission in the water drops and is likely to be the reason many stars in the Universe are in close binary systems. Galactic disks are unstable to nonaxisymmetric perturbations that lead to bars and to spiral structure. And a related instability of a variety of nonaxisymmetric modes, driven by gravitational waves, the Chandrasekhar-Friedman-Schutz (CFS) instability (Chandrasekhar 1970; Friedman & Schultz 1978; Friedman 1978), may limit the rotation of young neutron stars. The existence of this gravitational-wave driven instability in rotating stars was first found by Chandrasekhar (1970) in the case of the $l = 2$ mode in uniformly rotating, uniform density Maclaurin spheroids. Subsequently, Friedman and Schutz (1978) showed that all rotating self-gravitating perfect fluid configurations are generically unstable to the emission of gravitational waves. Along a sequence of stars, the instability sets in when the frequency of a nonaxisymmetric mode vanishes in the frame of an inertial observer at infinity, and such zero-frequency modes of rotating perfect-fluid stellar models are marginally stable.

This review begins with a discussion of the CFS instability for perfect-fluid models and then outlines the work that has been done to decide whether the instability is present in young neutron stars and in old neutron stars spun up by accretion. For very rapid rotation and for slower but highly differential rotation, nonaxisymmetric modes can be *dynamically unstable*, with growth times comparable to the period of a star's fundamental modes, and the review ends with a brief discussion of these related dynamical instabilities.

To understand the way the CFS instability arises, consider first a stable spherical star. All its modes have positive energy, and the sign of a mode's angular momentum J_c about an axis depends on whether the mode moves clockwise or counterclockwise around the star. That is, a mode with angular and time dependence of the form $\cos(m\phi - \sigma_0 t)e^{-\alpha_0 t}$, has positive angular momentum J_c about the z-axis if and only if the mode moves in a positive direction: $\dfrac{\sigma_0}{m}$ is positive. Because the wave moves in a positive direction relative to an observer at infinity, the star radiates positive angular momentum to infinity, and the mode is damped. Similarly, a mode with negative angular momentum has negative pattern speed $\dfrac{\sigma_0}{m}$ and radiates negative angular momentum to infinity, and the mode is again damped.

Now consider a slowly rotating star with a backward-moving mode, a mode that moves in a direction opposite to the star's rotation. Because a short-wavelength fluid mode (a mode with a Newtonian counterpart, not a w-mode) is essentially a wave in the fluid, the wave moves with nearly the same speed relative to a rotating observer that it had in the spherical star. That means that an observer at infinity sees the mode dragged forward by the fluid. The frequency σ_r seen in a rotating frame is the frequency associated with the ϕ coordinate $\phi_r = \phi - \Omega t$ of a rotating observer, $\sigma_r = \sigma - m\Omega$. Then

$$m\phi - \sigma t = m\phi_r - (\sigma + m\Omega)t = m\phi_r - \sigma_r t,$$

implying that the frequency seen by the rotating observer is

$$\sigma_r = \sigma - m\Omega. \tag{36}$$

For a slowly rotating star, $\sigma_r \approx \sigma_0$. When the star rotates with an angular velocity greater than $|\sigma_r/m|$, the backward-going mode is dragged *forward* relative to an observer at infinity:

$$\frac{\sigma}{m} = \frac{\sigma_r}{m} + \Omega \tag{37}$$

is positive.

Because the pattern speed σ/m is now positive, the mode radiates positive angular momentum to infinity. But the canonical angular momentum is still negative, because the mode is moving backward relative to the fluid: The angular momentum of the perturbed star is smaller than the angular momentum of the star without the backward-going mode. As the star radiates positive angular momentum to infinity, J_c becomes increasingly negative, implying that the amplitude of the mode grows in time: *Gravitational radiation now drives the mode instead of damping it.*

For large m or small σ_0, σ/m will be positive when $\Omega \approx |\sigma_0/m|$. This relation suggests two classes of modes that are unstable for arbitrarily slow rotation: Backward-moving modes with large values of m and modes with any m whose frequency is zero in a spherical star. Both classes of perturbations exist. The usual p-modes and g-modes have finite frequencies for a spherical star and are unstable for $\Omega \gtrsim \sigma_0/m$; and r-modes, which have zero frequency for a non-rotating barotropic star, are unstable for all values of m and

Ω (that is, those r-modes are unstable that are backward-moving in the rotating frame of a slowly rotating star).

We have so far not mentioned the canonical energy, but our key criterion for the onset of instability is a negative E_c. If we ignore the imaginary part of the frequency, the change in the sign of E_c follows immediately from the relation $J_c = -\sigma_p E_c$. To take the imaginary part $\mathrm{Im}\sigma = \alpha \neq 0$ of the frequency into account, we need to use the fact that energy is lost at a rate $\dot{E}_c \propto \dddot{Q}^2 \propto \sigma^6$ for quadrupole radiation, with $\dot{E}_c$ proportional to higher powers of σ for radiation into higher multipoles. Because E_c is quadratic in the perturbation, it is proportional to $e^{-2\alpha t}$, implying $\alpha \propto \sigma^6$. Thus $\alpha/\sigma \to 0$ as $\sigma \to 0$, implying that for a normal mode E_c changes sign when σ_p changes sign.

Although the argument we have given so far is heuristic, there is a precise form of the statement that a stable, backward-moving mode becomes unstable when it is dragged forward relative to an inertial observer (Friedman & Schultz 1978; Friedman & Stergioulas 2011).

Theorem. Consider an outgoing mode $(h_{\alpha\beta}(\lambda), \xi^\alpha(\lambda))$, that varies smoothly along a family of uniformly rotating perfect-fluid equilibria, labeled by λ. Assume that it has t and ϕ dependence of the form $e^{i(m\phi-\sigma t)}$, that $\sigma = \mathrm{Re}\{\sigma\}$ satisfies $\sigma/m - \Omega < 0$ for all λ, and that the sign of σ/m is negative for $\lambda < \lambda_0$ and positive for $\lambda > \lambda_0$. Then in a neighborhood of λ_0, $\alpha := \mathrm{Im}\{\sigma\} \leq 0$; and if the mode has at least one nonzero asymptotic multipole moment at future null infinity, the mode is unstable ($\alpha < 0$) for $\lambda > \lambda_0$.

A corresponding result that does not rely on existence or completeness of normal modes is the statement that one can always choose canonical initial data to make $E_c < 0$ (Friedman 1978; Friedman & Stergioulas 2011).

The growth time τ_{GR} of the instability of a perfect fluid star is governed by the rate $\left.\dfrac{dE}{dt}\right|_{\mathrm{GR}}$ at which energy is radiated in gravitational waves:

$$\frac{1}{\tau_{\mathrm{GR}}} = -\frac{1}{2E_c}\left.\frac{dE_c}{dt}\right|_{\mathrm{GR}}, \tag{38}$$

where (Thorne 1980)

$$\left.\frac{dE}{dt}\right|_{\mathrm{GR}} = -\sigma(\sigma + m\Omega)\sum_{l\geq 2} N_l \sigma^{2l}\left(|\delta D_{lm}|^2 + |\delta J_{lm}|^2\right), \tag{39}$$

where D_{lm} and J_{lm} are the asymptotically defined mass and current multipole moments of the perturbation and $N_l = \dfrac{4\pi(l+1)(l+2)}{l(l-1)[(2l+1)!!]^2}$ is, for low l, a constant of order unity. In the Newtonian limit,

$$\delta D_{lm} = \int \delta\rho\, r^l Y_{lm} d^3x. \tag{40}$$

For a star to be unstable, the growth time τ_{GR} must be shorter than the viscous damping time $\tau_{\mathrm{viscosity}}$ of the mode, and the implications of this are discussed below. In particular because the growth time is longer for larger l, only low multipoles can be unstable in neutron stars.

Modes with polar and axial parity

The spherical symmetry of a nonrotating star and its spacetime implies that perturbations can be labeled by fixed values l, m labeling an angular harmonic: The quantities $h_{\alpha\beta}, \xi^\alpha, \delta\rho, \delta\epsilon, \delta p, \delta s$ that describe a perturbation are all proportional to scalar, vector and tensor spherical harmonics constructed from Y_{lm}, and perturbations with different l, m values decouple. Similarly, because spherical stars are invariant under parity (a map of each point P of spacetime to the diametrically opposite point on the symmetry sphere through P), perturbations with different parity decouple, the parity of a perturbation is conserved, and normal modes have definite parity. Perturbations associated with an l, m angular harmonic are said to have *polar* parity if they have the same parity as the function Y_{lm}, $(-1)^l$. Perturbations having parity $(-1)^{l+1}$, opposite to that of Y_{lm} have axial parity. In the Newtonian literature, modes of a rotating star that are continuously related to polar modes of a spherical star are commonly called *spheroidal*; while modes whose spherical limit is axial are called *toroidal*.

Every rotational scalar – ϵ, p, ρ, and the components of the perturbed metric $h_{\alpha\beta}$ and the perturbed fluid velocity δu^α in the t-r subspace – can be expressed as a superposition of scalar spherical harmonics $Y_{\ell m}$. As a result, modes of spherical stars that involve changes in any scalar are polar. On the other hand, the angular components of velocity perturbations can have either polar parity, with

$$\delta v = f(r)\nabla Y_{lm} \tag{41}$$

or axial parity, with Newtonian form

$$\delta v = f(r)\mathbf{r} \times \nabla Y_{lm}, \tag{42}$$

and the relativistic form $\delta u^\alpha \propto \epsilon^{\alpha\beta\gamma\delta}\nabla_\beta t \nabla_\gamma r \nabla_\delta Y_{lm}$.

There are two families of polar modes of perfect-fluid Newtonian stars, p-modes (pressure modes) and g-modes (gravity modes). For short wavelengths, the p-modes are sound waves, with pressure providing the restoring force and frequencies

$$\sigma = c_s k, \tag{43}$$

where k is the wavenumber and c_s is the speed of sound. The short-wavelength g-modes are modes whose restoring force is buoyancy, and their frequencies are proportional to the Brunt-Väisälä frequency, related to the difference between $dp/d\epsilon$ in the star and $c_s^2 = \partial p(\epsilon, s)/\partial\epsilon$. The fundamental modes of oscillation of a star (f-modes), with no radial nodes, can be regarded as a bridge between g-modes and p-modes.

Because axial perturbations of a spherical star involve no change in density or pressure, there is no restoring force in the linearized Euler equation, and the linear perturbation is a time-independent velocity field – a zero-frequency mode.[3] In a rotating star, the axial

[3] Axial perturbations of the spacetime of a spherical star include both axial perturbations of the fluid and gravitational waves with axial parity. The axial-parity waves do not couple to the fluid perturbation, which is stationary in the sense that $\partial_t \delta u_\alpha = 0$.

modes acquire a nonzero frequency proportional to the star's angular velocity Ω, a frequency whose Newtonian limit has the simple form

$$\sigma = \frac{(l-1)(l+2)}{l(l+1)} m\Omega, \tag{44}$$

where the harmonic time and angular dependence of the mode is $e^{i(m\phi - \sigma t)}$. These modes are called r-modes, their name derived from the Rossby waves of oceans and planetary atmospheres. The term r-mode can be usefully regarded as a mnemonic for a *rotationally restored* mode. Equation (36) implies that the r-mode associated with every nonaxisymmetric multipole obeys the instability condition for every value of Ω: It is forward moving in an inertial frame and backwards moving relative to a rotating observer:

$$\sigma_r = -\frac{2m}{l(l+1)}\Omega, \tag{45}$$

with sign opposite to that of σ and m. Because the rate at which energy is radiated is greatest for the $l = m = 2$ r-mode, that is the mode whose instability grows most quickly and which determines whether an axial-parity instability can outpace viscous damping.

The instability of low-multipole r-modes for arbitrarily slow rotation is strikingly different from the behavior of the low-multipole f- and p-modes, which are unstable only for large values of Ω. The reason is that the frequencies of f- and p-modes are high, and, from Eq. (37), a correspondingly high angular velocity is needed before a mode that moves backward relative to the star is dragged forward relative to an inertial observer at infinity. Of the polar modes, f-modes with $l = m$ have the fastest growth rates; their instability points for uniformly rotating relativistic stars, found by Stergioulas (Friedman & Stergioulas 2011), are shown in Figure 1. (Work on these stability points of relativistic stars is reported in (Stergioulas & Friedman 1998; Yoshida & Eriguchi 1997; Yoshida & Eriguchi 1999; Zink *et al.* 2010; Gaertig *et al.* 2011)

The figure shows that, for uniform rotation, the $l = m = 2$ f-mode is unstable only for stars with high central density and therefore with masses greater than 1.4 $M_\odot$. Neutron stars, however, rotate differentially at birth, and the $l = 2$ mode, as well as f-modes with $l \geq 5$, could be initially unstable.

Implications of the instability

The nonaxisymmetric instability may limit the rotation of nascent neutron stars and of old neutron stars spun up by accretion; and the gravitational waves emitted by unstable modes may be observable by gravitational wave detectors. Whether a limit on spin is in fact enforced depends on whether the instability of perfect-fluid models implies an instability of neutron stars; and the observability of gravitational waves also requires a minimum amplitude and persistence of an unstable mode. We briefly review observational support for an instability-enforced upper limit on spin and then turn to the open theoretical issues.

Evidence for an upper limit on neutron-star spin smaller than the Keplerian frequency Ω_K comes from nearly 30 years of observations of neutron stars with millisecond periods, seen as pulsars and as X-ray binaries. The observations reveal rotational frequencies

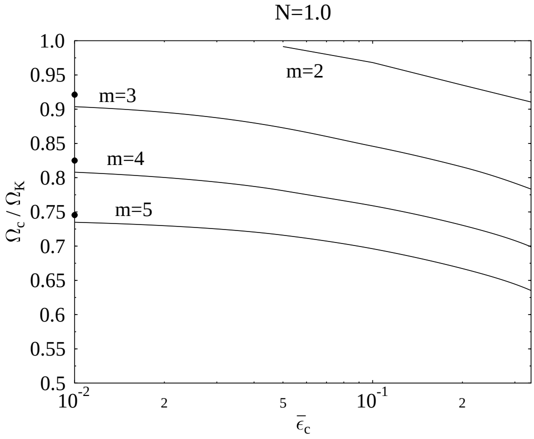


Figure 1. Critical angular velocity Ω/Ω_K vs. the dimensionless central energy density $\bar{\epsilon}_c$ for the $m = 2$, 3, 4 and 5 neutral modes of $N = 1.0$ polytropes. The filled circles on the vertical axis are the Newtonian values of the neutral points for each mode.

ranging upward to 716 Hz and densely populating a range of frequencies below that. Selection biases against detection of the fastest millisecond radio pulsars have made conclusions about an upper limit on spin uncertain, but Chakrabarty argues that the class of sources whose pulses are seen in nuclear bursts (nuclear powered accreting millisecond X-ray pulsars) constitute a sample without significant bias (Chakrabarty 2008); their distribution of spins is shown in Fig. 1 of that paper, reproduced as Fig. 2 below.

Summarizing his analysis, Chakrabarty writes, “There is a sharp cutoff in the population for spins above 730 Hz. RXTE has no significant selection biases against detecting oscillations as fast as 2000 Hz, making the absence of fast rotators extremely statistically

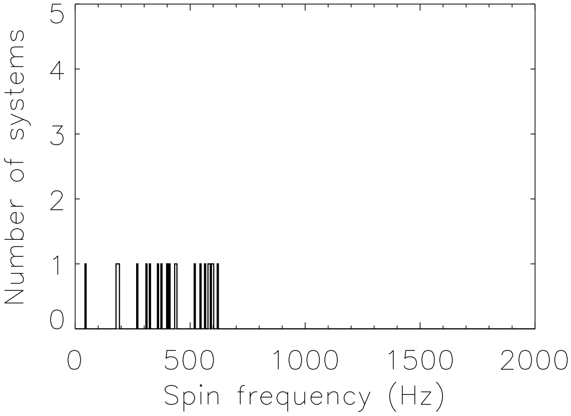


Figure 2. The spin frequency distribution of accreting millisecond X-ray pulsars. (From Chakrabarty 2008.)

significant." Even for a $1.4M_\odot$ star, 800 Hz is well below Ω_K for all but the stiffest candidate equations of state, and accreting pulsars are likely to have larger masses and still higher values of Ω_K.

A magnetic field of order 10^8 G can limit the spin of an accreting millisecond pulsar. Because matter within the magnetosphere corotates with the star, only matter that accretes from outside the magnetosphere can spin up the star, leading to an equilibrium period given approximately by (Ghosh & Lamb 1979)

$$P_{\rm eq} \sim \left(\frac{B}{10^{12}{\rm G}}\right)^{6/7} \left(\frac{\dot{M}}{10^{-9}M_\odot {\rm yr}^{-1}}\right)^{-3/7} . \tag{46}$$

Because this period depends on the magnetic field, a sharp cutoff in the frequency of accreting stars is not an obvious prediction of magnetically limited spins; and a cutoff at a rotation rate of order 700-800 Hz is not consistent with a range of magnetic field strengths presumed to extend below 10^8 G.

Under what circumstances the CFS instability could limit the spin of recycled pulsars has now been studied in a large number of papers. References to this work can be found in the treatment in (Friedman & Stergioulas 2011) on which the present review is based and in comprehensive earlier discussions by Stergioulas (2003), by Andersson and Kokkotas (2001), and by Kokkotas and Ruoff (2001, 2002) briefer reviews of more recent work are given in (Andersson *et al.* 2011; Owen 2010). References in the present review are generally limited to initial work and to a late paper that contains intervening references.

Whether the instability survives the complex physics of a real neutron star has been the focus of most recent work, but it remains an open question. Studies have focused on:

- Dissipation from bulk and shear viscosity and mutual friction in a superfluid interior;
- magnetic field wind-up;
- nonlinear evolution and the saturation amplitude; and
- the possiblity that a continuous spectrum replaces r-modes in relativistic stars.

We discuss these in turn and then summarize the implications for nascent, rapidly rotating stars and for old stars spun up by accretions.

Viscosity

When viscosity is included, the growth-time or damping time τ of an oscillation has the form

$$\frac{1}{\tau} = \frac{1}{\tau_{GR}} + \frac{1}{\tau_b} + \frac{1}{\tau_s}, \tag{47}$$

with τ_b and τ_s the damping times due to bulk and shear viscosity. Bulk viscosity is large at high temperatures, shear viscosity at low temperatures. This leaves a window of opportunity in which a star with large enough angular velocity can be unstable. The window for the $l = m = 2$ r-mode is shown in Fig. 3, for a representative computation of viscosity. The highest solid curves on left and right mark the critical angular velocity Ω_c above which the $l = m = 2$ r-mode is unstable. The curves on the left, show the effect of shear viscosity at low temperature, allowing instability when $\Omega < \Omega_K$ only for $T > 10^6$K; the curve on the right shows the corresponding effect of bulk viscosity, cutting off the instability at temperatures above about 4×10^{10}K.

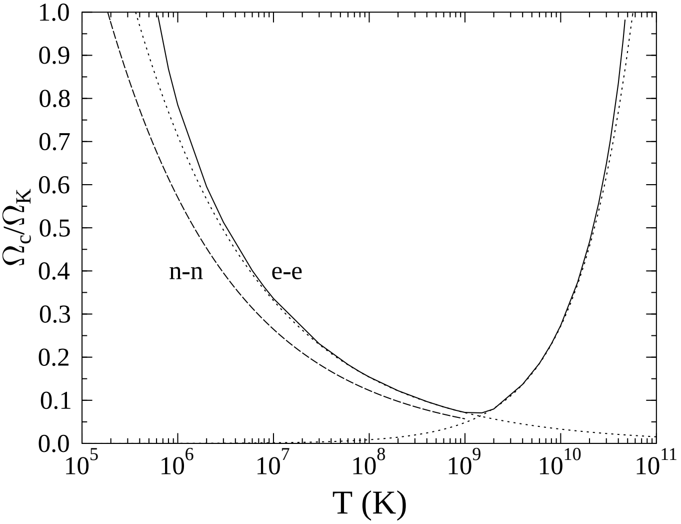


Figure 3. Critical angular velocity for the onset of the r-mode instability as a function of temperature (for a 1.5 $M_\odot$ neutron star model). The solid line corresponds to the $O(\Omega^2)$ result using electron-electron shear viscosity, and modified URCA bulk viscosity. The dashed line corresponds to the case of neutron-neutron shear viscosity. Dotted lines are $O(\Omega)$ approximations.

There is substantial uncertainty in the positions of both of these curves.

Bulk viscosity arises from nuclear reactions driven by the changing density of an oscillating fluid element, with neutrons decaying, $n \to p + e + \bar{\nu}_e$, as the fluid element expands and protons capturing electrons, $p + e \to n + \nu_e$, as it contracts. The neutrinos leave the star, draining energy from the mode. The rates of these *URCA* reactions increase rapidly with temperature and are fast enough to be important above about 10^9K, with an expected damping time τ_b given by

$$\frac{1}{\tau_{\rm b}} = \frac{1}{2E_c} \int \zeta(\delta\theta)^2 d^3x, \tag{48}$$

where $\theta = \nabla_\alpha u^\alpha$ is the divergence of the fluid velocity and the coefficient of bulk viscosity ζ is given by (Cutler, Lindblom & Splinter 1990)

$$\zeta = 6 \times 10^{25} \rho_{15}^2 T_9^6 \left(\frac{\omega_r}{1\text{Hz}}\right)^{-2} \ \text{g cm}^{-1}\ \text{s}^{-1}, \tag{49}$$

where $T_9 = T/(10^9\text{K})$. With these values, bulk viscosity kills the instability in all modes above a few times 10^{10}K (Ipser & Lindblom 1991 a, b; Yoshida & Eriguchi 1995).

These equations and Fig. 3 assume that only *modified URCA* reactions can occur, that the URCA reactions require a collision to conserve four-momentum, and this will be true when the proton fraction is less than about $1/9$. If the equation of state turns out to be unexpectedly soft (and the mass is large enough), direct URCA reactions would be allowed, suppressing the instability for uniformly rotating stars at roughly 10^9K (Zdunik 1996). A soft equation of state is also more likely to lead to stars with hyperons in their core with an additional set of nuclear reactions that dissipate energy and increase the bulk viscosity (Jones 2010; Lindblom & Owen 2002; Haensel, Levenfish & Yakovlev 2002; Nayyar & Owen 2006; Haskell & Andersson 2010) or quarks (Madsen 1998; Madsen 2000; Andersson *et al.*, 2002; Jaikumar *et al.* 2008; Rupak & Jaikumar 2010).

In contrast to bulk viscosity, shear viscosity increases as the temperature drops. In terms of the shear tensor $\sigma_{\alpha\beta} = (\delta^\gamma_\alpha + u_\alpha u^\gamma)(\delta^\delta_\beta + u_\beta u^\delta)(\nabla_\gamma u_\delta + \nabla_\delta u_\gamma - \frac{2}{3} g_{\gamma\delta}\nabla_\epsilon u^\epsilon)$, the damping time is given by

$$\frac{1}{\tau_{\rm s}} = \frac{1}{E_c}\int \eta\delta\sigma^{\alpha\beta}\delta\sigma_{\alpha\beta}\, d^3x, \tag{50}$$

where η is the coefficient of shear viscosity. For nascent neutron stars hotter than the superfluid transition temperature (about 10^9K), the neutron-neutron shear viscosity coefficient is (Flowers & Itoh 1976)

$$\eta_n = 2\times 10^{18}\rho_{15}^{9/4}T_9^{-2}\ \text{g cm}^{-1}\ \text{s}^{-1}, \tag{51}$$

where $\rho_{15} = \rho/(10^{15}\text{g cm}^{-3})$. Below the superfluid transition temperature, electron-electron scattering determines the shear viscosity in the superfluid core, giving (Cutler & Lindblom 1987)

$$\eta_e = 6\times 10^{18}\rho_{15}^{2}T_9^{-2}\ \text{g cm}^{-1}\ \text{s}^{-1}. \tag{52}$$

Shear viscosity may be greatly enhanced after formation of the crust in a boundary layer (Ekman layer) between crust and core (Ushomirsky & Bildsten 1998; Lindblom *et al.* 2000; Anderson *et al.* 2000; Glampedakis & Anderson 2006a, Glampedakis & Anderson 2006b). The enhancement depends on the extent to which the core participates in the oscillation, parametrized by the slippage at the boundary. The uncertainty in this slippage appears to be the greatest current uncertainty in dissipation of the mode by shear viscosity, and it significantly affects the critical angular velocity of the r-mode instability in accreting neutron stars.

For f-modes, the part of the instability window in Fig. 3 to the left of 10^9 K is thought to be removed by another dissipative mechanism that comes into play below the superfluid transition temperature. Called mutual friction, it arises from the scattering of electrons off magnetized neutron vortices. Work by Lindblom and Mendell (1995) shows that mutual friction in the superfluid core completely suppresses f- and p-mode instabilities below the transition temperature. For the r-mode instability, subsequent work by the same authors (2000) finds that the mutual friction is much smaller, with a damping time of order 10^4 s, too long to be important.

In a recent paper, Gaertig *et al.* point out the possibility of an interaction between vortices and quantized flux tubes that would result in a much smaller value for the mutual friction. They argue that the resulting uncertainty is great enough that shear viscosity could be the dominant dissipative mechanism for f-modes as well as r-modes.

Magnetic field windup

At second-order in the perturbation, the nonlinear evolution of an unstable mode includes an axisymmetric part that describes a growing differential rotation. Because differential rotation will wind up magnetic field lines, the mode's energy could be transferred to the star's magnetic field (Spruit 1999; Rezzolla *et al.* 2000; Rezzolla *et al.* 2001b; Rezzolla *et al.* 2001a; Cuofano & Drago 2010). Again there is large uncertainty about the strength of a toroidal magnetic field that will be generated by the differential rotation, what magnetic instabilities will arise, and what the effective dissipation will be. Apart from the studies cited here (all of which deal with r-modes) nearly all the remaining work on the evolution of unstable modes ignores magnetic fields.

Relativistic r-modes and a possible continuous spectrum

Relativistic r-modes have been computed by a number of authors (Kojima 1998; Kojima & Hosonuma 1999; Kojima & Hosonuma 2000; Lockitch, Andersson & Friedman 2001; Lockitch, Friedman & Andersson 2003; Lockitch, Andersson & Watts 2004; Andersson 1998; Ruoff & Kokkotas 2001; Ruoff & Kokkotas 2002; Ruoff, Stavridis & Kokkotas 2003; Kokkotas & Ruoff 2002; Yoshida & Lee 2002; Kastaun 2008). Where the Newtonian approximation has purely axial $l = m$ r-modes for barotropic stars at lowest order in Ω, in the full theory all rotationally restored modes include a polar part. The change in the structure of the computed r-modes are small, but that may not be the end of the story.

For non-barotropic stars Kojima found a single second-order eigenvalue equation for the frequency, to lowest nonvanishing order in Ω. The coefficient of the highest derivative term in that equation vanishes at some value of the radial coordinate r, for typical candidate neutron-star equations of state, and that singular behavior gives a continuous spectrum. Lockitch, Andersson & Watts (2004) consider the question of the continuous spectrum and the existence of r-modes in some detail. They argue that the singularity in the Kojima equation is an artifact of the slow-rotation approximation and is not present if one includes terms of order Ω^2. Their work is a strong argument for the existence of r-modes in non-barotropic models.

Showing the existence of the mode, however, does not decide the question of whether a continuous spectrum is also present or whether the existence of a continuous or nearly continuous spectrum significantly alters the evolution of an initial perturbation.

Nonlinear evolution

Linear perturbation theory is valid only for small-amplitude oscillations; as the amplitude of an unstable mode grows, couplings to other modes become increasingly important, and the mode ultimately reaches a saturation amplitude or is disrupted, losing coherence. The first nonlinear studies of the r-mode instability involved fully nonlinear 3+1 evolutions in which the r-mode was set at a large initial amplitude (Stergioulas & Font 2001) or was driven to large amplitude by an artificially large gravitational-radiation reaction term (Lindblom, Tohline & Vallisneri 2001, Lindblom, Tohline & Vallisneri 2002). On a few tens of dynamical timescales, saturation was seen only at an amplitude of order unity. Subsequently, simulations on longer timescales showed a coupling to daughter modes (Gressman *et al.* 2002; Lin & Suen 2006), suggesting that the actual saturation amplitude of the r-mode is smaller than the amplitude at which gravitational-radiation reaction was switched off in the short-timescale simulations.

The resolution of 3+1 simulations, however, is too low to see couplings to short-wavelength modes, and they cannot run for a time long enough to see the growth from a realistic radiation-reaction term. The alternative is to examine the nonlinear evolution in the context of higher-order perturbation theory. To do this, the Cornell group (initially with S. Morsink) (Arras *et al.* 2003; Schenk *et al.* 2002; Morsink 2002) constructed a second-order perturbation theory for rotating Newtonian stars, and then used the formalism to study the nonlinear evolution of an unstable r-mode. Their series of papers leaves little doubt that nonlinear couplings sharply limit the amplitude of an unstable r-mode, with a possible range of 10^{-1}–10^{-5} (see (Bondarescu, Teukolsky & Wasserman 2007) and references therein).

The nonlinear development of the f-mode instability has been modeled in three-dimensional, hydrodynamical simulations (in a Newtonian framework) by Ou, Tohline & Lindblom (2004) and by Shibata & Karino (2004), essentially confirming previous approximate results obtained in (Lai & Shapiro 1995). Kastaun *et al.* (2010) report an initial nonlinear study of f-modes in general relativity. In the framework of a 3+1 simulation in a Cowling approximation (a fixed background metric of the unperturbed rotating star), they find limits on the amplitude of less than 0.1, set by wave-breaking and by coupling to inertial modes. This can be regarded as an upper limit on the amplitude, with second-order perturbative computations still to be done.

Instability scenarios in nascent neutron stars and in old accreting stars

Both r-modes and f-modes may be unstable in nascent neutron stars that are rapidly rotating at birth. Recent work on f-modes in relativistic models (Gaertig *et al.*; Gaertig & Kokkotas 2010) finds growth times substantially shorter than previously computed Newtonian values. In particular, the $l = m = 3$ and $l = m = 4$ f-modes have growth times of 10^3-10^5 s for Ω near Ω_K. In a typical scenario, a star with rotation near the Kepler limit becomes unstable within a minute of formation, when the temperature has dropped below 10^{11}K. As the temperature drops further, the instability grows to saturation amplitude in days or weeks. Loss of angular momentum to gravitational waves spins down the star until the critical angular velocity is reached below which the star is stable, at or before the time at which the core becomes a superfluid. The $l = m = 3$ mode could be a source of observable gravitational waves for supernovae in or near the Galaxy.

The time over which the instability is active depends on the saturation amplitude, the cooling rate, and the superfluid transition temperature, and all of these have large uncertainties. The time at which a superfluid transition occurs could be shorter than a year, but recent analyses of the cooling of a neutron star in Cassiopeia A (Page *et al.* 2011; Shternin *et al.* 2011) suggest a superfluid transition time for that star of order 100 years.

The scenario for the $l = m = 2$ r-mode instability of a nascent star is similar. The r-mode instability itself was pointed out by Andersson (1998), with a mode-independent proof for relativistic stars given by Friedman and Morsink (1998). First computations of the growth and evolution were reported by Lindblom *et al.* (1998) and Andersson *et al.* (1999), with effects of a crust discussed in Lindblom *et al.* (2000). Intervening work is referred to in a recent paper by Bondarescu *et al.* (2008); the simulations reported by Bondarescu *et al.* include nonlinear couplings that saturate the amplitude and the alternative possibilities for viscosity that we have discussed above. The r-mode's saturation amplitude is likely to be lower than that of the f-modes, and it is likely to persist longer because of its low mutual friction.

As mentioned above, the r-mode instability of neutron stars spun up by accretion has been more intensively studied in connection with the observed spins of LMXBs. Papaloizou & Pringle (1978) suggested the possibility of accretion spinning up a star until it becomes unstable to the emission of gravitational waves and reaches a steady state, with the angular momentum gained by accretion equal to the angular momentum lost to gravitational waves. Following the discovery of the first millisecond pulsar, Wagoner examined the mechanism in detail for CFS unstable f-modes (Wagoner 2002). Although mutual friction appears to rule out the steady-state picture for f-modes, it remains a possibility for r-modes (Bildsten 1998; Andersson *et al.* 1999; Andersson *et al.* 2000; Wagoner 2002). Levin (1999) and (independently) Spruit (1999), however, pointed out that viscous heating of the neutron star by its unstable oscillations will lower the shear viscosity and so increase the mode's growth rate, leading to a runaway instability. The resulting scenario is a cycle in which a cold, stable neutron star is spun up over a few million years until it becomes unstable; the star then heats up, the instability grows, and the star spins down until it is again stable, all within a few months; the star then cools, and the cycle repeats.

This scenario would rule out r-modes in LMXBs as a source of detectable gravitational waves because the stars would radiate for only a small fraction of the cycle. A small saturation amplitude, however, lengthens the time spent in the cycle, possibly allowing observability (Heyl 2002). The steady state itself remains a possible alternative in stars whose core contains hyperons or free quarks (or if the "neutron stars" are really strange quark stars) (Andersson *et al.* 2002; Lindblom & Owen 2002; Wagoner 2002; Reisenegger & Bonacić 2003; Nayyar & Owen 2006; Haskell & Andersson 2010). Heating the core increases the bulk viscosity, and with an exotic core, this growth in the bulk viscosity is large enough to prevent the thermal runaway and allow a steady state. Recent work by Bondarescu *et al.* (2007) constructs nonlinear evolutions (restricted to 3 coupled modes) that include neutrino cooling, shear viscosity, hyperon bulk viscosity and dissipation at the core-crust boundary layer, with parameters to span a range of uncertainty in these various quantities. They display the regions of parameter space associated with the alternative scenarios just outlined – steady state, cycle, and fast and slow runaways. In all cases, the r-mode amplitude remains very small ($\sim 10^{-5}$), but because of the long duration of the instability, such systems are still good candidates for gravitational wave detection by advanced LIGO class interferometers (Bondarescu *et al.* 2007; Watts & Krishnan 2009; Owen 2010).

Dynamical nonaxisymmetric instability

Work on dynamical nonaxisymmetric instabilities is largely outside the scope of this review. They are most likely to be relevant to protoneutron stars and to the short-lived hypermassive neutron stars that form in the merger of a double neutron star system. Unless the star has unusually high differential rotation, instability requires a large value of the ratio $T/|W|$ of rotational kinetic energy to gravitational binding energy: comparable to the value $T/|W| = 0.27$ that marks the dynamical instability of the $l = m = 2$ mode of uniformly rotating uniform density Newtonian models (the Maclauring spheroids). This bar instability, if present, will emit strong gravitational waves with frequencies in the kHz regime. The development of the instability and the resulting waveform have been computed numerically in the context of both Newtonian gravity and in full general relativity (see (Houser *et al.* 1994; Tohline *et al.* 1985; Shibata *et al.* 2000; Manca *et al.* 2007) for representative studies).

Uniformly rotating neutron stars have maximum values of $T/|W|$ smaller than 0.14, apparently precluding dynamical nonaxisymmetric instability. For highly differential rotation, however, Centrella *et al.* (2001) found a one-armed ($m = 1$) instability for smaller rotation, for $T/|W| \sim 0.14$, but for a polytropic index of $N = 3$ which is not representative for neutron stars. Remarkably, Shibata *et al.* (2002, 2003) then found an $m = 2$ instability for $T/|W|$ as low as 0.01, for models with polytropic index $N = 1$, representing a stiffness appropriate to neutron stars. These instabilities appear to be related to the existence of corotation points, where the pattern speed of the mode matches the star's angular velocity (Watts, Anderson & Jones 2005; Saijo & Yoshida 2006); Ou and Tohline tie the growth of the instability to a resonant cavity associated with a minimum in the vorticity to density ratio (the so-called vortensity) (Ou & Tohline 2006). Collapsing cores in supernovae are differentially rotating, and these instabilities of proto-neutron stars arise in simulations of rotating core collapse (Ott *et al.* 2005; Ott 2009). Because they can radiate more energy in gravitational waves than the post-bounce burst signal itself, interest in these dynamical instabilities is strong.

Acknowledgments

J.F.'s work is supported in part by NSF Grants PHY 0503366 and PHY 1001515.

References

Abramowicz M. A., 2004 Rayleigh & Solberg criteria reversal near black holes, the optical geometry explanation. ArXiv Astrophysics e-prints, November 2004

Andersson N., 1998, ApJ, 502, 708

Andersson N., Ferrari V., Jones D. I., Kokkotas K. D., Krishnan B., Read J. S., Rezzolla L., Zink B., 2011, General Relativity and Gravitation, 43, 409

Andersson N., Jones D. I., Kokkotas K. D., 2002, MNRAS, 337, 1224

Andersson N., Jones D. I., Kokkotas K. D., Stergioulas N., 2000, ApJ, 534, L75

Andersson N., Kokkotas K. D., 2001, Intl. J. Modern Phys. D, 10, 381

Andersson N., Kokkotas K. D., Schutz B. F., 1999, ApJ, 510, 846

Andersson N., Kokkotas, K. D., Stergioulas N., 1999, ApJ, 516, 307

Arras P., Flanagan É. É., Morsink S. M., Schenk A. K., Teukolsky S. A., Wasserman I., 2003, ApJ, 591, 1129

Bardeen J. M., 1970, ApJ, 162, 71

Bildsten L., 1998, ApJ, 501, L89

Bondarescu R., Teukolsky S. A., Wasserman I., 2007, Phys. Rev. D, 76, 064019

Bondarescu R., Teukolsky S. A., Wasserman I., 2008, eprint arXiv, 0809.3448v2

Calkin M. G., 1963, Can. J. Phys., 41, 2241

Carter B., 1973, Comm. Math. Phys., 30, 261

Centrella J. M., New K. C. B., Lowe L. L., Brown J. D., 2001, ApJ, 550, L193

Chakrabarty D., 2008, The spin distribution of millisecond X-ray pulsars. In R. Wijnands *et al.*, eds, A decade of accreting millisecond x-ray pulsars, volume 1068 of AIP Conference Proceedings, p. 67

Chandrasekhar S., 1964, ApJ, 140, 417

Chandrasekhar S., 1964, ApJ, 139, 664

Chandrasekhar S., 1970, Phys. Rev. Lett., 24, 611

Chandrasekhar S., Friedman J. L., 1972, ApJ, 175, 379

Chandrasekhar S., Friedman J. L., 1972, ApJ, 176, 745

Clement M. J., 1964, ApJ, 140, 1045

Cook G. B., Shapiro S. L., Teukolsky S. A., 1992, ApJ, 398, 203

Cuofano C., Drago A., 2010, Phys. Rev. D, 82, 084027

Cutler C., Lindblom L., 1987, ApJ, 314, 234

Cutler C., Lindblom L., Splinter R. J., 1990, ApJ, 363, 603

Detweiler S. L., Ipser J. R., 1973, ApJ, 185, 685

Flowers E., Itoh N., 1976, ApJ, 206, 218

Fowler W. A., 1966, ApJ, 144, 180

Friedman J. L., 1978, Communications in Mathematical Physics, 62, 247

Friedman J. L., Ipser J. R., Sorkin R. D., 1988, ApJ, 325, 722

Friedman J. L., Schutz B. F., 1975, ApJ, 200, 204

Friedman J. L., Schutz B. F., 1978, ApJ, 221, 937

Friedman J. L., Schutz B. F., 1978, ApJ, 222, 281

Friedman J.L., Morsink S.M., 1998, ApJ., 502, 714

Friedman J. L., Stergioulas N., 2011, Rotating Relativistic Stars. Cambridge University Press, Cambridge

Gaertig, E., Glampedakis K., Kokkotas K. D., Zink B., 2011. The f-mode instability in relativistic neutron stars. eprint arXiv

Gaertig E., Kokkotas K. D., 2011 Gravitational wave asteroseismology with fast rotating neutron stars. eprint arXiv, 1005.5228

Ghosh P., Lamb F. K., 1979, ApJ, 232, 259

Glampedakis K., Andersson N., 2006a, Phys. Rev. D, 74, 044040

Glampedakis K., Andersson N., 2006b, MNRAS, 371, 1311

Gressman, P., Lin L.-M., Suen W.-M., Stergioulas N., Friedman J. L., 2002, Nonlinear r-modes in neutron stars, instability of an unstable mode

Haensel P., Levenfish K. P., Yakovlev D. G., 2002, A&A, 381, 1080

Harrison B. K., Thorne K. S., Wakano M., Wheeler J. A., 1965, Gravitation Theory & Gravitational Collapse. University of Chicago Press, Chicago

Haskell B., Andersson N., 2010, MNRAS, 408, 1897

Heyl J., 2002, ApJ, 574, L57

Houser J. L., Centrella J. M., Smith S. C., 1994, Phys. Rev. Lett., 72, 1314

Ipser J. R., Lindblom L., 1991, ApJ, 373, 213

Ipser J. R., Lindblom L., 1991, ApJ, 379, 285

Jaikumar P., Rupak G., Steiner A. W., 2008, Phys. Rev. D, 78, 123007

Jones P. B., 2010, Phys. Rev. Lett., 86, 1384

Kastaun W., 2008, Phys. Rev. D, 77, 124019

Kastaun W., Willburger B., Kokkotas K. D., 2010, Phys. Rev. D, 82, 104036

Kojima Y., 1998, MNRAS, 293, 49
Kojima Y., Hosonuma M., 1999, ApJ, 520, 788
Kojima Y., Hosonuma M., 2000, Phys. Rev. D, 62, 044006
Kokkotas K. D., Ruoff J., 2001, Instabilities of relativistic stars. In 2001, A relativistic spacetime Odyssey, 25th Johns Hopkins Workshop, 2002. Firenze 2001
Kovetz A., 1967, Zeitschrift fur Astrophysik, 66, 446
Lai D., Shapiro S. L., 1995, ApJ, 442, 259
Lebovitz N. R., 1965, ApJ, 142, 229
Levin Y., 1999, ApJ, 517, 328
Lin L.-M., Suen W.-M., 2006, MNRAS, 370, 1295
Lindblom L., Mendell G., 1995, ApJ, 444, 804
Lindblom L., Mendell G., 2000, Phys. Rev. D, 61, 104003
Lindblom L., Owen B. J., 2002, Phys. Rev. D., 65, 063006
Lindblom L., Owen B. J., Ushomirsky G., 2000, Phys. Rev. D, 62, 084030
Lindblom L., Owen B.J., Morsink S. M., 1998, Phys. Rev. Lett., 80, 4843
Lindblom L., Tohline J. E., Vallisneri M., 2001, Phys. Rev. Lett., 86, 1152
Lindblom L., Tohline J. E., Vallisneri M., 2002, Phys. Rev. D., 65, 084039
Lockitch K. H., Andersson N., Friedman J. L., 2001, Phys.Rev. D, 63, 024019
Lockitch K. H., Andersson N., Watts A. L., 2004, Class. Quant. Grav, 21, 4661
Lockitch K. H., Friedman J. L., Andersson N., 2003, Phys. Rev. D, 68, 124010
Lynden-Bell D., Ostriker J. P., 1967, MNRAS, 136, 293
Madsen J., 1998, Phys. Rev. Lett., 81, 3311
Madsen J., 2000, Phys. Rev. Lett., 85, 10
Manca G. M., Baiotti L., De Pietri R., Rezzolla L., 2007, Class. Quant. Grav, 24, S171
S. M. Morsink, 2002, ApJ, 571, 435
Nayyar M., Owen B. J., 2006, Phys. Rev. D, 73, 084001
Oppenheimer J. R., Volkoff G. M., 1939, Phys. Rev., 55, 374
Ott C. D., 2009, Class. Quant. Grav, 26, 063001
Ott C. D., Ou S., Tohline J. E., Burrows A., 2005, ApJ, 625, L119
Ou S., Tohline J. E., 2006, ApJ, 651, 1068
Ou S., Tohline J. E., Lindblom L., 2004, ApJ, 617, 490
Owen B. J., 2010, Phys. Rev. D, 82(10), 104002
Owen B. J., 2010, Phys. Rev. D, 82, 104002
Page D., Prakash M., Lattimer J. M., Steiner A. W., 2011, Phys. Rev. Lett., 106(8), 081101
Papaloizou J., Pringle J. E., 1978, MNRAS, 184, 501
Poincaré H., 1885, Acta Math., 7, 259
Read J. S., Markakis C., Shibata M., Uryū K., Creighton J. D. E., Friedman J. L., 2009, Phys. Rev. D, 79, 124033
Reisenegger A., Bonacić A., 2003, Phys. Rev. Lett., 91, 201103
Rezzolla L., Lamb F. K., Marković, D., Shapiro S. L., 2001a, Phys. Rev. D, 64, 104014
Rezzolla, L., Lamb F. K., Marković, D., Shapiro S. L., 2001b, Phys. Rev. D, 64, 104013
Rezzolla L., Lamb F. K., Shapiro S. L., 2000, ApJ, 531, L142
Rezzolla L., Takami K., Yoshida S., 2011, private communication.
Ruoff J., Kokkotas K. D., 2001, MNRAS, 328, 678
Ruoff J., Kokkotas K. D., 2002, MNRAS, 330, 1027
Ruoff J., Stavridis A., Kokkotas K. D., 2003, MNRAS, 339, 1170
Rupak G., Jaikumar P., 2010, Phys. Rev. C, 82, 055806
Shin'ichirou Saijo, Yoshida M., 2006, MNRAS, 368, 1429

Schenk A. K., Arras P., Flanagan É. É., Teukolsky S. A., Wasserman I., 2002, Phys. Rev. D, 65, 024001

Schutz B. F. Jr., 1970, ApJ, 161, 1173

Seguin F. H., 1975, ApJ, 197, 745

Shibata M., Baumgarte T. W., Shapiro S. L., 2000, ApJ, 542, 453

Shibata M., Karino S., 2004, Phys. Rev. D, 70, 084022

Shibata M., Karino S., Eriguchi Y., 2002, MNRAS, 334, L27

Shibata M., Karino S., Eriguchi Y., 2003, MNRAS, 343, 619

Shternin P. S., Yakovlev D. G., Heinke C. O., Ho W. C. G., Patnaude D. J., 2011, MNRAS, 412, L108

Solberg H., 1936, Proces Verbaux de I 1' Association de Météorologie, International Union of Geodesy and Geophysics, 6th General Assembly (Edinburg), 2, 66

Sorkin R. D., 1981, ApJ, 249, 254

Sorkin R. D., 1982, ApJ, 257, 847

Spruit H. C., 1999, A&A, 341, L1

Stergioulas N., 2003, Living Reviews in Relativity, 6, 3

Stergioulas N., Font J. A., 2001, Phys. Rev. Lett., 86, 1148

Stergioulas N., J. L. Friedman, 1998, ApJ, 492, 301

Taub A. H., 1954, Phys. Rev., 94, 1468

Taub A. H., 1969, Comm. Math. Phys., 15, 235

Thorne K. S., 1966, ApJ, 144, 201

Thorne K. S., 1967, in Relativistic stellar structure & dynamics. In E. Schatzman, P. Véron C. DeWitt, editor, High Energy Astrophysics, volume III of Les Houches Summer School of Theoretical Physics, p. 259, Gordon & Breach

Thorne K. S., 1978, General-relativistic astrophysics. In N. R. Lebovitz, ed., Theoretical Principles in Astrophysics & Relativity, p. 149

Thorne K. S., 1980, Rev. Mod. Phys., 52, 299

Tohline J. E., Durisen R. H., McCollough M., 1985, ApJ, 298, 220

Ushomirsky G., Bildsten L., 1998, ApJ Lett., 497, L101

Wagoner R. V., 1984, ApJ, 278, 345

Wagoner R. V., 2002, ApJ, 578, L63

Watts A. L., Andersson N., Jones D. I., 2005, ApJ, 618, L37

Watts A. L., Krishnan B., 2009, Adv. Space Res., 43, 1049

Yoshida S., Eriguchi Y., 1995, ApJ, 438, 830

Yoshida S., Eriguchi Y., 1997, ApJ, 490, 779

Yoshida S., Eriguchi Y., 1999, ApJ, 515, 414

Yoshida S., U. Lee. ApJ., 567, 1112-1120, 2002

Zdunik J. L., 1996, A&A, 308, 828

Zel'dovich Ya. B., Novikov I. D., 1971, Relativistic Astrophysics, Volume 1. University of Chicago Press, Chicago

Zink B., Korobkin O., Schnetter E., Stergioulas N., 2010, Phys. Rev. D, 81(8), 084055

Bull. Astr. Soc. India (2011)

Key problems in black hole physics today

Pankaj S. Joshi[*]

Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India

Received April 4, 2011; accepted April 14, 2011

Abstract. We review here some of the major open issues and challenges in black hole physics today, and the current progress on the same. It is pointed out that to secure a concrete foundation for the basic theory as well as astrophysical applications for black hole physics, it is essential to gain a suitable insight into these questions. In particular, we discuss the recent results investigating the final fate of a massive star within the framework of the Einstein gravity, and the stability and genericity aspects of the gravitational collapse outcomes in terms of black holes and naked singularities. Recent developments such as spinning up a black hole by throwing matter into it, and physical effects near naked singularities are considered. It is pointed out that some of the new results obtained in recent years in the theory of gravitational collapse imply interesting possibilities and understanding for the theoretical advances in gravity as well as towards new astrophysical applications.

Keywords : black hole physics – gravitation

1. Introduction

The fundamental question of the final fate of a massive star, when it exhausts its internal nuclear fuel and collapses continually under the force of its own gravity, was highlighted by Chandrasekhar way back in 1934 (Chandrasekhar 1934), who pointed out:

"Finally, it is necessary to emphasize one major result of the whole investigation, namely, that the life-history of a star of small mass must be essentially different from the life-history of a star of large mass. For a star of small mass the natural white-dwarf stage is an initial step towards complete extinction. A star of large mass ($> M_c$) cannot pass into the white-dwarf stage, and one is left speculating on other possibilities."

We can see the seeds of modern black hole physics already present in the inquiry made above on the final fate of massive stars. The issue of endstate of large mass stars has, however, remained unresolved and elusive for a long time of many decades after that. In fact, a review of the status of the subject many decades later notes, "Any stellar core with a mass exceeding the upper limit that undergoes gravitational collapse must collapse to

[*] e-mail: psj@tifr.res.in

indefinitely high central density... to form a (spacetime) singularity" (Report of the Physics Survey Committee 1986).

The reference above is to the prediction by general relativity, that under reasonable physical conditions, the gravitationally collapsing massive star must terminate in a spacetime singularity (Hawking & Ellis 1973). The densities, spacetime curvatures, and all physical quantities must typically go to arbitrarily large values close to such a singularity. The above theoretical result on the existence of singularities is, however, of a rather general nature, and provides no information on the nature and structure of such singularities. In particular, it gives us no information as to whether such singularities, when they form, will be covered in horizons of gravity and hidden from us, or alternatively these could be visible to external observers in the Universe.

One of the key questions in black hole physics today therefore is, are such singularities resulting from collapse, which are super-ultra-dense regions forming in spacetime, visible to external observers in the Universe? This is one of the most important unresolved issues in gravitation theory currently. Theorists generally believed that in such circumstances, a black hole will always form covering the singularity, which will then be always hidden from external observers. Such a black hole is a region of spacetime from which no light or particles can escape. The assumption that spacetime singularities resulting from collapse would be always covered by black holes is called the Cosmic Censorship Conjecture (CCC; Penrose 1969). As of today, we do not have any proof or any specific mathematical formulation of the CCC available within the framework of gravitation theory.

If the singularities were always covered in horizons and if CCC were true, that would provide a much needed basis for the theory and astrophysical applications of black holes. On the other hand, if the spacetime singularities which result from a continual collapse of a massive star were visible to external observers in the Universe, we would then have the opportunity to observe and investigate the super-ultra-dense regions in the Universe, which form due to gravitational collapse and where extreme high energy physics and also quantum gravity effects will be at work.

My purpose here is to review the above and some of the related key issues in gravitation theory and black hole physics today. This will be of course from a perspective of what I think are the important problems, and no claim to completeness is made. In Section 2, we point out that in view of the lack of any theoretical progress on CCC, the important way to make any progress on this problem is to make a detailed and extensive study of gravitational collapse in general relativity. Some recent progress in this direction is summarized. While we now seem to have a good understanding of the black hole and naked singularity formations as final fate of collapse in many gravitational collapse models, the key point now is to understand the genericity and stability of these outcomes, as viewed in a suitable framework. Section 3 discusses these issues in some detail. Recent developments on throwing matter into a black hole and the effect it may have on its horizon are pointed out in Section 4, and certain quantum aspects are also discussed. The issue of predictability or its breakdown in gravitational collapse is discussed in Section 5. We conclude by giving a brief idea of the future outlook and possibilities in the final section.

2. What is the final fate of a massive star?

While Chandra's work pointed out the stable configuration limit for the formation of a white dwarf, the issue of the final fate of a star which is much more massive (e.g. tens of solar masses) remains very much open even today. Such a star cannot settle either as a white dwarf or as a neutron star.

The issue is clearly important both in high energy astrophysics and in cosmology. For example, our observations today on the existence of dark energy in the Universe and its acceleration are intimately connected to the observations of Type Ia supernovae in the Universe. The observational evidence coming from these supernovae, which are exploding stars in the faraway Universe, tells us on how the Universe may be accelerating away and the rate at which such an acceleration is taking place. While Type Ia supernovae result from the explosion of a white dwarf star, at the heart of a Type II supernova underlies the phenomenon of a catastrophic gravitational collapse of a massive star, wherein a powerful shock wave is generated, blowing off the outer layers of the star.

If such a star is able to throw away enough of matter in such an explosion, it might eventually settle as a neutron star. But otherwise, or if further matter is accreted onto the neutron star, there will be a further continual collapse, and we shall have to then explore and investigate the question of the final fate of such a massive collapsing star. But other stars, which are more massive and well above the normal supernova mass limits must straightaway enter a continual collapse mode at the end of their life cycle, without an intermediate neutron star stage. The final fate of the star in this case must be decided by general relativity alone.

The point here is, more massive stars which are tens of times the mass of the Sun burn much faster and are far more luminous. Such stars then cannot survive more than about ten to twenty million years, which is a much shorter life span compared to stars like the Sun, which live billions of years. Therefore, the question of the final fate of such short-lived massive stars is of central importance in astronomy and astrophysics.

What happens then, in terms of the final outcome, when such a massive star dies after exhausting its internal nuclear fuel? As we indicated above, the general theory of relativity predicts that the collapsing massive star must terminate in a spacetime singularity, where the matter energy densities, spacetime curvatures and other physical quantities blow up. It then becomes crucial to know whether such super-ultra-dense regions, forming in stellar collapse, are visible to an external observer in the Universe, or whether they will be always hidden within a black hole and an event horizon that could form as the star collapses. This is one of the most important issues in black hole physics today.

The issue has to be probed necessarily within the framework of a suitable theory of gravity, because the strong gravity effects will be necessarily important in such a scenario. This was done for the first time in the late 1930s, by the works of Oppenheimer and Snyder, and Datt (Oppenheimer & Snyder 1939; Datt 1938). They used the general theory of relativity to examine the final fate of an idealized massive matter cloud, which was taken to be a spatially homogeneous ball which had no rotation or internal pressure, and was assumed to be spherically symmetric. The dynamical collapse studied here resulted in the formation of a spacetime singularity, which was preceded by the development of an event horizon, which created a black hole in the spacetime. The singularity was hidden

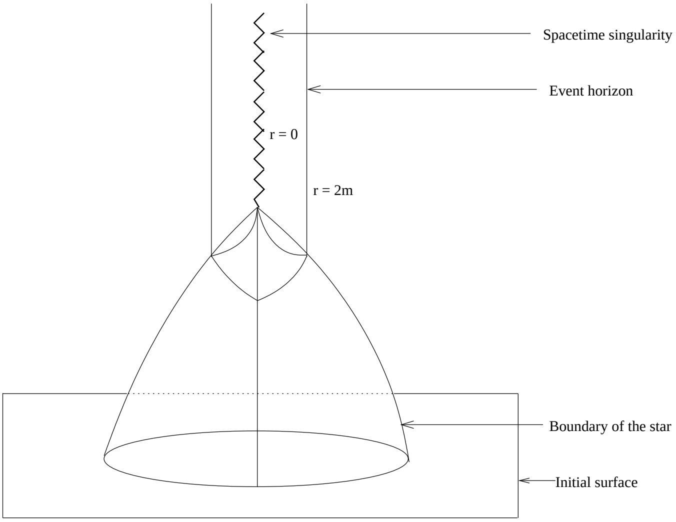


Figure 1. Dynamical evolution of a homogeneous spherical dust cloud collapse, as described by the Oppenheimer-Snyder-Datt solution.

inside such a black hole, and the collapse eventually settled into a final state which was the Schwarzschild geometry (see Fig. 1).

There was, however, not much attention paid to this model at that time, and it was widely thought by gravitation theorists as well as astronomers that it would be absurd for a star to reach such a final ultra-dense state of its evolution. It was in fact only as late as 1960s, that a resurgence of interest took place in the topic, when important observational developments in astronomy and astrophysics revealed several very high energy phenomena in the Universe, such as quasars and radio galaxies, where no known physics was able to explain the observations of such extremely high energy phenomena in the cosmos. Attention was drawn then to dynamical gravitational collapse and its final fate, and in fact the term ‘black hole’ was also popularized just around the same time in 1969, by John Wheeler.

The CCC also came into existence in 1969. It suggested and assumed that what happens in the Oppenheimer-Snyder-Datt (OSD) picture of gravitational collapse, as discussed above, would be the generic final fate of a realistic collapsing massive star in general. In other words, it was assumed that the collapse of a realistic massive star will terminate in a black hole, which hides the singularity, and thus no visible or naked singularities will develop in gravitational collapse. Many important developments then took place in black hole physics which started in earnest, and several important theoretical aspects as well as astrophysical applications of black holes started developing. The classical as well as quantum aspects of black holes were then explored and interesting thermodynamic analogies for black holes were also developed. Many astrophysical applications for the real Universe then started developing for black holes, such as making models using black holes for phenomena such as jets from the centres of galaxies and the extremely energetic gamma rays bursts.

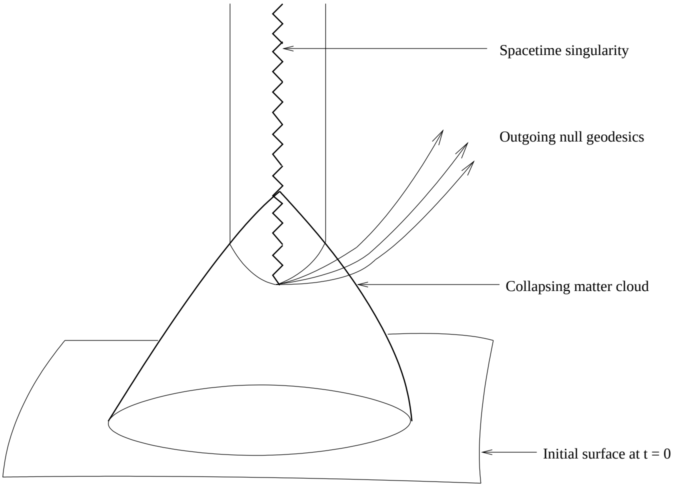


Figure 2. A spacetime singularity of gravitational collapse which is visible to external observers in the Universe, in violation to the cosmic censorship conjecture.

The key issue raised by the CCC, however, still remained very much open, namely whether a real star will necessarily go the OSD way for its final state of collapse, and whether the final singularity will be always necessarily covered within an event horizon. This is because real stars are inhomogeneous, have internal pressure forces and so on, as opposed to the idealized OSD assumptions. This remains an unanswered question, which is one of the most important issues in gravitation physics and black hole physics today. A spacetime singularity that is visible to faraway observers in the Universe is called a naked singularity (see Fig. 2). The point here is, while general relativity predicts the existence of singularity as the endstate for collapse, it gives no information at all on the nature or structure of such singularities, and whether they will be covered by event horizons, or would be visible to external observers in the Universe.

There is no proof, or even any mathematically rigorous statement available for CCC after many decades of serious effort. What is really needed to resolve the issue is gravitational collapse models for a realistic collapse configuration, with inhomogeneities and pressures included. The effects need to be worked out and studied in detail within the framework of Einstein gravity. Only such considerations will allow us to determine the final fate of collapse in terms of either a black hole or a naked singularity final state.

Over the past couple of decades, many such collapse models have been worked out and studied in detail. The generic conclusion that has emerged out of these studies is that both the black holes and naked singularity final states do develop as collapse endstates, for a realistic gravitational collapse that incorporates inhomogeneities as well as non-zero pressures within the interior of the collapsing matter cloud. Subject to various regularity and energy conditions to ensure the physical reasonableness of the model, it is the initial data, in terms of the initial density, pressures, and velocity profiles for the collapsing shells, that determine the final fate of collapse as either a naked singularity or a black hole (for further detail and references see e.g. Joshi 2008).

3. The genericity and stability of collapse outcomes

While general relativity may predict the existence of both black holes and naked singularities as collapse outcome, an important question then is, how would a realistic continual gravitational collapse of a massive star in nature end up. Thus the key issue under active debate now is the following: Even if naked singularities did develop as collapse end states, would they be generic or stable in some suitably well-defined sense, as permitted by the gravitation theory? The point here is, if naked singularity formation in collapse was necessarily 'non-generic' in some appropriately well-defined sense, then for all practical purposes, a realistic physical collapse in nature might always end up in a black hole, whenever a massive star ended its life.

In fact, such a genericity requirement has been always discussed and desired for any possible mathematical formulation for CCC right from its inception. However, the main difficulty here has again been that, there is no well-defined or precise notion of genericity available in gravitation theory and the general theory of relativity. Again, it is only various gravitational collapse studies that can provide us with more insight into this genericity aspect also.

A result that is relevant here is the following (Joshi & Dwivedi, 1999; Goswami & Joshi, 2007). For a spherical gravitational collapse of a rather general (type I) matter field, satisfying the energy and regularity conditions, given any regular density and pressure profiles at the initial epoch, there always exist classes of velocity profiles for the collapsing shells and dynamical evolutions as determined by the Einstein equations, that, depending on the choice made, take the collapse to either a black hole or naked singularity final state (see e.g. Fig. 3 for a schematic illustration of such a scenario).

Such a distribution of final states of collapse in terms of the black holes and naked singularities can be seen much more transparently when we consider a general inhomogeneous dust collapse, for example, as discussed by Mena, Tavakol & Joshi (2000) (see Fig. 4).

What determines fully the final fate of collapse here are the initial density and velocity profiles for the collapsing shells of matter. One can see here clearly how the different choices of these profiles for the collapsing cloud distinguish between the two final states of collapse, and how each of the black hole and naked singularity states appears to be 'generic' in terms of their being distributed in the space of final states. Typically, the result we have here is, given any regular initial density profile for the collapsing dust cloud, there are velocity profiles that take the collapse to a black hole final state, and there are other velocity profiles that take it to naked singularity final state. In other words, the overall available velocity profiles are divided into two distinct classes, namely the ones which take the given density profile into black holes, and the other ones that take the collapse evolution to a naked singularity final state. The same holds conversely also, namely if we choose a specific velocity profile, then the overall density profile space is divided into two segments, one taking the collapse to black hole final states and the other taking it to naked singularity

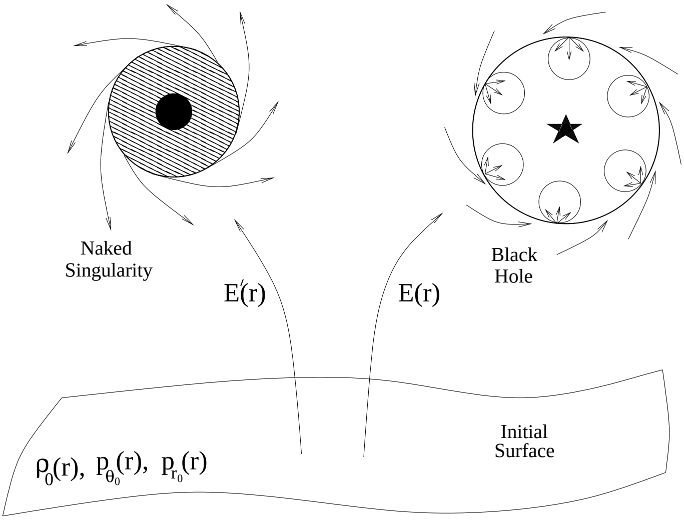


Figure 3. Evolution of spherical collapse for a general matter field with inhomogeneities and non-zero pressures included.

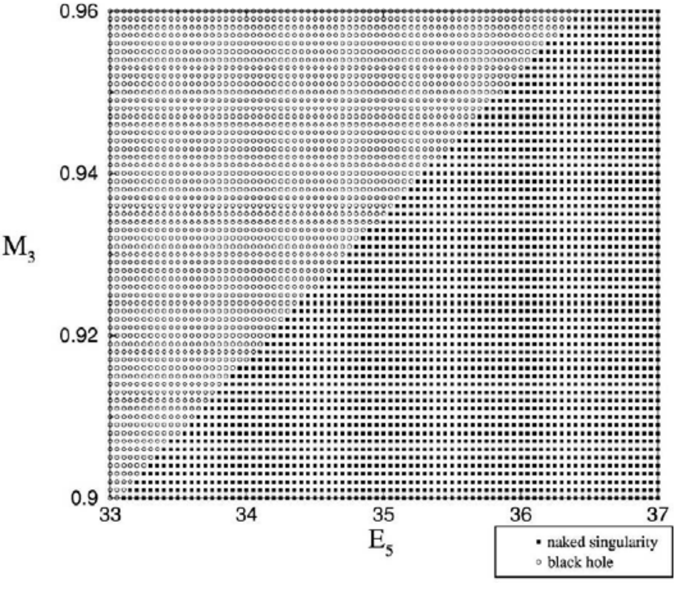


Figure 4. Collapse final states for inhomogeneous dust in terms of initial mass and velocity profiles for the collapsing shells.

final states. The clarity of results here gives us much understanding on the final fate of a collapsing matter cloud.

Typically, all stars have a higher density at the centre, which slowly decreases as one moves away. So it is very useful to incorporate inhomogeneity into dynamical collapse considerations. However, much more interesting is the collapse with non-zero pressures which are very important physical forces within a collapsing star. We briefly consider below a typical scenario of collapse with a non-zero pressure component, and for further details we refer to Joshi & Malafarina (2011).

For a possible insight into genericity of naked singularity formation in collapse, we investigated the effect of introducing small tangential pressure perturbations in the collapse dynamics of the classic Oppenheimer-Snyder-Datt gravitational collapse, which is an idealized model assuming zero pressure, and which terminates in a black hole final fate as discussed above. Thus we study the stability of the OSD black hole under introduction of small tangential stresses.

It is seen explicitly that there exist classes of stress perturbations such that the introduction of a smallest tangential pressure within the collapsing OSD cloud changes the endstate of collapse to formation of a naked singularity, rather than a black hole. What follows is that small stress perturbations within the collapsing cloud change the final fate of the collapse from being a black hole to a naked singularity. This can also be viewed as perturbing the spacetime metric of the cloud in a small way. Thus we can understand here the role played by tangential pressures in a well-known gravitational collapse scenario. A specific and physically reasonable but generic enough class of perturbations is considered so as to provide a good insight into the genericity of naked singularity formation in collapse when the OSD collapse model is perturbed by introduction of a small pressure. Thus we have an important insight into the structure of the censorship principle which as yet remains to be properly understood.

The general spherically symmetric metric describing the collapsing matter cloud can be written as,

$$ds^2 = -e^{2\sigma(t,r)}dt^2 + e^{2\psi(t,r)}dr^2 + R(t,r)^2 d\Omega^2, \tag{1}$$

with the stress-energy tensor for a generic matter source being given by, $T^t_t = -\rho,\ T^r_r = p_r,\ T^\theta_\theta = T^\phi_\phi = p_\theta$. The above is a general scenario, in that it involves no assumptions about the form of the matter or the equation of state.

As a step towards deciding the stability or otherwise of the OSD collapse model under the injection of small tangential stress perturbations, we consider the dynamical development of the collapsing cloud, as governed by the Einstein equations. The visibility or otherwise of the final singularity that develops in collapse is determined by the behaviour of the apparent horizon in the spacetime, which is the boundary of the trapped surface region that develops as the collapse progresses. First, we define a scaling function $v(r,t)$ by the relation $R = rv$. The Einstein equations for the above spacetime geometry can then be written as,

$$p_r = -\frac{\dot{F}}{R^2\dot{R}},\ \rho = \frac{F'}{R^2R'}\ , \tag{2}$$

$$\sigma' = 2\frac{p_\theta - p_r}{\rho + p_r}\frac{R'}{R} - \frac{p_r'}{\rho + p_r}\ , \tag{3}$$

$$2\dot{R}' = R'\frac{\dot{G}}{G} + \dot{R}\frac{H'}{H} \ , \tag{4}$$

$$F = R(1 - G + H) \ , \tag{5}$$

The functions H and G in the above are defined as, $H = e^{-2\sigma(r,v)}\dot{R}^2, \ G = e^{-2\psi(r,v)}R'^2$. The above are five equations in seven unknowns, namely $\rho, \ p_r, \ p_\theta, \ R, \ F, \ G, \ H$. Here ρ is the mass-energy density, p_r and p_θ are the radial and tangential stresses respectively, R is the physical radius for the matter cloud, and F is the Misner-Sharp mass function.

It is possible now, with the above definitions of v, H and G, to substitute the unknowns R, H with $v, \ \sigma$. Then, without loss of generality, the scaling function v can be written as $v(t_i, r) = 1$ at the initial time $t_i = 0$, when the collapse begins. It then goes to zero at the spacetime singularity t_s, which corresponds to $R = 0$, and thus we have $v(t_s, r) = 0$. This amounts to the scaling $R = r$ at the initial epoch of the collapse, which is an allowed freedom. The collapse condition here is $\dot{R} < 0$ throughout the evolution, and this is equivalent to $\dot{v} < 0$.

One can integrate the Einstein equations, at least up to one order, to reduce them to a first order system, to obtain the function $v(t, r)$. This function, which is monotonically decreasing in t can be inverted to obtain the time needed by a matter shell at any radial value r to reach the event with a particular value v. We can then write the function $t(r, v)$ as,

$$t(r, v) = \int_v^1 \frac{e^{-\sigma}}{\sqrt{\frac{F}{r^3\tilde{v}} + \frac{be^{2rA}-1}{r^2}}} d\tilde{v} \ . \tag{6}$$

The function $A(r, v)$ in the above depends on the nature of the tangential stress perturbations chosen. The time taken by the shell at r to reach the spacetime singularity at $v = 0$ is then $t_s(r) = t(r, 0)$.

Since $t(r, v)$ is in general at least C^2 everywhere in the spacetime (because of the regularity of the functions involved), and is continuous at the centre, we can write it as,

$$t(r, v) = t(0, v) + r\chi(v) + O(r^2) \tag{7}$$

Then, by continuity, the time for a shell located at any r close to the centre to reach the singularity is given as,

$$t_s(r) = t_s(0) + r\chi(0) + O(r^2) \tag{8}$$

Basically, this means that the singularity curve should have a well-defined tangent at the center. Regularity at the center also implies that the metric function σ cannot have constant or linear terms in r in a close neighborhood of $r = 0$, and it must go as $\sigma \sim r^2$ near the center. Therefore the most general choice of the free function σ is,

$$\sigma(r, v) = r^2 g(r, v) \tag{9}$$

Since $g(r, v)$ is a regular function (at least C^2), it can be written near $r = 0$ as,

$$g(r, v) = g_0(v) + g_1(v)r + g_2(v)r^2 + ... \tag{10}$$

We can now investigate how the OSD gravitational collapse scenario, which gives rise to a black hole as the final state, gets altered when small stress perturbations are introduced in the dynamical evolution of collapse. For that we first note that the dust model is obtained if $p_r = p_\theta = 0$ in the above. In that case, $\sigma' = 0$ and together with the condition $\sigma(0) = 0$ gives $\sigma = 0$ identically. In the OSD homogeneous collapse to a black hole, the trapped surfaces and the apparent horizon develop much earlier before the formation of the final singularity. But when density inhomogeneities are allowed in the initial density profile, such as a higher density at the centre of the star, then the trapped surface formation is delayed in a natural manner within the collapsing cloud. Then the final singularity becomes visible to faraway observers in the Universe (e.g. Joshi, Dadhich & Maartens 2002).

The OSD case is obtained from the inhomogeneous dust case, when we assume further that the collapsing dust is necessarily homogeneous at all epochs of collapse. This is of course an idealized scenario because realistic stars would have typically higher densities at the centre, and they also would have non-zero internal stresses. The conditions that must be imposed to obtain the OSD case from the above are given by $M = M_0$ $v = v(t)$ $b_0(r) = k$. Then we have $F' = 3M_0r^2$, $R' = v$, the energy density is homogeneous throughout the collapse, and the density is given by $\rho = \rho(t) = 3M_0/v^3$. The spacetime geometry then becomes the Oppenheimer-Snyder metric, which is given by,

$$ds^2 = -dt^2 + \frac{v^2}{1+kr^2}dr^2 + r^2v^2d\Omega^2, \tag{11}$$

where the function $v(t)$ is a solution of the equation of motion, $\frac{dv}{dt} = \sqrt{(M_0/v)+k}$, obtained from the Einstein equation. In this case we get $\chi(0) = 0$ identically. All the matter shells then collapse into a simultaneous singularity, which is necessarily covered by the event horizon that developed in the spacetime at an earlier time. Therefore the final fate of collapse is a black hole.

To explore the effect of introducing small pressure perturbations in the above OSD scenario and to study the models thus obtained which are close to the Oppenheimer-Snyder, we can relax one or more of the above conditions. If the collapse outcome is not a black hole, the final collapse to singularity cannot be simultaneous. We can thus relax the condition $v = v(t)$ above, allowing for $v = v(t,r)$. We keep the other conditions of the OSD model unchanged, so as not to depart too much from the OSD model, and this should bring out more clearly the role played by the stress perturbations in the model. We know that the metric function $\sigma(t,r)$ must identically vanish for the dust case. On the other hand, the above amounts to allowing for small perturbations in σ, and allowing it to be non-zero now. This is equivalent to introducing small stress perturbations in the model, and it is seen that this affects the apparent horizon developing in the collapsing cloud. We note that taking $M = M_0$ leads to $F = r^3M_0$.

In this case, in the small r limit we obtain $G(r,t) = b(r)e^{2\sigma(r,v)}$. The radial stress p_r vanishes here as $\dot{F} = 0$, and the tangential pressure turns out to have the form, $p_\theta = p_1r^2 + p_2r^3 + \ldots$, where p_1, p_2 are evaluated in terms of coefficients of m, g, and R and its derivatives, and we get,

$$p_\theta = 3\frac{M_0g_0}{vR'^2}r^2 + \frac{9}{2}\frac{M_0g_1}{vR'^2}r^3 + \ldots \tag{12}$$

As seen above, the choice of the sign of the functions g_0 and g_1 is enough to ensure positivity or negativity of the pressure p_θ. The first order coefficient χ in the equation of the time curve of the singularity $t_s(r)$ is now obtained as,

$$\chi(0) = -\int_0^1 \frac{v^{\frac{3}{2}} g_1(v)}{(M_0 + vk + 2vg_0(v))^{\frac{3}{2}}} dv \ . \tag{13}$$

As we have noted above, it is the quantity $\chi(0)$ that governs the nature of the singularity curve, and whether it is increasing or decreasing away from the center. It can be seen from above that the initial data matters in terms of the density and stress profiles, the velocity of the collapsing shells, and the allowed dynamical evolutions that govern and fix the value of $\chi(0)$.

The apparent horizon in the spacetime and the trapped surface formation as the collapse evolves is also governed by the quantity $\chi(0)$, which in turn governs the nakedness or otherwise of the singularity. The equation for the apparent horizon is given by $F/R = 1$. This is analogous to that of the dust case since $F/R = rM/v$ in both these cases. So the apparent horizon curve $r_{ah}(t)$ is given by

$$r_{ah}^2 = \frac{v_{ah}}{M_0}, \tag{14}$$

with $v_{ah} = v(r_{ah}(t), t)$, which can also be inverted as a time curve for the apparent horizon $t_{ah}(r)$. The visibility of the singularity at the center of the collapsing cloud to faraway observers is determined by the nature of this apparent horizon curve which is given by,

$$t_{ah}(r) = t_s(r) - \int_0^{v_{ah}} \frac{e^{-\sigma}}{\sqrt{\frac{M_0}{v} + \frac{be^{2\sigma}-1}{r^2}}} dv \tag{15}$$

where the $t_s(r)$ is the singularity time curve, and its initial point is $t_0 = t_s(0)$. Near $r = 0$ we then get,

$$t_{ah}(r) = t_0 + \chi(0)r + o(r^2) \ . \tag{16}$$

From these considerations, it is possible to see how the stress perturbations affect the time of formation of the apparent horizon, and therefore the formation of a black hole or naked singularity. A naked singularity would typically occur as a collapse endstate when a comoving observer at a fixed r value does not encounter any trapped surfaces before the time of singularity formation. For a black hole to form, trapped surfaces must develop before the singularity. Therefore it is required that,

$$t_{ah}(r) \le t_0 \ \text{ for } \ r > 0, \ \text{ near } \ r = 0 \ . \tag{17}$$

As can be seen from above, for all functions $g_1(v)$ for which $\chi(0)$ is positive, this condition is violated and in that case the apparent horizon is forced to appear after the formation of the central singularity. In that case, the apparent horizon curve begins at the central singularity $r = 0$ at $t = t_0$ and increases with increasing r, moving to the future. Then we have $t_{ah} > t_0$ for $r > 0$ near the center. The behaviour of outgoing families of null geodesics has been analyzed in detail in the case when $\chi(0) > 0$ and we can see that the geodesics terminate at the singularity in the past. Thus timelike and null geodesics come out from the singularity, making it visible to external observers (Joshi & Dwivedi 1999).

One thus sees that it is the term g_1 in the stresses p_θ which decides either the black hole or naked singularity as the final fate for the collapse. We can choose it to be arbitrarily small, and it is then possible to see how introducing a generic tangential stress perturbation in the model would change drastically the final outcome of the collapse. For example, for all non-vanishing tangential stresses with $g_0 = 0$ and $g_1 < 0$, even the slightest perturbation of the Oppenheimer-Snyder-Datt scenario, injecting a small tangential stress would result in a naked singularity. The space of all functions g_1 that make $\chi(0)$ positive, which includes all the strictly negative functions g_1, causes the collapse to end in a naked singularity. While this is an explicit example, it is by no means the only class. The important feature of this class is that it corresponds to a collapse model for a simple and straightforward perturbation of the Oppenheimer-Snyder-Datt spacetime metric.

In this case, the geometry near the centre can be written as,

$$ds^2 = -(1 - 2g_1 r^3)dt^2 + \frac{(v + rv')^2}{1 + kr^2 - 2g_1 r^3}dr^2 + r^2 v^2 d\Omega^2 \ , \tag{18}$$

The metric above satisfies the Einstein equations in the neighborhood of the center of the cloud when the function $g_1(v)$ is small and bounded. We can take $0 < |g_1(v)| < \epsilon$, so that, the smaller we take the parameter ϵ, the bigger will be the radius where the approximation is valid. We can consider here the requirement that a realistic matter model should satisfy some energy conditions ensuring the positivity of mass and energy density. The weak energy condition would imply restrictions on the density and pressure profiles. The energy density as given by the Einstein equation must be positive. Since R is positive, to ensure positivity of ρ we require $F > 0$ and $R' > 0$. The choice of positive $M(r)$, which clearly holds for $M_0 > 0$, and is physically reasonable, ensures positivity of the mass function. Here $R' > 0$ is a sufficient condition for the avoidance of shell crossing singularities. The tangential stress can now be written, with $p_r = 0$, and is given by

$$p_\theta = \frac{1}{2}\frac{R}{R'}\rho\sigma' \tag{19}$$

So the sign of the function σ' would determine the sign of p_θ. Positivity of $\rho + p_\theta$ is then ensured for small values of r throughout collapse for any form of p_θ. In fact, regardless of the values taken by M and g, there will always be a neighbourhood of the center $r = 0$ for which $|p_\theta| < \rho$ and therefore $\rho + p_\theta \geq 0$.

What we see here is that, in the space of initial data in terms of the initial matter densities and velocity profiles, any arbitrarily small neighborhood of the OSD collapse model, which is going to end as a black hole, contains collapse evolutions that go to naked singularities. Such an existence of subspaces of collapse solutions, that go to a naked singularity rather than a black hole, in the arbitrary vicinity of the OSD black hole, presents an intriguing situation. It gives an idea of the richness of structure present in the gravitation theory, and indicates the complex solution space of the Einstein equations which are a complicated set of highly non-linear partial differential equations. What we see here is the existence of classes of stress perturbations such that an arbitrarily small change from the OSD model is a solution going to a naked singularity.

This then provides an intriguing insight into the nature of cosmic censorship, namely that the collapse must necessarily be properly fine-tuned if it is to produce a black hole only as the final endstate. Traditionally it was believed that the presence of stresses or pressures in the collapsing matter cloud would increase the chance of black hole formation, thereby ruling out dust models that were found to lead to a naked singularity as the collapse endstate. It now becomes clear that this is actually not the case. The model described here not only provides a new class of collapses ending in naked singularities, but more importantly, shows how the bifurcation line that separates the phase space of 'black hole formation' from that of 'naked singularity formation' runs directly over the simplest and most studied of black hole scenarios such as the OSD model.

It has to be noted of course that the general issue of stability and genericity of collapse outcomes has been a deep problem in gravitation theory, and requires mathematical breakthroughs as well as evolving further physical understanding of the collapse phenomenon. As noted above, this is again basically connected with the main difficulty of cosmic censorship itself, which is the issue of how to define censorship. However, it is also clear from the discussion above, that consideration of various collapse models along the lines as discussed here does yield considerable insight and inputs in understanding gravitational collapse and its final outcomes.

4. Spinning up a black hole and quantum aspects

It is clear that the black hole and naked singularity outcomes of a complete gravitational collapse for a massive star are very different from each other physically, and would have quite different observational signatures. In the naked singularity case, if it occurs in nature, we have the possibility of observing the physical effects happening in the vicinity of the ultra-dense regions that form in the very final stages of collapse. However, in a black hole scenario, such regions are necessarily hidden within the event horizon of gravity. The fact that a slightest stress perturbation of the OSD collapse could change the collapse final outcome drastically, as we noted in the previous section, changing it from black hole formation to a naked singularity, means that the naked singularity final state for a collapsing star must be studied very carefully to deduce its physical consequences, which are not well understood so far.

It is, however, widely believed that when we have a reasonable and complete quantum theory of gravity available, all spacetime singularities, whether naked or those hidden inside black holes, will be resolved away. As of now, it remains an open question if quantum gravity will remove naked singularities. After all, the occurrence of spacetime singularities could be a purely classical phenomenon, and whether they are naked or covered should not be relevant, because quantum gravity will possibly remove them all any way. But one may argue that looking at the problem this way is missing the real issue. It is possible that in a suitable quantum gravity theory the singularities will be smeared out, though this has not been realized so far. Also there are indications that in quantum gravity also the singularities may not after all go away.

In any case, the important and real issue is, whether the extreme strong gravity regions formed due to gravitational collapse are visible to faraway observers or not. It is quite clear that gravitational collapse would certainly proceed classically, at least till quantum

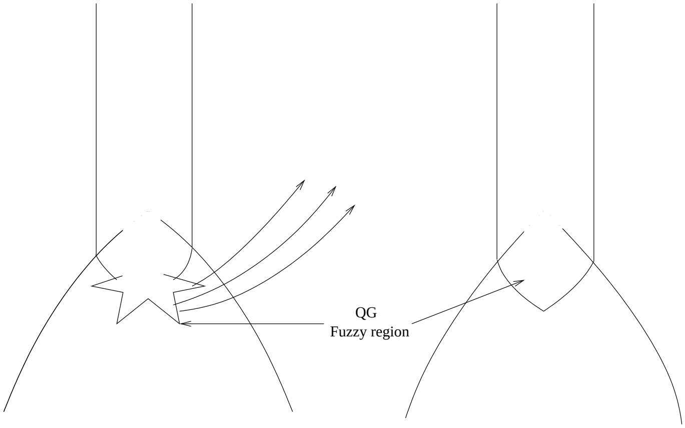


Figure 5. Even if the naked singularity is resolved by the quantum gravity effects, the ultra-strong gravity region that developed in gravitational collapse will still be visible to external observers in the Universe.

gravity starts governing and dominating the dynamical evolution at scales of the order of the Planck length, *i.e.* till extreme gravity configurations have been already developed due to collapse. The key point is the visibility or otherwise of such ultra-dense regions whether they are classical or quantum (see Fig. 5).

What is important is, classical gravity implies necessarily the existence of ultra-strong gravity regions, where both classical and quantum gravity come into their own. In fact, if naked singularities do develop in gravitational collapse, then in a literal sense we come face-to-face with the laws of quantum gravity, whenever such an event occurs in the Universe (Wald 1997).

In this way, the gravitational collapse phenomenon has the potential to provide us with a possibility of actually testing the laws of quantum gravity. In the case of a black hole developing in the collapse of a finite sized object such as a massive star, such strong gravity regions are necessarily hidden behind an event horizon of gravity, and this would happen well before the physical conditions became extreme near the spacetime singularity. In that case, the quantum effects, even if they caused qualitative changes closer to singularity, will be of no physical consequence. This is because no causal communications are then allowed from such regions. On the other hand, if the causal structure were that of a naked singularity, then communications from such a quantum gravity dominated extreme curvature ball would be visible in principle. This will be so either through direct physical processes near a strong curvature naked singularity, or via the secondary effects, such as the shocks produced in the surrounding medium.

Independently of such possibilities connected with gravitational collapse as above, let us suppose that the collapse terminated in a black hole. It is generally believed that such a black hole will be described by the Kerr metric. A black hole, however, by its very nature accretes matter from the surrounding medium or from a companion star. In that case, it is worth noting here that there has been an active debate in recent years about whether a black hole can survive as it is, when it accretes charge and angular momentum from the surrounding medium.

The point is, there is a constraint in this case for the horizon to remain undisturbed, namely that the black hole must not contain too much of charge and it should not spin too fast. Otherwise, the horizon cannot be sustained. It will breakdown and the singularity within will become visible. The black hole may have formed with small enough charge and angular momentum to begin with; however, we have the key astrophysical process of accretion from its surroundings, of the debris and outer layers of the collapsing star. This matter around the black hole will fall into the same with great velocity, which could be classical or quantized, and with either charge or angular momentum or perhaps both. Such in-falling particles could 'charge-up' or 'over-spin' the black hole, thus eliminating the event horizon. Thus, the very fundamental characteristic of a black hole, namely its trait of gobbling up the matter all around it and continuing to grow becomes its own nemesis and a cause of its own destruction.

Thus, even if a massive star collapsed into a black hole rather than a naked singularity, important issues remain such as the stability against accretion of particles with charge or large angular momentum, and whether that can convert the black hole into a naked singularity by eliminating its event horizon. Many researchers have claimed this is possible, and have given models to create naked singularities this way. But there are others who claim there are physical effects which would save the black hole from over-spinning this way and destroying itself, and the issue is very much open as yet. The point is, in general, the stability of the event horizon and the black hole continues to be an important issue for black holes that developed in gravitational collapse. For a recent discussion on some of these issues, we refer to Matsas & da Silva (2007), Matsas *et al.* (2009), Hubeny (1999), Hod (2008), Richartz & Saa (2008), Jacobson & Sotiriou (2009, 2010a,b), Barausse, Cardoso & Khanna (2010), and references therein.

The primary concern of the cosmic censorship hypothesis is the formation of black holes as collapse endstates. Their stability, as discussed above, is only a secondary issue. Therefore, what this means for cosmic censorship is that the collapsing massive star should not retain or carry too much charge or spin; otherwise it could necessarily end up as a naked singularity, rather than a black hole.

5. Predictability, Cauchy horizons and all that

A concern sometimes expressed is that if naked singularities occurred as the final fate of gravitational collapse, that would break the predictability in the spacetime. A naked singularity is characterized by the existence of light rays and particles that emerge from the same. Typically, in all the collapse models discussed above, there is a family of future directed non-spacelike curves that reach external observers, and when extended in the past these meet the singularity. The first light ray that comes out from the singularity marks the boundary of the region that can be predicted from a regular initial Cauchy surface in the spacetime, and that is called the Cauchy horizon for the spacetime. The causal structure of the spacetime would differ significantly in the two cases, when there is a Cauchy horizon and when there is none. A typical gravitational collapse to a naked singularity, with the Cauchy horizon forming is shown in Fig. 6.

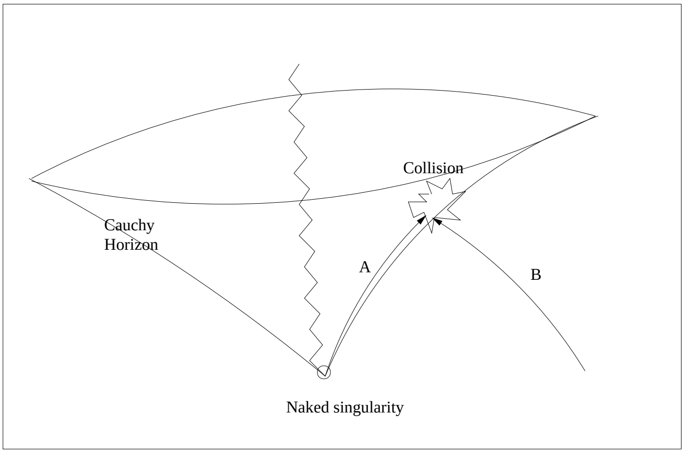


Figure 6. The existence of a naked singularity is typically characterized by existence of a Cauchy horizon in the spacetime. Very high energy particle collisions can occur close to such a Cauchy horizon.

The point here is, given a regular initial data on a spacelike hypersurface, one would like to predict the future and past evolutions in the spacetime for all times (see for example, Hawking & Ellis (1973) for a discussion). Such a requirement is termed the global hyperbolicity of the spacetime. A globally hyperbolic spacetime is a fully predictable universe. It admits a Cauchy surface, the data on which can be evolved for all times in the past as well as in future. Simple enough spacetimes such as the Minkowski or Schwarzschild are globally hyperbolic, but the Reissner-Nordstrom or Kerr geometries are not globally hyperbolic. For further details on these issues, we refer to Hawking & Ellis (1973) or Joshi (2008).

Here we would like to mention certain recent intriguing results in connection to the existence of a Cauchy horizon in a spacetime when there is a naked singularity resulting as final fate of a collapse. Let us suppose the collapse resulted in a naked singularity. In that case, there are classes of models where there will be an outflow of energy and radiations of high velocity particles close to the Cauchy horizon, which is a null hypersurface in the spacetime. Such particles, when they collide with incoming particles, would give rise to a very high center of mass energy of collisions. The closer we are to the Cauchy horizon, higher is the center of mass energy of collisions. In the limit of approach to the Cauchy horizon, these energies approach arbitrarily high values and could reach the Planck scale energies (see for example, Patil & Joshi 2010, 2011a,b; Patil, Joshi & Malafarina 2011).

It has been observed recently that in the vicinity of the event horizon for an extreme Kerr black hole, if the test particles arrive with fine-tuned velocities, they could undergo very high energy collisions with other incoming particles. In that case, the possibility arises that one could see Planck scale physics or ultra-high energy physics effects near the event horizon, given suitable circumstances (Banados, Silk & West 2009; Berti *et al.* 2009; Jacobson & Sotiriou 2010a,b; Wei *et al.* 2010; Grib & Pavlov 2010; Zaslavskii, 2010).

What we mentioned above related to the particle collisions near Cauchy horizon is similar to the situation where the background geometry is that of a naked singularity. These results could mean that in strong gravity regimes, such as those of black holes or naked singularities developing in gravitational collapse, there may be a possibility to observe ultra-high energy physics effects, which would be very difficult to see in the near future in terrestrial laboratories.

While these phenomena give rise to the prospect of observing Planck scale physics near the Cauchy horizon in the gravitational collapse framework, they also raise the following intriguing question. If extreme high energy collisions do take place very close to the null surface which is the Cauchy horizon, then in a certain sense it is essentially equivalent to creating a singularity at the Cauchy horizon. In that case, all or at least some of the Cauchy horizon would be converted into a spacetime singularity, and would effectively mark the end of the spacetime itself. In this case, the spacetime manifold terminates at the Cauchy horizon, whenever a naked singularity is created in gravitational collapse. Since the Cauchy horizon marks in this case the boundary of the spacetime itself, predictability is then restored for the spacetime, because the rest of the spacetime below and in the past of the horizon was predictable before the Cauchy horizon formed.

6. Future perspectives

We have pointed out in the considerations here that the final fate of a massive star continues to be a rather exciting research frontier in black hole physics and gravitation theory today. The outcomes here will be fundamentally important for the basic theory and astrophysical applications of black hole physics, and for modern gravitation physics. We highlighted certain key challenges in the field, and also several recent interesting developments were reviewed. Of course, the issues and the list given here are by no means complete or exhaustive in any manner, and there are several other interesting problems in the field as well.

In closing, as a summary, we would like to mention here a few points which we think require the most immediate attention, and which will have possibly maximum impact on future development in the field.

1. The genericity of the collapse outcomes, for black holes and naked singularities need to be understood very carefully and in further detail. It is by and large well-accepted now, that the general theory of relativity does allow and gives rise to both black holes and naked singularities as the final outcome of continual gravitational collapse, evolving from a regular initial data, and under reasonable physical conditions. What is not fully clear yet is the distribution of these outcomes in the space of all allowed outcomes of collapse. The collapse models discussed above and considerations we gave here would be of some help in this direction, and may throw some light on the distribution of black holes and naked singularity solutions in the initial data space.

2. Many of the models of gravitational collapse analyzed so far are mainly of spherical symmetric collapse. Therefore, the non-spherical collapse needs to be understood in a much better manner. While there are some models which illustrate what the departures from spherical symmetry could do (see e.g. Joshi & Krolak 1996), on the whole, not very much is known for non-spherical collapse. Probably numerical relativity could be of help in this direction (see for example Baiotti & Rezzolla 2006), for a discussion on

the evolving developments as related to applications of numerical methods to gravitational collapse issues. Also, another alternative would be to use global methods to deal with the spacetime geometry involved, as used in the case of singularity theorems in general relativity.

3. At the very least, the collapse models studied so far do help us gain much insight into the structure of the cosmic censorship, whatever final form it may have. But on the other hand, there have also been attempts where researchers have explored the physical applications and implications of the naked singularities investigated so far (see e.g. Harada, Iguchi & Nakao 2000, 2002; Harada *et al.* (2001) and also references therein).

If we could find astrophysical applications of the models that predict naked singularities, and possibly try to test the same through observational methods and the signatures predicted, that could offer a very interesting avenue to get further insights into this problem as a whole.

4. An attractive recent possibility in that regard is to explore the naked singularities as possible particle accelerators as we pointed out above.

Also, the accretion discs around a naked singularity, wherein the matter particles are attracted towards or repulsed away from the singularities with great velocities could provide an excellent venue to test such effects and may lead to predictions of important observational signatures to distinguish the black holes and naked singularities in astrophysical phenomena (see Kovacs & Harko 2010; Pugliese, Quevedo & Ruffini 2011).

5. Finally, further considerations of quantum gravity effects in the vicinity of naked singularities, which are super-ultra-strong gravity regions, could yield intriguing theoretical insights into the phenomenon of gravitational collapse (Goswami, Joshi & Singh 2006).

Acknowledgments

Over past years, discussions with many colleagues and friends have contributed greatly to shape my understanding of the questions discussed here. In particular, I would like to thank I. H. Dwivedi, N. Dadhich, R. Goswami, T. Harada, S. Jhingan, R. Maartens, K. Nakao, R. Saraykar, T. P. Singh, R. Tavakol, and also many other friends with whom I have extensively discussed these issues. Figure 4 here is from Mena, Tavakol & Joshi (2000), and Fig. 6 is from Patil, Joshi & Malafarina (2011). The rest of the figures are from Joshi (2008).

References

Baiotti L., Rezzolla L., 2006, Phys. Rev. Lett., 97, 141101
Banados M., Silk J., West S.M., 2009, Phys. Rev. Lett., 103, 111102
Barausse E., Cardoso V., Khanna G., 2010, Phys. Rev. Lett., 105, 261102
Berti E., Cardoso V., Gualtieri L., Pretorius F., 2009, Phys. Rev. Lett., 103, 239001
Chandrasekhar S., 1934, Observatory, 57, 373
Datt B., 1938, Z. Physik, 108, 314
Goswami R., Joshi P.S., 2007, Phys. Rev. D, 76, 084026
Goswami R., Joshi P.S., Singh P., 2006, Phys. Rev. Lett., 96, 031302
Grib A.A., Pavlov Y. V., 2010, arXiv:1004.0913 [gr-qc].

Hawking S. W., Ellis G.F.R., 1973, The Large Scale Structure of Space-time, Cambridge University Press, Cambridge

Harada T., Iguchi H., Nakao K., 2000, Phys. Rev. D, 61, 101502

Harada T., Iguchi H., Nakao K., 2002, Prog. Theor. Phys., 107, 449

Harada T., Iguchi H., Nakao K., Singh T.P., Tanaka T, Vaz C., 2001, Phys. Rev. D, 64, 041501

Hod S., 2008, Phys. Rev. Lett., 100, 121101

Hubeny V.E., 1999, Phys. Rev. D, 59, 064013

Jacobson T., Sotiriou T.P., 2009, Phys. Rev. Lett., 103, 141101

Jacobson T., Sotiriou T.P., 2010a, J. Phys. Conf. Ser., 222, 012041

Jacobson T., Sotiriou T.P., 2010b, Phys. Rev. Lett., 104, 021101

Joshi P.S., 2008, Gravitational Collapse and Spacetime Singularities, Cambridge University Press, Cambridge.

Joshi P.S., Dwivedi I.H., 1999, Class. Quantum Grav., 16, 41

Joshi P.S., Krolak A., 1996, Class. Quant. Grav., 13, 3069

Joshi P.S., Malafarina D., 2011, Phys. Rev. D, 83, 024009

Joshi P.S., Dadhich N., Maartens R., 2002, Phys. Rev. D, 65, 101501

Kovacs Z., Harko T., 2010, Phys. Rev. D, 82, 124047

Matsas G. E.A., da Silva A.A.R., 2007, Phys. Rev. Lett., 99, 181301

Matsas G.E.A., Richartz M., Saa A., da Silva A.A.R., Vanzella D.A.T., 2009, Phys. Rev. D, 79, 101502

Mena F.C., Tavakol R., Joshi P.S., 2000, Phys. Rev. D, 62, 044001

Oppenheimer J.R., Snyder H., 1939, Phys. Rev., 56, 455.

Patil M., Joshi P.S., 2010, Phys. Rev. D, 82, 104049

Patil M., Joshi P.S., 2011a, arXiv:1103.1082 [gr-qc]

Patil M., Joshi P.S., 2011b, arXiv:1103.1083 [gr-qc]

Patil M., Joshi P.S., Malafarina D., 2011, Phys. Rev. D., 83, 064007

Penrose R., 1969, Riv. Nuovo Cimento Soc. Ital. Fis., 1, 252

Pugliese D., Quevedo H., Ruffini R., 2011, arXiv:1103.1807 [gr-qc]

Richartz, M., Saa A., 2008, Phys. Rev. D, 78, 081503

Report of the Physics Survey Committee 1986, Physics through the 1990s: gravitation, cosmology, and cosmic-ray physics, National Academy Press, Washington, D.C.

Wald R.M., 1997, gr-qc/9710068.

Wei S.-W., Liu Y.-X., Heng G., Fu C.-E., 2010, Phys. Rev. D, 82, 103005

Zaslavskii O.B., 2010, JETP Lett., 92, 571

This page intentionally left blank

Bull. Astr. Soc. India (2011)

Problems of collisional stellar dynamics

D. C. Heggie[*]
University of Edinburgh, School of Mathematics and the Maxwell Institute for Mathematical Sciences, King's Buildings, Edinburgh EH9 3JZ, U.K.

Received February 15, 2011; accepted March 1, 2011

Abstract. The discovery of dynamical friction was Chandrasekhar's best known contribution to the theory of stellar dynamics, but his work ranged from the few-body problem to the limit of large *N* (in effect, galaxies). Much of this work was summarised in the text "Principles of Stellar Dynamics" (Chandrasekhar 1942, 1960), which ranges from a precise calculation of the time of relaxation, through a long analysis of galaxy models, to the behaviour of star clusters in tidal fields. The later edition also includes the work on dynamical friction and related issues. In this review we focus on progress in the collisional aspects of these problems, i.e. those where few-body interactions play a dominant role, and so we omit further discussion of galaxy dynamics.[1] But we try to link Chandrasekhar's fundamental discoveries in collisional problems with the progress that has been made in the 50 years since the publication of the enlarged edition.

Keywords : binaries: general – galaxies: kinematics and dynamics – globular clusters: general – open clusters and associations: general

1. Introduction

Chandrasekhar's "Principles of Stellar Dynamics" is not his best-known text, but it had few competitors for many years, and covered a broad range of topics. The later edition (Chandrasekhar 1960) added a number of lengthy and significant papers mainly on the statistical approach to collisional stellar dynamics, and was published just over 50 years ago. In this review we consider a few of the topics considered by Chandrasekhar, and try to connect his view of the subject with current research.

Since the book is not so well known, and has been virtually supplanted by the book by Binney and Tremaine (Binney & Tremaine 1987, 2008), it is worth looking over its contents list. After an observational review, the theory begins with a careful derivation of the relaxation time of stellar systems, including all the geometry of two-body encounters

[*] e-mail: d.c.heggie@ed.ac.uk

[1] There is one other such problem to which Chandrasekhar contributed, though the paper in question (Chandrasekhar 1944) was not reprinted in the book. See Section 2. For more on the collisionless problems studied by Chandrasekhar, see the paper by N. Wyn Evans (2011) in the present volume.

and a discussion of the origin of the Coulomb logarithm. It contains a derivation of a formula for what came to be known as “dynamical friction”, possibly Chandrasekhar’s most significant and enduring discovery in the field of stellar dynamics. The history and context of Chandrasekhar’s work in these two topics is nicely discussed in Padmanabhan (1996), and in the present paper we consider more recent developments in the theory of dynamical friction in Section 4.

In Chandrasekhar’s book there follows two chapters on collisionless stellar dynamics, or rather the dynamics of galaxies. After some preliminaries, the first of these considers the problem of determining what galactic potentials are consistent with the assumption of a Schwarzschild (Gaussian) distribution of velocities, while the second turns to the problem of spiral structure. Then there is a long and interesting chapter on collisional stellar dynamics; specifically, the dynamics of star clusters, a subject which we review here in Section 3.

Apart from two appendices, at this point the old and new editions of the book differ. The latter now includes several reprints, some on dynamical friction (a topic we take up here in Section 4), and a final long paper on the statistics of the gravitational field of a distribution of point masses, together with its applications to dynamical friction and star clusters. It is amusing to find that Chandrasekhar titled this last paper “New Methods in Stellar Dynamics”. One wonders if this was a conscious echo of Poincaré’s famous “Les Méthodes Nouvelles de la Mécanique Celeste” (see also Hénon 1967). At any rate, one of Chandrasekhar’s applications of his theory is the starting point for the next section of this review.

2. The dynamics of binary stars

We begin with a slim paper “On the Stability of Binary Systems” (Chandrasekhar 1944). It did not make it into his book, but it appears to be Chandrasekhar’s only work on a topic which has become one of the pillars of collisional stellar dynamics. In Chandrasekhar’s paper, it is set in the context of a critique by Ambartsumian (Ambartsumian 1937) of an older idea by Jeans, who had used information on the distribution of binary stars to argue that the Milky Way was well relaxed.

2.1 The statistical effect of encounters

Chandrasekhar’s estimate for the disruption time scale, τ, of a binary was based on his theory of the two-point distribution for the gravitational acceleration due to a random distribution of stars, which led to the formula

$$\tau = \frac{(M_1 + M_2)^{1/2}}{4\pi G^{1/2} M N a^{3/2}},$$

where M_1, M_2 are the component masses, a is the semi-major axis, and M, N are the average individual mass and number density of the field stars. Ambartsumian’s formula, by contrast, was

$$\tau = \frac{v}{4\pi G M a N \ln\left(1 + \dfrac{a^2 v^4}{4G^2 M^2}\right)},$$

where v is some average speed which we take here to be the velocity dispersion. Notice that, by contrast, Chandrasekhar's formula includes no information on the velocity dispersion, because the underlying theory describes the spatial correlation of fluctuations but not their temporal correlation. He gives the impression that the absence of any dependence on the velocities is a merit, but it later turned out that the velocities of the stars are crucial. Jeans himself (Jeans 1918) had argued (incorrectly, as it later emerged) that encounters with field stars would lead to equipartition of kinetic energies, giving all binaries a period of order

$$P = \frac{G(M_1 + M_2)}{v^3}. \tag{1}$$

This conclusion was, however, turned upside down by Gurevich & Levin (1950), who used arguments akin to those of Ambartsumian and obtained formulae for the average rate of change of the binding energy of a binary as a result of encounters. They concluded that, if a binary has period much longer than equation (1), then its period will tend to become longer still (eventually leading to disruption), while if its period is much shorter then it becomes still shorter. Their conclusion was correct, and was arrived at independently by Hills (Aarseth & Hills 1972; Hills 1975) using numerical methods and by Heggie (Heggie 1975) using mainly analytical approximations. Heggie also used the terms "hard" and "soft" to signify binaries whose internal binding energy, ε, was larger or smaller, respectively, than the mean kinetic energy of the field stars.

The case of equal masses has been worked out in a lot of detail, especially in a series of papers by Hut and colleagues, much of it summarised with tables, figures and formulae in Heggie & Hut (1993). Applications, of course, require unequal masses in general, and here our knowledge is much more patchy. Heggie, Hut & McMillan (1996) were able to give a general formula for exchange encounters with hard binaries (i.e. encounters in which the incoming third star takes the place of one of the original binary components). They used analytical arguments to establish the mass dependence for extreme cases (e.g. one component of very low mass), and filled in the gaps by interpolation in results of numerical experiments. For this purpose they used the starlab package (http://www.manybody.org/manybody/starlab.html, McMillan & Hut (1996)), which has very well organised tools for the computation of scattering cross sections. Large numbers of other results for various specific combinations of masses will be found scattered in the literature, including Sigurdsson & Phinney (1993) and especially the compendious book of Valtonen & Karttunen (2006).

A number of extreme parameter ranges have become important for astrophysical reasons, and also these are situations in which the complexities of the mass-dependence of a cross section may be considerably reduced. On the other hand, as we shall see, the situation can be considerably richer than the simple notion that soft binaries soften and hard binaries harden.

The case of a third body (intruder) of relatively low mass has been studied in the context of the hardening of a black hole-black hole binary in a galactic nucleus (Mikkola & Valtonen 1992). It led to an interesting debate (Hills 1990; Gould 1991) on whether it is really true that hard binaries (defined by the ratio of the binding energy of the binary to the energy of relative motion of the intruder) tend to harden and soft binaries tend to soften.

Hills had argued that it was the ratio of *speeds* that mattered, but Quinlan (1996) eventually vindicated the earlier position. There is now a considerable literature on the problem of a massive binary in a system of particles of low mass which uses N-body simulations rather than cross sections.

Another important context where the distinction between fast encounters and energetic encounters is significant is the study of stellar encounters with planetary systems. This has been studied by several groups, including Laughlin & Adams (1998); Hills & Dissly (1989); Malmberg, Davies & Heggie (2010); Spurzem *et al.* (2009), but here we focus briefly on the study of Fregeau, Chatterjee & Rasio (2006). In the case under consideration, let us denote the incoming velocity which distinguishes hard from soft binaries (in the sense of energies) by v_c (the "critical" velocity, in the sense that slower encounters cannot destroy the binary). Because the planetary mass is so low, v_c is much smaller than the orbital speed of the planet (v_{orb}). These authors find that, indeed, when the encounter speed is less than a speed of order v_c, the average change in the binding energy of the binary is positive, i.e. the encounter hardens the binary. But at the same time the planetary system has been destroyed, because the most likely outcome of a close encounter in this regime is an exchange encounter leaving the two stars bound. Similarly, in the regime of encounter speeds between v_c and v_{orb} an encounter indeed tends to soften a planetary system, but not to disrupt it (until encounter speeds of order v_{orb} are reached). A careful reading of Fregeau, Chatterjee & Rasio (2006) is recommended for a proper appreciation of the issues.

The last case of extreme masses that we shall mention is another highly topical one: that of a single black hole encountering a binary with stellar-mass components. This is thought to be of importance in the creation of high-velocity stars by scattering off the black hole at the Galactic Centre (Hills 1988). The literature is considerable, but among those studies focusing on the three-body aspects of the problem are Zhang, Lu & Yu (2010); Gvaramadze, Gualandris & Portegies Zwart (2009); Sari, Kobayashi & Rossi (2010); Gualandris, Portegies Zwart & Sipior (2005); Yu & Tremaine (2003).

2.2 Wide binaries in the Milky Way

Chandrasekhar's interest in binary stars was focused on the dynamical evolution of field binaries, a topic which remains lively up to the present day, with an extensive literature, especially on the observational side. Before turning to recent developments, however, it is worth mentioning that the topic had already been considered, before the work of Chandrasekhar and Ambartsumian, by Öpik (1932).

This thread of research was then taken up by Oort (1950) who, like Öpik, was concerned with binaries with one massless component (a comet or meteoroid). While all these studies used analytical estimates, soon after this numerical integrations came into routine use, and were applied to this problem by Yabushita (1966) and Cruz-González & Poveda (1971). The latter authors found that the lifetime exceeded the estimates given by any of the previous theories which they tested. There were also theoretical developments, however. Except for Chandrasekhar's theory, that earlier work had been based on the computation of the mean square change of velocity (i.e. the relative velocity between the two components of a binary), or the mean change in the energy of the binary. As Chandrasekhar himself would have recognised, however, it is also necessary to take into account the second moment of the change in energy (i.e. its mean square value), to construct a kinetic theory based on a Fokker-Planck treatment. This was accomplished in King (1977) and Retterer & King (1982).

Further theoretical developments have mainly involved improvements in the physical model, i.e. the inclusion of significant additional processes, such as encounters with giant molecular clouds (Weinberg, Shapiro & Wasserman 1987) and with dark matter particles (Wasserman & Weinberg 1987). Nor are the wide binaries of the Galactic field lacking interest even after they have dissolved. Then they are also strongly subject to galactic perturbations, which impose an interesting (and potentially detectable) correlation in density with a peak when the two components have separated by 100-300pc (Jiang & Tremaine 2010). Finally, it is not self-evident how wide binaries can emerge from the relatively dense star-cluster environment in which most stars are thought to form, and indeed it seems likely that significant numbers form during the cluster dissolution process itself (Kouwenhoven *et al.* 2010).

3. The dynamics of star clusters

The title of this section is also the title Chandrasekhar chose for the last chapter of his book. As usual, it opens with a number of generalities, but then it moves on to the important problem of the escape rate from a star cluster, including the differential escape of stars of low and high mass. Implicit in this theory is the assumption that the cluster is isolated, but the next section of his book moves on to consider the effect on a cluster of its galactic environment. This section begins with an excellent derivation of equations of motion, “energy”-integral and virial theorem.

3.1 Tidal stability

After the preliminaries, Chandrasekhar takes up the stability of star clusters, using as his model an ellipsoidal cluster of uniform density ρ. The reason for this is that the gravitational field (including the tidal field of the galaxy) becomes linear, and the motions of the stars can be computed explicitly. The frequencies become imaginary when ρ is sufficiently low, and Chandrasekhar interprets this as the onset of instability. A somewhat similar approach was taken by Angeletti, Capuzzo-Dolcetta & Giannone (1983), who studied orbits in the nearly constant-density core of a cluster. They used Floquet analysis to determine the stability limit and were thus able to extend results to the case of a cluster on a non-circular galactic orbit.

This section of Chandrasekhar’s book is of particular interest to the author of the present paper because it turns out to be possible to construct a self-consistent ellipsoidal model of uniform density by superposition of these exact orbits (Mitchell & Heggie 2007), though they are indeed unstable when the density is low enough (Fellhauer & Heggie 2005). Though these models are artificial,[2] they are of interest because examples of self-consistent

[2] The later paper has never been cited so far, much as the referee predicted. The present paper will probably provide its one and only citation.

cluster models in a tidal field are rare (Heggie & Ramamani 1995; Bertin & Varri 2008; Varri & Bertin 2009). These models also give a pointer for the construction of better models of star clusters than any in existence, in the following sense. These models consist of the familiar galactic epicycles, but modified by the attraction of the cluster. They are therefore retrograde orbits, and it has been known for a long time that there should exist *stable* retrograde orbits in the vicinity of a star cluster, but outside its tidal radius (Hénon 1970). Thus one can imagine a sequence of self-consistent cluster models with varying proportions of stars inside and outside the cluster tidal radius, with (say) the models of Heggie & Ramamani (1995) (which are generalised King models) at one end of the sequence, and the models of Mitchell & Heggie (2007) at the other.

3.2 Fokker-Planck dynamics

Towards the end of his chapter on star clusters, in which Chandrasekhar has discussed both escape and relaxation, he laments, "A rigorous theory of galactic clusters must therefore take both these factors into account. But such a theory is not yet available." It was not too long in coming, the essential formalism being established by Kuzmin (1957). But its power was first demonstrated by Hénon (1961), in a landmark paper which, among other things, produced a solution of the Fokker-Planck equation (for the relaxation) with a tidal boundary condition (producing escape). Not only this, but Hénon also realised the critical role played by binaries.

Hénon's model was of a very special type, but one which all reasonable solutions would approach asymptotically. It took almost another 20 years before a general numerical solution of the equation became feasible (Cohn 1979), though initially restricted to the case of stars of equal mass, as in Hénon's model. This refinement was added relatively quickly, however (Merritt 1981, 1983), albeit in the context of *galaxy* clusters (as opposed to galactic ones). The subsequent development of this tool was steady: rotation (Goodman 1983), binary stars formed in three-body encounters (quoted in Cohn 1985) or those formed tidally (Statler, Ostriker & Cohn 1987) or primordially (Gao et al. 1991) and stellar evolution (Kim, Chun & Min 1991) were all added; until it became a tool which could be applied to the modelling of individual star clusters and quite detailed comparison with observations.

This was not the first kind of code to reach this goal, however. In advance of the development of Fokker-Planck methods was a method of treating the dynamics of a star cluster as if it was a self-gravitating gas (Larson 1970). This technique developed with comparable rapidity, and after only 10 years it was possible to produce synthetic surface density profiles for comparison with observation (Angeletti, Dolcetta & Giannone 1980). These gas models remained of importance, and were responsible for the discovery of gravothermal oscillations (Sugimoto & Bettwieser 1983), which are the response of a system to an unstable balance between the relaxation-driven flow of energy and the redistribution of energy by binary interactions in the core. Interest in this behaviour slowly waned, but has recently been invoked as a significant mechanism for understanding the variety of surface brightness profiles exhibited by well observed star clusters (Fig. 1).

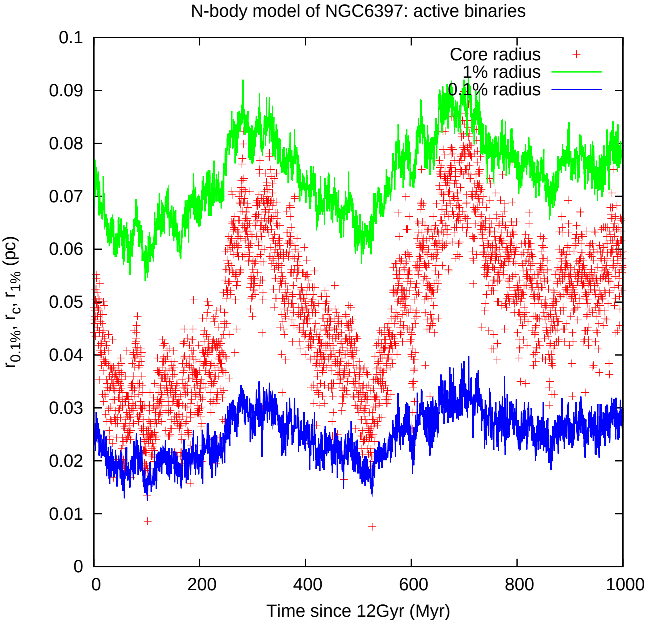


Figure 1. Evolution of the core in a direct N-body model of the globular cluster NGC6397 (from Heggie & Giersz (2009)). From top to bottom, the plotted radii are the 1% Lagrangian radius (i.e. the radius of a sphere, centred at the densest part of the cluster, which contains 1% of the total mass), the core radius (i.e. the radius at which the density of the cluster falls to a certain fraction of its central value, though it is actually calculated by a different procedure), and the 0.1% Lagrangian radius. Radii are given in parsecs, and the horizontal axis is time (in units of 1Myr) after the present. The cluster core is alternately compact and more extended, on a time scale which is long by comparison with the central relaxation time; this, and other arguments, suggest that the oscillations are essentially gravothermal.

3.3 Monte Carlo models

From the numerical point of view, both the Fokker-Planck and gas models are of finite difference type. It is also possible to solve the former equation with at least two kinds of Monte Carlo technique. One of these was pioneered by Spitzer and his students (see Spitzer & Hart (1971) and subsequent papers in the series). Its last serious application appeared many years ago (Spitzer & Mathieu 1980), and it is probably ripe for revival, as it better adapted to some important situations (e.g. a time-dependent tide) than some competitors.

An alternative Monte Carlo technique was developed at about the same time (Hénon 1967, 1971), but has been taken up and developed by others, right up to the present (Stodolkiewicz 1982; Giersz 2006; Chatterjee *et al.* 2010). It now includes a rich mix of important ingredients, not only relaxation and escape, but also the internal (stellar) evolution of single stars and binaries, as well as interactions involving binaries. Despite its limitations to a steady tidal field, spherical symmetry and zero rotation, it is the method of choice for studying virtually all globular star clusters, because it is so fast, without sacrificing much realism. Even the evolution of a rich star cluster like 47 Tuc, which is thought to have almost 2×10^6 stars and a few percent of binaries, can be modelled for a Hubble time in less than a week on an ordinary computer (Giersz & Heggie 2010). Such modelling makes possible a range of investigations at the interface with observations, and is very useful for the planning and interpretation of some kinds of observational programmes, such as searches for radial velocity binaries (Sommariva *et al.* 2009). The speed is important, because we do not know *ab initio* what initial conditions to use to match a given cluster. Repeated trial and error, or grid searches, require a fast method.

3.4 *N*-body methods

Naturally enough, there is nothing in *Principles of Stellar Dynamics* which prepares us for the explosion of interest in *N*-body methods in the subject, even if we restrict ourselves to direct summation methods. It started about 20 years after the publication of the book (von Hoerner 1960), and in 50 years has brought us to the point where it begins to be possible to study the entire life history of the easiest globular clusters (Hasani Zonoozi *et al.* 2011), though most still lie beyond our capabilities.

The main problem is the number of stars, N. Figure 2 shows the steady but slow progress that has been made since 1960. The mean mass of the Galactic globular clusters being (Mandushev, Staneva & Spasova 1991) of order $1.9 \times 10^5 M_\odot$ (and the median mass is lower still), it might be thought that a large fraction of them are within reach of N-body simulation. However, they lose large amounts of mass through evolution over about 12Gyr, and so, except for a few sparse and large clusters, the initial stages of the evolution prevents them from being simulated in a reasonable time.

Actually, it is not hard to evolve larger models than those shown in Fig. 2 to well beyond core collapse. Figure 3 shows a simulation using as initial conditions those suggested for the cluster M4 by Heggie & Giersz (2008), except that there are *no primordial binaries.* If these had been included (and the suggested abundance is only about 7%) the progress of the simulation would have been slower by about a factor of 20.

3.5 The escape rate from star clusters

Chandrasekhar (1943a,b) produced two papers on this topic in quick succession. His motivation for this was not simply to understand the lifetime of star clusters, but to elucidate the role played by dynamical friction (Section 4). Dynamical friction is an aspect of two-body relaxation which tends to reduce the energy of stars, especially those of high speed, and which therefore particularly affects escaping stars. Without it, Chandrasekhar showed, the lifetime of a star cluster would be too short to explain the existence of star clusters with ages of order a Gyr. Somewhat analogously, it has also been invoked in studies of the escape of black holes from a galaxy (Kapoor 1985a,b).

In Chandrasekhar's papers he was, in effect, solving the Fokker-Planck equation assuming that escape took place at some fixed speed (which he estimated from the virial theorem). Similar calculations have been carried out by Spitzer & Harm (1958); King (1960) and Lemou & Chavanis (2010). Long ago, however, King (1958) pointed out a

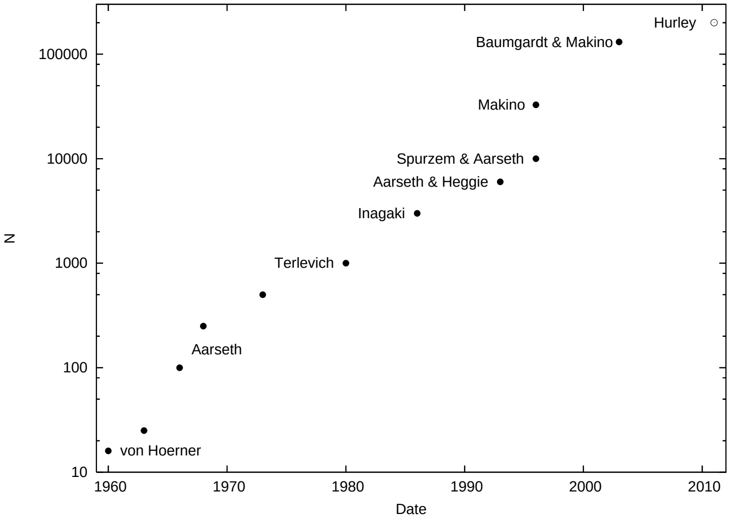


Figure 2. Largest direct N-body simulation to date, plotted against publication date. Only dynamically well evolved simulations (roughly speaking, to or beyond "core collapse"; see, for example, Heggie & Hut (2003)) are included. The last simulation is not yet published at the time of writing (February 2011), in fact see Hurley *et al.* (2008).

number of shortcomings of the Chandrasekhar model, and proceeded to investigate some of them in subsequent papers. One of these was the spatial inhomogeneity of the star cluster model, which, in the same paper, he investigated by integrating the escape rate over the cluster. The escape rate formula which he integrated was a more primitive estimate than Chandrasekhar's, dating back to the earlier work of Ambartsumian (1938) and Spitzer (1940). A similar treatment, but based on more elaborate formulae for the local escape rate were later presented by Saito (1976) and Johnstone (1993).

All the formulae in the papers cited are based on the theory of relaxation, and therefore include as a factor the Coulomb logarithm. A completely different view of the situation was taken by Hénon (1960), who showed that relaxation cannot lead to escape from an isolated cluster, essentially because the "period" of a star's orbit tends to infinity as its energy approaches zero (from below). He obtained a formula for the escape rate due to discrete, individual encounters (rather than the diffusive effect of many encounters). Tellingly, it does not contain the Coulomb logarithm.

One of the comments made by King (1958) was that, however one computes the escape rate, it will change as the cluster evolves. Spitzer & Shapiro (1972) pointed out that relaxation changes the distribution function of the stars in a cluster, and then a single encounter, as envisaged by Hénon, may raise the energy of a star above the energy of escape. Therefore it is possible that the relaxation time scale does, after all, control the escape rate from an isolated system, as is commonly assumed.

There have been many numerical studies addressing the problem, but we shall mention only one (Baumgardt, Hut & Heggie 2002), which showed that another, additional mecha-

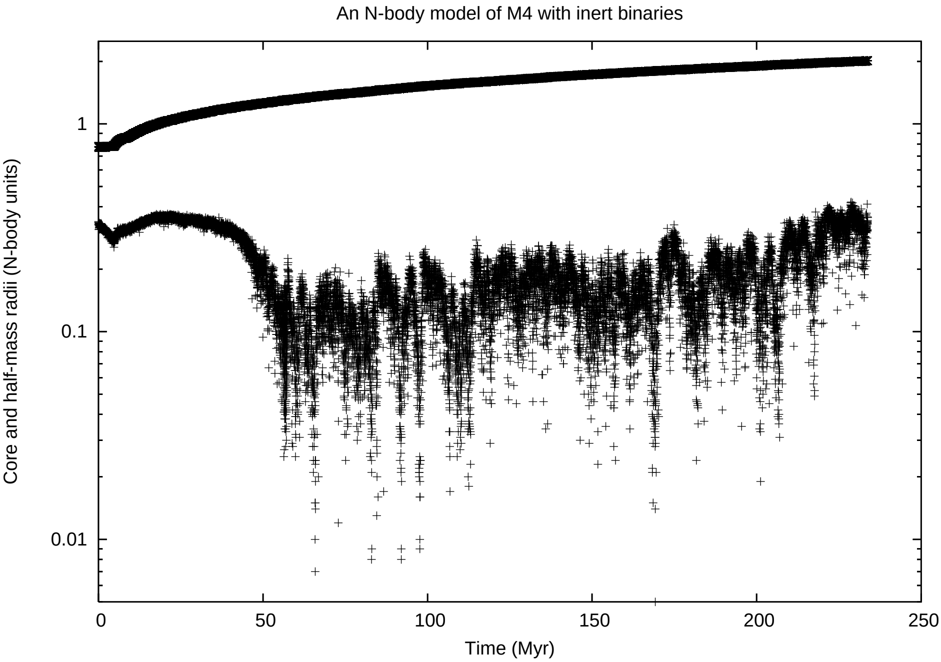


Figure 3. Core- and half-mass radii of an N-body model with 453 000 particles initially. The initial conditions are described in the text. The model includes stellar evolution and the effect of the Galactic tide, but *no binaries* (initially). The initial brief reduction of the core radius is caused by mass segregation. After about 5Myr, stellar evolution temporarily halts the collapse of the core, but this resumes, and reaches completion after about 50Myr. Thereafter the small core is sustained by the formation and dynamical evolution of binary stars, but the core is now unstable to "gravothermal oscillations". Both stellar evolution (or rather its associated mass loss) and the evolution of binaries increase the energy of the cluster, which leads to the expansion of the half-mass radius. The tidal radius (not shown), which marks the effective boundary between motions dominated by the cluster and those dominated by the Galaxy, decreases by only about 15% in the time shown. The units of radius are "N-body units" (see Heggie & Hut (2003)). This simulation took about 2 months.

nism is at play. While it is true that most escapees emerge from encounters deep inside the cluster, some occur because this process causes the potential well of the cluster to become shallower, and this in turn causes stars with energies just below the energy of escape to drift across the escape boundary.

All these issues change when one considers a cluster limited by the tidal field of a galaxy. At the simplest level, the energy of escape drops and the rate of escape increases, as was found numerically many years ago (Wielen 1968; Hayli 1970). But the very notion of escape is complicated. It is possible, at least on the standard model of a star cluster on a circular galactic orbit, for an escapee to recede arbitrarily far from from the cluster and still return to it (Ross, Mennim & Heggie 1997). Stars can exist on stable orbits with energies above the escape energy, and even on orbits which lie outside the conventional tidal radius of the cluster (Hénon 1970); see also Section 3.1. Such stars have important effects on the observable velocity dispersion profile of a globular cluster (Küpper *et al.* 2010). Matters are complicated further for clusters on *non-circular* galactic orbits, where there is no conserved quantity analogous to energy (and therefore no notion of escape energy or even of an escape boundary). Nevertheless the common view is that, even in these cases, the time scale of escape is determined by the time of relaxation.

Among early indications that this is not so were numerical results by Vesperini & Heggie (1997), who showed that the escape rate depended systematically on N, even when scaled by the relaxation time. N-body models by T. Fukushige and J. Makino (Heggie et al. 1998) showed clearly that escape scales with N in a different way from relaxation. Further N-body results (Heggie 2001a,b) gave an escape time scale proportional (empirically) to about $N^{0.63}$, whereas in the same units the relaxation time scales approximately as $N/\log N$.

The problem was greatly clarified by the work of Baumgardt (2001). He showed that the scaling could be understood by noting that stars may remain inside a static cluster for an arbitrarily long time, even with energies above the escape energy (Fukushige & Heggie 2000), and that this changes the escape time scale from the relaxation time, t_r, to approximately $t_r^{3/4}t_{cr}^{1/4}$, where t_{cr} is the crossing time. For the range of N studied in N-body simulations of the time, this results in a dependence close to $N^{0.63}$, in units such that t_{cr} is constant.

It is the interaction between this buffer of "potential escapees" and the processes of relaxation and escape which complicates the overall escape time scale. The effect of this buffer is sometimes referred to as a "retardation effect", after a study by King (1959), which was in turn suggested by a remark of Chandrasekhar (1960, p.209). The point is that a star which has gained enough energy to escape may, on its way out of the cluster, experience another encounter which brings it below the escape energy once more. But the N-dependence of this effect is different from that described in Baumgardt (2001), and there it is not encounters which retain an escapee, but the dynamics of stars in the field of tidal and inertial forces.

The scaling does depend on factors such as the initial concentration of the cluster (Tanikawa & Fukushige 2005, 2010), the extent to which the cluster initially underfills its tidal radius (Gieles & Baumgardt 2008), and the galactic environment (i.e. the strength of the tidal field, Lamers, Gieles & Portegies Zwart (2005)). Amazingly, it does not depend significantly on the assumption that the cluster orbit is circular (Baumgardt & Makino 2003); if the orbit is non-circular, the cluster appears to behave like one on a circular orbit of intermediate radius. Understanding this fact from a theoretical point of view is a significant unsolved problem in this area, despite some empirical advances with the aid of N-body simulations (Küpper *et al.* 2010). There is also growing observational evidence on the mass-dependence of cluster disruption, and it is consistent with these theoretical developments (Boutloukos & Lamers 2003; Gieles *et al.* 2005; Lamers & Gieles 2006), though it has to be recognised that other processes come into play beyond the relatively gentle evaporation of escapees created in encounters. That is a long and old story which we shall not review here.

Equally old and long is the theory of what is called *preferential* or *differential* escape, i.e. its dependence on the mass of the escapee. The common opinion is that the escape rate increases with decreasing mass, but Chandrasekhar's finding (Chandrasekhar 1960) (p.209f) was more subtle. His result was that the escape rate is fastest for stars with a mass of about 40% of the mean mass. This result was based on the assumption that stars are in energy equipartition in the cluster, which is inconsistent with a fixed escape velocity. Nevertheless the result still turns out to hold in star clusters which include stellar evolution and a low remnant retention fraction (Kruijssen 2009), if the total disruption time of the cluster is short enough.

4. Dynamical friction

This is another subject with a long and rich history, and it takes us beyond star cluster dynamics into the dynamics of galaxies and galaxy clusters. In that context, which we come to at the end of this section, it also takes us away from the collisional problems to which this review has been devoted. Within collisional stellar dynamics, dynamical friction is simply part of the mainstream of the theory of relaxation, and does not often receive separate, explicit mention. There is an interesting experimental check of what is, in effect, the coefficient of dynamical friction in Theuns (1996). Within limits the comparison is satisfactory, but this study also shows that direct comparison is not an easy task.

One current problem of collisional stellar dynamics involving dynamical friction explicitly (and linking with the topics of Section 2) is the fate of black holes in merging galaxies. Their evolution was outlined in a famous paper of Begelman, Blandford & Rees (1980), and much subsequent attention has been paid to a protracted period of evolution under the action of dynamical friction, often referred to as “the final parsec problem”. Some theoretical studies relevant to this problem (scattering of low-mass objects off a massive binary) are referred to in Section 2.1, and others which refer specifically to the galactic context include Polnarev & Rees (1994) and Vecchio, Colpi & Polnarev (1994). (Of course it is a big assumption to suppose that this process of the evolution of pairs of black holes can be understood entirely in terms of stellar dynamics; the effect of galactic gas and accretion disks around the black holes may be decisive; but we shall continue to ignore these in our further review.)

In the stellar dynamical problem Chandrasekhar’s formula has been extended in several ways, e.g. to a non-uniform background medium (Just & Peñarrubia 2005), one with an anisotropic velocity distribution (Ideta 2002), or one with a mass spectrum (Ciotti 2010). N-body techniques are possible, but demanding, because it is known on theoretical grounds that it is necessary to include the effect of the black holes on the stellar distribution self-consistently (Iwasawa *et al.* 2011; Sesana 2010). To reach a regime which can be scaled robustly to galactic nuclei is a computational challenge comparable to the simulation of globular clusters (Section 3.4). Progress has been faster than for the globular cluster problem, however, partly because there is no need (it is assumed) to follow also a binary population in the stellar distribution (Berczik, Merritt & Spurzem 2005; Berczik *et al.* 2006; Berentzen *et al.* 2009).

Black holes are point masses, but the notion of dynamical friction has been extended (in numerous studies) to problems of the orbital evolution of a satellite galaxy within a larger galaxy or halo, i.e. to extended bodies. From the theoretical point of view it seems clear that the behaviour of a rigid satellite (which is the basis of some theoretical studies) may differ essentially from that of a responsive satellite (Fujii, Funato & Makino 2006). A common approach is to use a more-or-less self-consistent N-body simulation and to summarise the results by a calibration of the Coulomb logarithm in the Chandrasekhar formula; examples include Chan, Mamon & Gerbal (1997); Cora, Muzzio & Vergne (1997); Spinnato, Fellhauer, & Portegies Zwart (2003) and Just *et al.* (2010).

While dynamical friction, as introduced by Chandrasekhar, is a mechanism of collisional stellar systems, galaxies are collisionless (at least, in the regime under discussion here). Indeed, since Chandrasekhar's time, it has become clear that there is a *collective* process which governs such phenomena as the decay of the orbit of a satellite galaxy in the halo of its parent galaxy (Tremaine & Weinberg 1984; Weinberg 1986; Colpi & Pallavicini 1998). It might even be concluded that the physical phenomenon which causes the orbital decay of satellite galaxies has no deeper connection with dynamical friction than the same dependence on the basic scales of density and velocity dispersion (which allows it to be expressed by a suitable choice of the Coulomb logarithm). Even more tenuous is the link between Chandrasekhar's theory and the decay of a satellite in a partly or purely *gaseous* medium, though this too is often referred to as "dynamical friction". While there is some danger of confusing the underlying physics, perhaps it is a measure of the appeal of Chandrasekhar's discovery that the term "dynamical friction" has been extended to encompass such a diversity of astrophysical processes.

Acknowledgments

I am very grateful to Don Goldsmith for permission to place his translation of Ambartsumian (1937) in the public domain.

References

Aarseth S. J., Hills J. G., 1972, A&A, 21, 255

Ambartsumian V. A., 1937, Russian Astron. J., 14, 207 (in Russian); transl. by D. Goldsmith, available at http://www.maths.ed.ac.uk/~heggie/Ambartsumian1937001.pdf

Ambartsumian V. A., 1938, An.. Leningrad State Univ., No.22, p.19; transl. in J. Goodman and P. Hut, eds, Dynamics of Star Clusters, IAUS113, D. Reidel, Dordrecht, p.521

Angeletti L., Capuzzo-Dolcetta R., Giannone P., 1983, A&A, 121, 183

Angeletti L., Dolcetta R., Giannone P., 1980, Ap&SS, 69, 45

Baumgardt H., 2001, MNRAS, 325, 1323

Baumgardt H., Hut P., Heggie D. C., 2002, MNRAS, 336, 1069

Baumgardt H., Makino J., 2003, MNRAS, 340, 227

Begelman M. C., Blandford R. D., Rees M. J., 1980, Natur, 287, 307

Berczik P., Merritt D., Spurzem R., 2005, ApJ, 633, 680

Berczik P., Merritt D., Spurzem R., Bischof H.-P., 2006, ApJ, 642, L21

Berentzen I., Preto M., Berczik P., Merritt D., Spurzem R., 2009, ApJ, 695, 455

Bertin G., Varri A. L., 2008, ApJ, 689, 1005

Binney J., Tremaine S., 1987, Galactic Dynamics, 1e, Princeton University Press, Princeton

Binney J., Tremaine S., 2008, Galactic Dynamics, 2e, Princeton University Press, Princeton

Boutloukos S. G., Lamers H. J. G. L. M., 2003, MNRAS, 338, 717

Chan R., Mamon G. A., Gerbal D., 1997, ApL&C, 36, 47

Chandrasekhar S., 1942, Principles of Stellar Dynamics, University of Chicago Press, Chicago
Chandrasekhar S., 1943a, ApJ, 97, 263
Chandrasekhar S., 1943b, ApJ, 98, 54
Chandrasekhar S., 1944, ApJ, 99, 54
Chandrasekhar S., 1960, Principles of Stellar Dynamics, Dover Publications Inc, New York
Chatterjee S., Fregeau J. M., Umbreit S., Rasio F. A., 2010, ApJ, 719, 915
Ciotti L., 2010, AIPC, 1242, 117
Cohn H., 1979, ApJ, 234, 1036
Cohn H., 1985, in J. Goodman and P. Hut, eds, Dynamics of Star Clusters, IAUS113, D. Reidel, Dordrecht, p.161
Colpi M., Pallavicini A., 1998, ApJ, 502, 150
Cora S. A., Muzzio J. C., Vergne M. M., 1997, MNRAS, 289, 253
Cruz-González C., Poveda A., 1971, Ap&SS, 13, 335
Evans N. W., 2011, BASI, 39, in press
Fellhauer M., Heggie D. C., 2005, A&A, 435, 875
Fregeau J. M., Chatterjee S., Rasio F. A., 2006, ApJ, 640, 1086
Fujii M., Funato Y., Makino J., 2006, PASJ, 58, 743
Fukushige T., Heggie D. C., 2000, MNRAS, 318, 753
Gao B., Goodman J., Cohn H., Murphy B., 1991, ApJ, 370, 567
Gieles M., Baumgardt H., 2008, MNRAS, 389, L28
Gieles M., Bastian N., Lamers H. J. G. L. M., Mout J. N., 2005, A&A, 441, 949
Giersz M., 2006, MNRAS, 371, 484
Giersz M., Heggie D. C., 2010, MNRAS, 1747
Goodman J. J., 1983, PhD Thesis, Princeton University
Gould A., 1991, ApJ, 379, 280
Gualandris A., Portegies Zwart S., Sipior M. S., 2005, MNRAS, 363, 223
Gurevich L. E., Levin B. Y., 1950, Astron. Zh., 27, 273 (in Russian; translation in NASA TT F-11, 541, NASA, Washington, 1968)
Gvaramadze V. V., Gualandris A., Portegies Zwart S., 2009, MNRAS, 396, 570
Hasani Zonoozi A., Kuepper A. H. W., Baumgardt H., Haghi H., Kroupa P., Hilker M., 2011, MNRAS, 411, 1989 (arXiv:1010.2210)
Hayli A., 1970, A&A, 7, 17
Heggie D. C., 1975, MNRAS, 173, 729
Heggie D. C., 2001a, ASPC, 228, 29
Heggie D. C., 2001b, in T. Ebisuzaki and J. Makino, eds, New Horizons of Computational Science, Kluwer, Dordrecht, p.59
Heggie D. C., Giersz M., 2008, MNRAS, 389, 1858
Heggie D. C., Giersz M., 2009, MNRAS, 397, L46
Heggie D. C., Giersz M., Spurzem R., Takahashi K., 1998, HiA, 11, 591
Heggie D. C., Hut P., 1993, ApJS, 85, 347
Heggie D., Hut P., 2003, The Gravitational Million Body Problem, Cambridge University Press, Cambridge
Heggie D. C., Hut P., McMillan S. L. W., 1996, ApJ, 467, 359
Heggie D. C., Ramamani N., 1995, MNRAS, 272, 317
Hénon M., 1960, AnAp, 23, 668
Hénon M., 1961, AnAp, 24, 369
Hénon M., 1967, Colloque "Les Nouvelles Méthodes de la Dynamique Stellaire", Editions CNRS, Paris
Hénon M., 1970, A&A, 9, 24

Hénon M. H., 1971, Ap&SS, 14, 151
Hills J. G., 1975, AJ, 80, 809
Hills J. G., 1988, Nature, 331, 687
Hills J. G., 1990, AJ, 99, 979
Hills J. G., Dissly R. W., 1989, AJ, 98, 1069
Hurley J. R., *et al.*, 2008, AJ, 135, 2129
Ideta M., 2002, in Ikeuchi S., Hearnshaw J., Hanawa T., eds, Proceedings of the IAU 8th Asian-Pacific Regional Meeting, Volume II, p. 255
Iwasawa M., An S., Matsubayashi T., Funato Y., Makino J., 2011, ApJ, 731, L9
Jeans J. H., 1918, MNRAS, 79, 100
Jiang Y.-F., Tremaine S., 2010, MNRAS, 401, 977
Johnstone D., 1993, AJ, 105, 155
Just A., Peñarrubia J., 2005, A&A, 431, 861
Just A., Khan F. M., Berczik P., Ernst A., Spurzem R., 2010, MNRAS, 1687
Kapoor R. C., 1985a, Ap&SS, 112, 347
Kapoor R. C., 1985b, Ap&SS, 117, 363
Kim C.-H., Chun M.-S., Min K. W., 1991, JASS, 8, 11
King I., 1958, AJ, 63, 109
King I., 1959, AJ, 64, 351
King I., 1960, AJ, 65, 122
King I. R., 1977, RMxAA, 3, 167
Kouwenhoven M. B. N., Goodwin S. P., Parker R. J., Davies M. B., Malmberg D., Kroupa P., 2010, MNRAS, 404, 1835
Kruijssen J. M. D., 2009, A&A, 507, 1409
Küpper A. H. W., Kroupa P., Baumgardt H., Heggie D. C., 2010, MNRAS, 407, 2241
Kuzmin G. G., 1957, Tartu Astron. Obs. Publ., 33, 75
Lamers H. J. G. L. M., Gieles M., 2006, A&A, 455, L17
Lamers H. J. G. L. M., Gieles M., Portegies Zwart S. F., 2005, A&A, 429, 173
Larson R. B., 1970, MNRAS, 147, 323
Laughlin G., Adams F. C., 1998, ApJ, 508, L171
Lemou M., Chavanis P.-H., 2010, PhyA, 389, 1021
McMillan S. L. W., Hut P., 1996, ApJ, 467, 348
Malmberg D., Davies M. B., Heggie D. C., 2010, arXiv:1009.4196
Mandushev G., Staneva A., Spasova N., 1991, A&A, 252, 94
Merritt D., 1981, PhD thesis, Princeton University
Merritt D., 1983, ApJ, 264, 24
Mikkola S., Valtonen M. J., 1992, MNRAS, 259, 115
Mitchell D. G. M., Heggie D. C., 2007, MNRAS, 376, 705
Oort J. H., 1950, BAN, 11, 91
Öpik E., 1932, Proc. Amer. Acad. Arts and Sc., 67, 169
Padmanabhan, T., 1996, Current Science, 70, 784
Polnarev A. G., Rees M. J., 1994, A&A, 283, 301
Quinlan G. D., 1996, NewA, 1, 35
Retterer J. M., King I. R., 1982, ApJ, 254, 214
Ross D. J., Mennim A., Heggie D. C., 1997, MNRAS, 284, 811
Saito H., 1976, A&A, 46, 171
Sari R., Kobayashi S., Rossi E. M., 2010, ApJ, 708, 605
Sesana A., 2010, ApJ, 719, 851
Sigurdsson S., Phinney E. S., 1993, ApJ, 415, 631

Sommariva V., Piotto G., Rejkuba M., Bedin L. R., Heggie D. C., Mathieu R. D., Villanova S., 2009, A&A, 493, 947

Spinnato P. F., Fellhauer M., Portegies Zwart S. F., 2003, MNRAS, 344, 22

Spitzer L., Jr., 1940, MNRAS, 100, 396

Spitzer L., Jr., Harm R., 1958, ApJ, 127, 544

Spitzer L., Jr., Hart M. H., 1971, ApJ, 164, 399

Spitzer L., Jr., Mathieu R. D., 1980, ApJ, 241, 618

Spitzer L., Jr., Shapiro S. L., 1972, ApJ, 173, 529

Spurzem R., Giersz M., Heggie D. C., Lin D. N. C., 2009, ApJ, 697, 458

Statler T. S., Ostriker J. P., Cohn H. N., 1987, ApJ, 316, 626

Stodolkiewicz J. S., 1982, AcA, 32, 63

Sugimoto D., Bettwieser E., 1983, MNRAS, 204, 19P

Tanikawa A., Fukushige T., 2005, PASJ, 57, 155

Tanikawa A., Fukushige T., 2010, PASJ, 62, 1215

Theuns T., 1996, MNRAS, 279, 827

Tremaine S., Weinberg M. D., 1984, MNRAS, 209, 729

Valtonen M., Karttunen H., 2006, The Three Body Problem, Cambridge University Press, Cambridge

Varri A. L., Bertin G., 2009, ApJ, 703, 1911

Vecchio A., Colpi M., Polnarev A. G., 1994, ApJ, 433, 733

Vesperini E., Heggie D. C., 1997, MNRAS, 289, 898

von Hoerner S., 1960, ZA, 50, 184

Wasserman I., Weinberg M. D., 1987, ApJ, 312, 390

Weinberg M. D., 1986, ApJ, 300, 93

Weinberg M. D., Shapiro S. L., Wasserman I., 1987, ApJ, 312, 367

Wielen R., 1968, BAst, 3, 127

Yabushita S., 1966, MNRAS, 133, 133

Yu Q., Tremaine S., 2003, ApJ, 599, 1129

Zhang F., Lu Y., Yu Q., 2010, ApJ, 722, 1744

Chandrasekhar and modern stellar dynamics

N. W. Evans[*]

Institute of Astronomy, Madingley Rd, Cambridge, CB3 0HA, UK

Received February 28, 2011; accepted March 21, 2011

Abstract. Stellar dynamics occupied Chandrasekhar's interest for a brief interlude between his more prolonged studies of stellar structure and radiative transfer. This paper traces the history of one of his ideas – namely, that the shape of the galactic potential controls the orientation of the stellar velocity dispersion tensor. It has its roots in papers by Eddington (1915) and Chandrasekhar (1939), and provoked a fascinating dispute between these two great scientists – less well-known than their famous controversy over the white dwarf stars. In modern language, Eddington claimed that the integral curves of the eigenvectors of the velocity dispersion tensor provide a one-dimensional foliation into mutually orthogonal surfaces. Chandrasekhar challenged this, and explicitly constructed a counter-example. In fact, the work of neither of these great scientists was without flaws, though further developments in stellar dynamics were to ultimately draw more on Eddington's insight than Chandrasekhar's. We conclude with a description of modern attempts to measure the orientation of the velocity dispersion tensor for populations in the Milky Way Galaxy, a subject that is coming into its own with the dawning of the age of precision astrometry.

Keywords : celestial mechanics – stellar dynamics – galaxies: kinematics and dynamics – galaxies: general – Galaxy: stellar populations

1. Introduction

Chandrasekhar was perhaps the most influential theoretical astrophysicist of his time. This influence was particularly felt through an outstanding series of research monographs that continue to be read today. In fact, most astronomers first encounter Chandrasekhar through the cheap Dover reprints of books like *Stellar Structure*, *Radiative Transfer*, *Hydrodynamic and Hydromagnetic Stability* and *Ellipsoidal Figures of Equilibrium*. These books bristle with formulae, equations, numerical tables, graphs and historical notes, leavened with an immaculate prose style. They make exciting reading still today because they contain so much classic astrophysics so lucidly explained.

In his Nobel lecture, Chandrasekhar (1984) wrote "*There have been seven periods in my life. They are briefly: 1) stellar structure, including the theory of white dwarfs*

[*] e-mail: nwe@ast.cam.ac.uk

(1929-1939); 2) stellar dynamics, including the theory of Brownian motion (1938-1943); 3) the theory of radiative transfer, the theory of the illumination and the polarization of sunlit sky (1943-1950); 4) hydrodynamic and hydromagnetic stability (1952-1961); 5) the equilibrium and stability of ellipsoidal figures of equilibrium (1961-1968); 6) the general theory of relativity and relativistic astrophysics (1962-71) and 7) the mathematical theory of black holes (1974-1983)."

So, Chandrasekhar's work on stellar dynamics occupied a brief interlude of time. It began in 1938 as a natural progression of his interests in the structure and evolution of stars. This was at the height of his famous controversy with Eddington over the fate of the white dwarf stars. It was over by 1943, when Chandrasekhar was commuting between his professorship at Yerkes Observatory and the University of Chicago, and the Aberdeen Proving Ground in Maryland, working on ballistics as part of the war effort. His research interests had moved towards radiative transfer – the subject which Chandrasekhar himself has described as the one giving him most satisfaction (Wali 1990).

Chandrasekhar's (1943) book *Principles of Stellar Dynamics* is not as well-known or as magisterial as some of his others. The work on dynamical friction and dynamics of star clusters has proved to be of long-lasting value (see e.g., Heggie's article in this issue). However, much of the book reads oddly today. There are two long and, to modern eyes, puzzling chapters devoted to problems in collisionless stellar dynamics, in particular, galaxy models consistent with the ellipsoidal hypothesis. This term is not much used nowadays, but was introduced by Eddington (1915) as a generalisation of the triaxial Gaussian distribution of velocities used by Schwarzschild (1908) to describe the velocities of stars in the solar neighbourhood. This is the work we shall examine here, and it is fair to say that this is not Chandrasekhar at his most memorable. But, its connection with the earlier work of Eddington is fascinating, especially considering the personal relations between these two great scientists. And even when Chandrasekhar was not at his brilliant best, he could still find much of interest that others had overlooked.

So, we shall trace out the twists and turns that take us from the founding of stellar dynamics by Jeans and Eddington at the beginning of the twentieth century to modern times. Chandrasekhar himself contributed both fresh footpaths and blind alleys to this mazy route.

2. Eddington and the ellipsoidal hypothesis

Eddington's (1915) paper that studies the ellipsoidal hypothesis is one of his great ones. We can do no better than use Eddington's own words:

"At any point of the system, the directions of the axes of the velocity ellipsoid determine three directions at right angles. The velocity ellipsoids thus define three orthogonal families of curves, each curve being traced by moving step by step always in the direction of an axis of the velocity ellipsoid at the point reached. These curves may be regarded as the intersections of a triply orthogonal family of surfaces, which we shall call the principal velocity surfaces. The axes of the velocity ellipsoid at any point are normals to the three principal velocity surfaces through any point".

In modern language, the theory of collisionless systems such as galaxies begins with the Boltzmann equation:

$$\frac{\partial F}{\partial t} + \mathbf{v} \cdot \frac{\partial F}{\partial \mathbf{x}} - \frac{\partial \Phi}{\partial \mathbf{x}} \cdot \frac{\partial F}{\partial \mathbf{v}} = 0, \tag{1}$$

where F is the phase space distribution function and Φ is the gravitational potential. At every point in the galaxy, we can define a velocity dispersion tensor

$$\sigma_{ij} = \langle (v_i - \langle v_i \rangle)(v_j - \langle v_j \rangle) \rangle, \tag{2}$$

where angled brackets denote averages over the distribution function. The velocity dispersion tensor σ_{ij} is real and symmetric, and therefore by a well-known theorem in linear algebra has mutually orthogonal eigenvectors. Eddington is asserting that the integral curves of the eigenvectors provide a one-dimensional foliation into surfaces, which he calls *the principal velocity surfaces.* We shall return to the assumptions underlying this assertion shortly, as it is precisely the point that troubled Chandrasekhar.

Eddington then shows via Lagrange's equations that a steady state distribution of stars moving in a gravitational potential Φ necessarily generates principal velocity surfaces that are confocal quadrics. Labelling the quadric surfaces by (λ, μ, ν), these are recognised as ellipsoidal coordinates (e.g., Morse & Feshbach 1953). Eddington now proves two further theorems. First, suppose that the distribution of velocities has exactly the Schwarzschild (1908) or triaxial Gaussian form

$$F \propto \exp\left(-\frac{v_\lambda^2}{2\sigma_\lambda^2} - \frac{v_\mu^2}{2\sigma_\mu^2} - \frac{v_\nu^2}{2\sigma_\nu^2} \right), \tag{3}$$

where $(v_\lambda, v_\mu, v_\nu)$ are velocity components referred to the locally orthogonal axes and $(\sigma_\lambda, \sigma_\mu, \sigma_\nu)$ are the semiaxes of the velocity ellipsoid. This is the ellipsoidal hypothesis. Eddington showed that the only solutions for the principal velocity surfaces are spheres. However, the gravitational potential need not be spherical, but can take the general form

$$\Phi(r, \theta, \phi) = f(r) + \frac{g(\theta)}{r^2} + \frac{h(\phi)}{r^2 \sin^2 \theta}, \tag{4}$$

where f, g and h are arbitrary functions of the indicated arguments. These have sometimes been called Eddington potentials in the astronomical literature.

Secondly, Eddington considered the more general case of a stellar population with an arbitrary distribution of velocities. Under the assumption of the existence of principal velocity surfaces, he showed that the potential can take the general form in ellipsoidal coordinates

$$\Phi(\lambda, \mu, \nu) = \frac{f(\lambda)}{(\lambda - \mu)(\lambda - \nu)} + \frac{g(\mu)}{(\mu - \lambda)(\mu - \nu)} + \frac{h(\nu)}{(\nu - \mu)(\nu - \lambda)}. \tag{5}$$

Eddington does not consider the fully triaxial case in detail, but he does study the degenerations of the ellipsoidal coordinates into spheroidal coordinates. Here, the stars have oblate or prolate density distributions, the principal velocity surfaces are prolate or oblate spheroids and the velocity dispersion tensor is in general anisotropic. This was the first attempt to build galaxy models using the separable potentials. Except in the spherical limit, Eddington did not write down the form of the integrals of motion, leaving that task to his student, G.L. Clark (1937).

Although Eddington's paper is not without its flaws, it turned out to be remarkably prescient, anticipating developments over half a century later.

3. Chandrasekhar's criticism

In retrospect, Chandrasekhar's venture into stellar dynamics seems both natural and brave. It is natural, as it is an obvious progression of his interests in stellar structure and evolution. It is brave, as it strays onto territory that Eddington had already made his own. The discipline had been founded by two people – Eddington in his book *Stellar Movements and the Structure of the Universe* published in 1914, and Jeans in his 1917 Adams Prize essay, published somewhat later in 1919 as *Problems of Cosmogony and Stellar Dynamics*. Eddington and Jeans had dominated the subject over the 1920s, with fundamental contributions, including Jeans' theorem, the equations of stellar hydrodynamics (sometimes called the Jeans' equations), and Eddington's inversion formula for the distribution function of a spherical galaxy. Given Chandrasekhar's worsening relationship with Eddington over these years, his incursions into this field were almost inevitably opening up a second front.

Chandrasekhar (1939, 1940) announced his entry into the field with two gigantic papers on the ellipsoidal hypothesis (summarised in Chapters 3 and 4 of *Principles of Stellar Dynamics* which themselves occupy over a hundred pages). Right away, he detected an error in Eddington's paper. Chandrasekhar's criticism is worth quoting in full:

"The fallacy in Eddington's argumentation is clear. It is true that we can regard the directions of the principal axes of the velocity ellipsoid at any given point as being tangential to the three curves which intersect orthogonally at the point considered. But it is not generally true that we can regard these curves as the intersections of a triply orthogonal system of surfaces. Consequently, the notion of principal velocity surfaces introduces severe restrictions on the problem, which are wholly irrelevant and certainly unnecessary."

Here, Chandrasekhar is completely correct. Eddington assumed that the eigenvectors of the velocity dispersion tensor are the tangent vectors of a triply orthogonal system of surfaces. This is a sufficient, but not a necessary, consequence of the orthogonality of the eigenvectors of the dispersion tensor. Eddington (1943) himself conceded as much in his review of Chandrasekhar's book. Writing in the journal *Nature*, he stated:

"Chandrasekhar rightly points out a fallacy in a theorem which I gave in 1915 and the correction makes the conclusion less general than has hitherto been assumed. But he does not take the opportunity of restating the position. Presumably it is still true that in a steady system with axial symmetry, the velocity surfaces are confocal quadrics and transverse star streaming is necessarily excluded, but there is no mention of this".

Where did Chandrasekhar's insight lead? Chandrasekhar first somewhat generalised the problem by asking for stellar dynamical models with distribution functions F of the form

$$F = F(Q), \tag{6}$$

where Q is a quadratic function of the velocities. The coefficients are arbitrary functions of position. More formally,

$$Q = \mathbf{v} \cdot \mathbf{M}(\mathbf{x}) \cdot \mathbf{v} + N(\mathbf{x}), \tag{7}$$

where $\mathbf{M}$ and N are matrix and scalar functions of position. This is a generalized ellipsoidal hypothesis, as Q and hence the phase space density F is constant on ellipsoids in velocity space.

Chandrasekhar proceeds by substituting his ansatz for the distribution function into the Boltzmann equation and separating term by term in the powers of velocity. He extracts a set of 20 partial differential equations, which he reduces to 6 integrability conditions. Note that Chandrasekhar does not impose the Poisson equation, as he is interested in finding the conditions that a stellar population has a distribution function of ellipsoidal form in an externally imposed potential. He reaches a very surprising conclusion that *for stellar systems in a steady state, the potential Φ must necessarily be characterised by helical symmetry. The case of axial symmetry is included as a special case.*

In other words, using cylindrical polar coordinates (R, ϕ, z), Chandrasekhar asserts that the only solutions for the gravitational potential compatible with the generalised ellipsoidal hypothesis are

$$\Phi = f(R, z + \alpha\phi), \tag{8}$$

where f is an arbitrary function of the indicated arguments and α is a constant (the reciprocal of the pitch of the helix). The integrals of motion are the energy E and the generalisation of the angular momentum component, namely

$$I = p_\phi - \alpha p_z \tag{9}$$

where p_ϕ and p_z are the canonical momenta conjugate to ϕ and z. Chandrasekhar then notes that such a potential can have bound orbits only if it is axisymmetric ($\alpha = 0$) and so he reaches his final conclusion. *For stellar systems with differential motions, which are in steady states and of finite spatial extent, the potential Φ must necessarily be characterized by axial symmetry*.

This is a strong claim and we shall shortly see that, like Eddington's work, it is not entirely correct. A surprising aspect is that, having realised that Eddington had introduced unnecessarily restrictive assumptions into the problem, Chandrasekhar is not troubled by that fact that his more general approach finds fewer solutions than Eddington – and indeed doesn't find the solutions with quadric principal velocity surfaces at all! Even more curiously, Chandrasekhar recognises that the phase space distribution F is an integral of motion, quoting Whittaker's (1936) book on *Analytical Dynamics* as a reference. He therefore knows that his problem is exactly equivalent to seeking all potentials that admit integrals of motion quadratic in the velocities. But, this problem is also (partly) solved in Whittaker's book, which provides a derivation of the separable potentials in spheroidal coordinates, though not ellipsoidal, from the assumption of quadratic integrals.

A new result of Chandrasekhar is that he provides an explicit counter-example to Eddington's assumption. The helically symmetric systems indeed remain the only ones known to us which do not possess mutually orthogonal principal velocity surfaces, but do satisfy the ellipsoidal hypothesis. They are not of much astrophysical interest as they do not resemble galaxies, but they remain of considerable intellectual interest.

Another insight of Chandrasekhar that has proved its worth is his repeated emphasis on the principle of equivalence. By this, he means that if several different models can be found sharing the same gravitational potential, then a more complex model that does not satisfy the ellipsoidal hypothesis can be built by weighted linear superposition. This idea has often been exploited in modern times to build realistic models by superpositions of analytic distribution functions (e.g., Fricke 1952; Dejonghe 1989; Emsellem, Dejonghe & Bacon 1999).

4. A modern approach

Let us now state and give the solution to Chandrasekhar's problem anew from the point of view of a modern dynamicist. Jeans' theorem tells us that the distribution function of a collisionless system depends only on the globally defined, isolating, integrals of motion. It therefore follows that Q must be an integral of motion. Chandrasekhar's problem is exactly equivalent to identifying all those potentials that admit integrals of motion at most quadratic in the velocities. This is a problem of widespread interest in both mathematics and physics, with an enormous literature and history.

Integrals of motion that are linear in the velocities always result from geometric symmetries of space. This is sometimes called Noether's theorem (see e.g., Landau & Lifshitz 1976; Arnold 1978). It follows from the fact that the Lagrangian is invariant with respect to the corresponding transformations, which are linear in the generators of the Euclidean group of symmetries. Examples include the invariance of the angular momentum component p_ϕ in axisymmetric potentials $\Phi(R, z)$, and the invariance of the linear momentum component p_z in translationally invariant potentials $\Phi(x, y)$. Chandrasekhar's helical solution is the most general possible, with rotationally and translationally invariant potentials given by the limits $\alpha \to 0$ and $\alpha \to \infty$ respectively.

Integrals of motion that are quadratic in the velocities always result from separability of the Hamilton-Jacobi equation in some coordinate system. Many authors discovered some or all of the potentials for which the Hamilton-Jacobi equation is separable in the confocal ellipsoidal coordinates or their degenerations (e.g., Eddington 1915; Weinacht 1924; Whittaker 1936; Clark 1937; Eisenhart 1948; Lynden-Bell 1962). These systems possess integrals of the motion quadratic in the velocities by construction, as the Hamilton-Jacobi equation only has such terms in it! The fact that separability of the Hamilton-Jacobi equation is both a necessary and sufficient condition is a much more difficult result to prove. It was done for the first time in Makarov *et al.* (1967).

Although written from the viewpoint of particle physicists, Makarov *et al.* (1967) follow essentially the same route as Chandrasekhar in Chapter 3 of *Principles of Stellar Dynamics*. That is, they ask for the Poisson bracket of the integral of motion Q with the Hamiltonian H to vanish. This is mathematically identical to requiring that Q satisfy the collisionless Boltzmann equation, as Chandrasekhar did. The main difference is that Makarov *et al.* substantially simplify Q by rotations and translations, before requiring that Q commute with the Hamiltonian H. This considerably reduces the mathematical complexity of the problem, enabling them to find all possible solutions (including the separable ones that Chandrasekhar had missed).

Before passing to later developments, it is worth remembering that Chandrasekhar and Eddington had disagreed over the white dwarf stars and the endpoints of stellar evolution (see Vibert Douglas 1956; Wali 1997; Chandrasekhar 1988 for various perspectives on this affair). In retrospect, it is clear that Eddington behaved badly over the white dwarfs, not so much because he was wrong – that can (and should) happen to every scientist! – but because he used his seniority to stifle the work of a younger colleague.

Is it possible that Chandrasekhar, hurt by the reception of what would ultimately prove to be a Nobel Prize winning achievement, was unable to appreciate fully the advantages in Eddington's approach in stellar dynamics? True, he had detected an error in Eddington's (1915) paper, but Eddington in the end saw closer to the truth of the matter in stellar dynamics. Eddington introduced a hypothesis – the principal velocity surfaces – that was not strictly-speaking necessary and would ultimately be discarded by later scientists. But, it proved to be a physically fruitful hypothesis that led Eddington to an important class of models. Consequently, later developments were to build more on Eddington's work than Chandrasekhar's, as we will now see.

5. Later developments

Further advances had to wait till the late fifties and early sixties, when the subject was revived by Lynden-Bell (1962) with a particularly original investigation. Rather than starting from an assumption that the integrals are quadratic in the velocities, Lynden-Bell permitted the integrals to have any form (polynomial or transcendent). Instead, he assumed that the steady-state is one of a set through which the system may secularly evolve whilst preserving the existence of the integrals of motion. This led to the enumeration of all potentials with such isolating integrals – prominent among them being the separable potentials in ellipsoidal coordinates and their degenerations. At the time, the flattening of elliptical galaxies was believed to be caused primarily by rotation rather than velocity anisotropy. Hence, the application of the potentials to galaxies remained unexplored in the West.

This was not true in the former Soviet Union, as a remarkable and sadly neglected paper by Kuzmin (1956) – citing the influences of Eddington (1915) and Clark (1937) – had already used the separable potentials in spheroidal coordinates to build an oblate, axisymmetric model of the Galaxy. Kuzmin (1973) was also the first to write down the fully triaxial case, and study its orbital structure, identifying the 4 characteristic classes of orbits: box, inner and outer long axis tubes and short axis tubes (see e.g., Binney & Tremaine 1987). These models became well-known in the West only after they had been re-discovered and extended by de Zeeuw (1985). Kuzmin (1973) and de Zeeuw (1985) showed that an ellipsoidally stratified model with density

$$\rho = \frac{\rho_0}{(1+m^2)^2}, \qquad m^2 = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \tag{10}$$

possesses an exactly separable gravitational potential in confocal ellipsoidal coordinates. The easiest way to demonstrate this is by making use of the methods and formulae in Chandrasekhar's (1968) finest and most beautiful book, *Ellipsoidal Figures of Equilibrium*. This is an important result as it showed that realistic and physically motivated models of elliptical galaxies could be built from separable potentials. De Zeeuw also demonstrated a number of beautiful properties of these models, including the classification of their orbits in integral and action space.[1] This led to a flowering of interest in the models, as evidenced by the papers in *IAU Symposium 127: The Structure and Dynamics of Elliptical Galaxies* (de Zeeuw 1987). This can be seen as the culmination of over seventy years of astronomical research on the subject, from Eddington, through Chandrasekhar, Lynden-Bell and Kuzmin to modern times.

Even though their mass density falls off faster than the luminosity density of giant ellipticals, and even though they are cored in the central parts rather than cusped, the separable models still occupy a special place in modern galactic dynamics. This is because the orbital structure of the models is generic for all flattened triaxial systems without figure rotation. Although the models do not contain any irregular or chaotic orbits, for many applications in galactic dynamics, this is unimportant, as the fraction of phase space occupied by truly irregular orbits is believed to be small (Goodman & Schwarzschild 1981).

6. The alignment of the velocity dispersion tensor

Modern interest in the subject (e.g., Smith, Evans & An 2009a,b; Binney & McMillan 2011) has been given additional impetus by large-scale photometric and spectroscopic surveys of hundreds of thousands of stars in the Milky Way Galaxy itself. If proper motions are also available, then this raises the possibility that all the components of the velocity dispersion tensor can be computed directly from the data. There have been a number of interesting recent attempts to do this, both for halo and disk populations. Although sample sizes are presently still small, and distance errors a serious hazard, matters will substantially improve in the next few years.

For example, the *Sloan Digital Sky Survey* (SDSS, York *et al.* 2000) carried out repeated photometric measurements in an equatorial stripe, known as Stripe 82, primarily with the aim of supernova detection. Bramich *et al.* (2008) then provided a public archive

[1] By now, these potentials had come to be known as Stäckel potentials in the astronomical literature. This seems unwarranted. First, it is poor practice in physics to associate a name with an equation if a perfectly adequate descriptive term exists. On these grounds alone, the term 'separable potential' is preferable to 'Stäckel potential'. And, second, there is no reason to associate the name of Paul Stäckel with coordinate systems and potentials that he never wrote down! Stäckel was a prominent differential geometer, later Professor of Mathematics at Heidelberg. In his *Habilitationschrift* in 1891 at Halle, Stäckel wrote down the condition for the Hamilton-Jacobi equation to separate in a given coordinate system on a general Riemannian manifold in the form of the vanishing of a determinant (which has reasonably enough come to be called the Stäckel determinant). Stäckel did not derive the coordinate systems in Euclidean 3-space for which his determinant vanishes, far less the form of the separable potentials in these coordinates. This work was left to Weinacht (1924) and Eisenhart (1948). In fact, Stäckel's result is limited, as it does not even provide a comprehensive test for separability. Stäckel's determinant for a separable system only vanishes if it is written down in the separable coordinate system itself. The finding of a general criterion for identifying whether a potential is separable in some coordinate system remains an outstanding research problem.

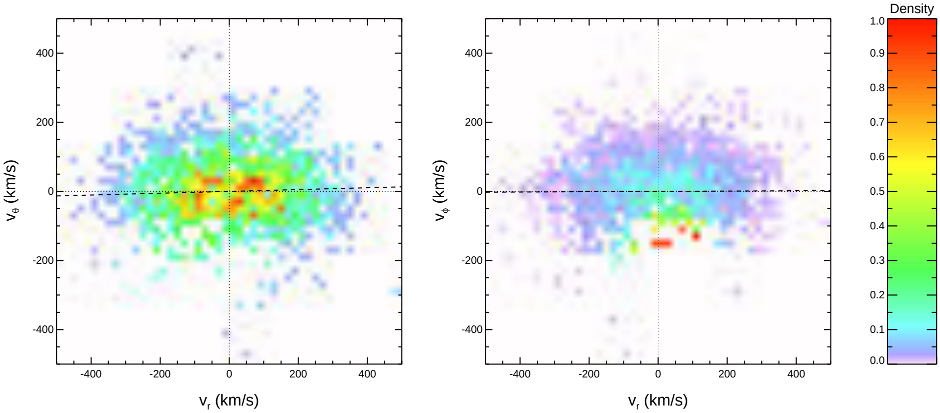


Figure 1. The efficiency corrected velocity distributions in the (v_r, v_θ) and (v_r, v_ϕ) planes for the sample of 1,600 subdwarfs with 1 kpc $< |z| <$ 4 kpc. The dashed lines show the orientation of the tilts, which are very close to spherical alignment. The apparent non-Gaussianity in the (v_r, v_ϕ) distribution is due to the variation of the efficiency correction across this plane. [From Smith et al. (2009b)].

of light-motion curves in Stripe 82 complete down to magnitude 21.5 in the u, g, r and i photometric bands, and to magnitude 20.5 in z. This reaches almost 2 magnitudes fainter than the SDSS/USNO-B catalogue (Munn *et al.* 2004), making it the deepest large-area photometric and astrometric catalogue available. Smith *et al.* (2009a,b) extracted a sample of $\sim$1,600 halo subdwarf stars via a reduced proper motion diagram. Their radial velocities are calculated from the SDSS spectra and their distances are estimated from photometric parallaxes, thus giving the full phase space information. Although the sample is not kinematically unbiased, the detection efficiency can be calculated and corrections made for any biases.

Figure 1 shows the velocity distributions of the SDSS Stripe 82 subdwarfs. These halo stars lie at Galactocentric cylindrical polar radii between 7 and 10 kpc, and at depths of 4.5 kpc or less below the Galactic plane. The good alignment of the velocity ellipsoid in spherical polars is already apparent from the velocity distributions in the $(v_r, v_\theta\ )$ and (v_r, v_ϕ) planes. Smith *et al.* find that the velocity dispersion tensor of the halo subdwarfs has semiaxes $(\sigma_r, \sigma_\phi, \sigma_\theta)$ = $(143 \pm 2, 82 \pm 2, 77 \pm 2)$ km s^{-1}. The misalignment from the spherical polar coordinate surfaces can then be described by the correlation coefficients and the tilt angles using

$$\mathrm{Corr}[v_i, v_j] = \frac{\sigma_{ij}^2}{(\sigma_{ii}^2 \sigma_{jj}^2)^{1/2}}, \qquad (11)$$

and

$$\tan(2\alpha_{ij}) = \frac{2\sigma_{ij}^2}{\sigma_{ii}^2 - \sigma_{jj}^2}. \qquad (12)$$

The tilt of the velocity ellipsoid with respect to the spherical polar coordinate system is found to be consistent with zero for two of the three tilt angles, and very small for the third. Specifically, Smith *et al.* find:

$$\begin{aligned}\mathrm{Corr}[v_r, v_\theta] &= 0.078\pm0.029, & \alpha_{r\theta} &= 3\overset{\circ}{.}4\pm1\overset{\circ}{.}3,\\ \mathrm{Corr}[v_r, v_\phi] &= -0.028\pm0.039, & \alpha_{r\phi} &= -2\overset{\circ}{.}2\pm3\overset{\circ}{.}3,\\ \mathrm{Corr}[v_\phi, v_\theta] &= -0.087\pm0.047, & \alpha_{\phi\theta} &= -37\overset{\circ}{.}4\pm20\overset{\circ}{.}4.\end{aligned} \tag{13}$$

In Eddington's language, these stars have spherical principal velocity surfaces to an excellent approximation. In a slight extension of the earlier results of Eddington (1915) and Chandrasekhar (1939), Smith *et al.* (2009b) prove that: *If the potential is nonsingular, it is a sufficient condition for spherical symmetry that one of the non-degenerate eigenvectors of the velocity dispersion tensor is aligned radially everywhere.*

Of course, Smith *et al.* (2009b) did not demonstrate that the velocity dispersion tensor is aligned everywhere in spherical polar coordinates. They showed that the alignment is very close to spherical for halo subdwarfs at heliocentric distances of < 5 kpc along the ~ 250 deg^2 covered by SDSS Stripe 82. Nonetheless, they argued that this is still a striking and unexpected result over a range of Galactic locations that provides a new line of attack on the awkward question of the shape of the Milky Way's dark halo. Binney & McMillan (2011) concur that local measurements are not enough to constrain the shape of the Galaxy's potential. Further work on the alignment of the velocity ellipsoid of halo populations is highly desirable.

By contrast, the behaviour of the velocity ellipsoid of disk populations has been more widely studied, not least because of its importance for calculations of the asymmetric drift and the Oort Limit. Based on evidence from orbit integrations, Binney & Tremaine (1987) suggest that the tilt may lie midway between spherical and cylindrical polar alignment. This is also the expectation from models based on potentials separable in spheroidal coordinates (Statler 1989). There have been three recent determinations directly from data by Siebert *et al.* (2008), Fuchs *et al.* (2009) and Smith, Evans & Whiteoak (2011).

Siebert *et al.* (2008) extracted 763 red clump stars from the *Radial Velocity Experiment* dataset (RAVE, Zwitter *et al.* 2008), spanning a distance interval from the Sun of 500 to 1500 pc. The tilt of the velocity ellipsoid of stars so close to the Galactic plane is affected both by the structure of the Galactic disk and and the flattening of the dark halo. Siebert *et al.* find that the velocity ellipsoid is tilted towards the Galactic plane with an inclination of $7\overset{\circ}{.}3 \pm 1\overset{\circ}{.}8$. This is entirely consistent with alignment in spherical polar coordinates. Siebert *et al.* compare this value to computed inclinations for two mass models of the Milky Way. The measurement is consistent with a short scalelength of the stellar disc (≈ 2 kpc) if the dark halo is oblate or with a long scalelength (≈ 3 kpc) if the dark halo is spherical or prolate.

Fuchs *et al.* (2009) used an enormous sample of ~ 2 million M dwarfs derived from the *Sloan Digital Sky Survey* Data Release 7 (Abazajian *et al.* 2009). Although the proper motions and photometric distances of these stars are available, unfortunately the radial velocities are not. Fuchs *et al.* estimated the radial velocities via the method of deprojection of proper motions. They found an anomalously large tilt reaching an inclination of 20° at heights above the Galactic plane of 800 pc, whereas spherical alignment would predict an inclination of $\approx 5^\circ$. McMillan & Binney (2009) have argued that this surprisingly large

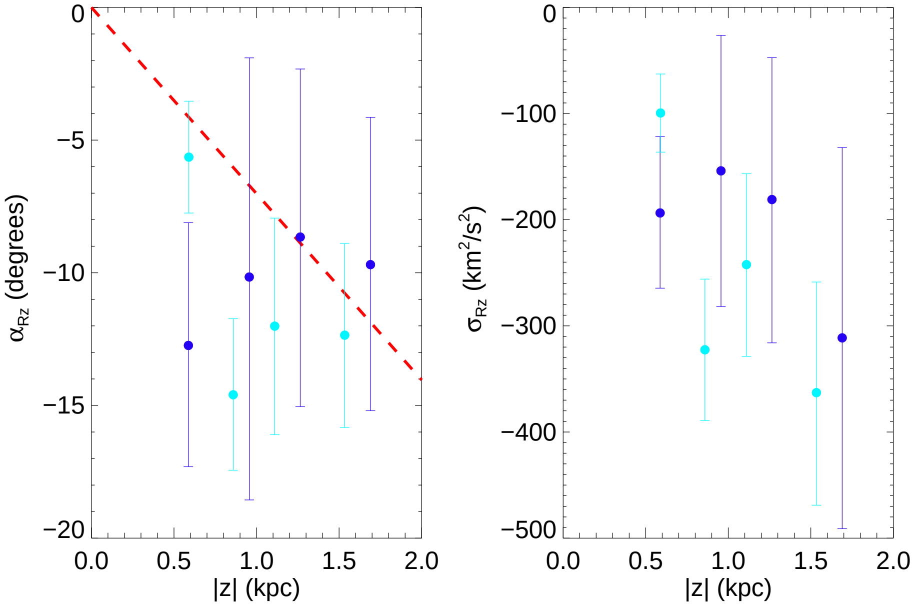


Figure 2. The variation of σ_{Rz} and the corresponding angle α_{Rz} as a function of height from the plane. The dashed red line is the assumed halo tilt (i.e. aligned in spherical polars). The blue and cyan points correspond to disc stars with metallicities $-0.8 \leq [\mathrm{Fe/H}] \leq -0.5$ and $[\mathrm{Fe/H}] \geq -0.5$, respectively. [From Smith *et al.* (2011)].

value may be spurious, a consequence of correlations between velocities and positions of stars, which renders the method of deprojection invalid.

Finally, Smith *et al.* (2011) again use the very deep light-motion catalogue for Stripe 82 (Bramich *et al.* 2008) to extract a sample of disk stars, complete with radial velocities from SDSS spectra and photometric metallicities. These stars are confined to a narrow range of cylindrical polar radius between $7 \leq R \leq 9$ kpc. However, there are enough stars to split the data into three ranges in metallicity ($-1.5 \leq [\mathrm{Fe/H}] \leq -0.8$, $-0.8 \leq [\mathrm{Fe/H}] \leq -0.5$ and $-0.5 \leq [\mathrm{Fe/H}]$), and for each metallicity bin to divide the data into four ranges in z ($0 \leq |\mathrm{z}| \leq 0.8$, $0.8 \leq |\mathrm{z}| \leq 1.1$, $1.1 \leq |\mathrm{z}| \leq 1.5$ and $1.5 \leq |\mathrm{z}| \leq 2.2$ kpc). This gives around 500 to 800 stars per bin. The variation with height and metallicity is shown in Figure 2. The dotted line corresponds to what we would expect for a velocity ellipsoid aligned in spherical polar coordinates. The metal-rich and medium-metallicity stars are arguably consistent with the dotted line, and hence consistent with the result of Siebert *et al.* (2008). In general, the stars in the lowest metallicity bins (not plotted in Figure 2) exhibit tilt angles which are larger than this, albeit with very large error bars.

Fortunately, the very-near future comprises the dawning of the Age of Precision Astrometry. The GAIA satellite (e.g., Gilmore 2007) will provide tangential velocities for 44 million stars and distances for 21 million stars with an accuracy better than 1 per cent. There is therefore a realistic prospect that the behaviour of the velocity ellipsoid for both disk and halo populations over a swathe of locations in the Milky Way Galaxy will be known shortly.

Acknowledgments

I am grateful to Donald Lynden-Bell and Tim de Zeeuw for first introducing me to this fascinating subject. I thank too my collaborators Jin An, Martin Smith and Hannah Whiteoak for their help and insight.

References

Abazajian K., *et al.*, 2009 ApJS, 182, 543

Arnold V.I., 1978, Mathematical Methods of Classical Mechanics, Springer

Binney J., Tremaine S., 1987, Galactic Dynamics, Princeton

Binney J., McMillan P., 2011, MNRAS, submitted (arXiv:1101.0747)

Bramich D., *et al.*, 2008, MNRAS, 386, 887

Chandraskehar S., 1939, ApJ, 90, 1

Chandraskehar S., 1940, ApJ, 92, 441

Chandrasekhar S., 1943, Principles of Stellar Dynamics, University of Chicago Press

Chandrasekhar S., 1968, Ellipsoidal Figures of Equilibrium, University of Yale Press

Chandrasekhar S., 1984, On Stars, Their Evolution and Their Stability, Nobel Foundation

Chandrasekhar S., 1988, Eddington: The Most Distinguished Astrophysicist of his Time, Cambridge University Press

Clark G.L., 1937, MNRAS, 97, 182

de Zeeuw P.T., 1985, MNRAS, 216, 273

de Zeeuw P.T., 1987, IAU Symposium 127: The Structure and Dynamics of Elliptical Galaxies, Reidel

Dejonghe H., 1989, ApJ, 343, 113

Eddington A.S., 1914, Stellar Movements and the Structure of the Universe, Macmillan.

Eddington A.S., 1915, MNRAS, 76, 37

Eddington A.S., 1943, Nature, 151, 91

Eisenhart L.P., 1948, Phys. Rev, 74, 87

Emsellem E., Dejonghe H., Bacon R., 1999, MNRAS, 303, 495

Fricke W., 1952, Ast Nach., 280, 193

Fuchs B., *et al.*, 2009, AJ, 137, 4149

Gilmore G., 2007, In Exploring the Cosmic Frontier: Astrophysical Instruments for the 21st Century., p. 205

Goodman J., Schwarzschild M., 1981, ApJ, 245, 1087

Jeans J.H., 1919, Problems of Cosmogony and Stellar Dynamics, Cambridge University Press

Kuzmin G.G., 1956, Astr. Zh, 33, 27

Kuzmin G.G., 1973, In Dynamics of Galaxies and Clusters, Materials of the All Union Conference in Alma Ata, p. 71, English translation in de Zeeuw (1987)

Landau L.D., Lifshitz E.M., 1976, Mechanics, Pergamon

Lynden-Bell D., 1962, MNRAS, 124, 95

Makarov A.A., Smorodinsk J.A., Valiev Kh., Winternitz P., 1967, Nuovo Cimento, 52, 1061

McMillan P., Binney J., 2009, MNRAS, 400, L103

Morse P., Feshbach H., 1953, Methods of Mathematical Physics, McGraw-Hill

Munn J.A., *et al.*, 2004, AJ, 127, 3034

Schwarzschild K., 1908, Göttingen Nachriten, p.191

Siebert A., *et al.*, 2008, MNRAS, 391, 793

Smith M.C., *et al.* 2009a, MNRAS, 399, 1223

Smith M.C., Evans N.W., An J.H. 2009b, ApJ, 698, 1110

Smith M.C., Evans N.W., Whiteoak H., 2011, MNRAS, submitted.

Stäckel P., 1891, Uber die Integration der Hamilton-Jacobischen Differential Gleichunger mittelst Separation der Variabeln, Habilitationschrift, Halle

Statler T.S, 1989, ApJ, 375, 544

Vibert Douglas A., 1956, The Life of Arthur Stanley Eddington, Nelson

Wali K.C., 1990, Chandra: a biography of S. Chandrasekhar, University of Chicago Press, Chicago

Wali K.C., 1997, S. Chandrasekhar: the man behind the legend, Imperial College Press, London

Weinacht J., 1924, Math Ann, 91, 279

Whittaker E.T., 1936 Analytical Dynamics, Cambridge University Press

York D., *et al.* 2000, AJ, 120, 1579

Zwitter T., *et al.* 2008, AJ, 136, 421

This page intentionally left blank

Bull. Astr. Soc. India (2011)

Monte Carlo radiative transfer

Barbara A. Whitney[*]
Astronomy Department, University of Wisconsin-Madison,
475 N. Charter St., Madison, WI 53706, USA
Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, Colorado 80301, USA

Received April 4, 2011; accepted April 14, 2011

Abstract. I outline methods for calculating the solution of Monte Carlo Radiative Transfer (MCRT) in scattering, absorption and emission processes of dust and gas, including polarization. I provide a bibliography of relevant papers on methods with astrophysical applications.

Keywords : radiative transfer – scattering – polarization – radiation mechanisms: general

1. Introduction

The Monte Carlo method was invented by Stanislaw Ulam and John von Neumann to study neutron transport during the atomic bomb program of World War II. According to Wikipedia, because the work was secret, a code name was needed, so they chose Monte Carlo, after the famous Casino in Monaco which Ulam's uncle frequented. At this time and for several decades after, the pressing radiative transfer problems in astrophysics were in stellar atmospheres and interiors, which fortunately are 1-D problems that could be solved with other, much faster methods. Many clever integral and differential equation techniques were devised to calculate sophisticated stellar atmosphere models, including line transfer and stellar winds. These methods are reviewed in several standard texts, e.g., Mihalas (1978). Scattering and polarization were always the most complicated aspects of these methods, and were therefore often ignored. Not surprisingly, these were tackled very early by S. Chandrasekhar (1946, 1960).

As radiative transfer began to be applied to other kinds of objects that are not as spherical as stars, it became necessary to consider multi-dimensional geometries and scattering. As an example, both forming and evolved stars are often surrounded by dusty disks and/or clumpy envelopes and outflows. The asymmetric circumstellar geometries produce very different spectral energy distributions (SEDs) than 1-D models can account for. Galaxies can appear bluer than expected if scattering from interstellar dust is not taken into account. A method that is ideally suited to solve these types of problems is the Monte Carlo method.

[*] e-mail: bwhitney@astro.wisc.edu

I was fortunate to have my thesis advisor, Art Code, suggest this method to study polarization in magnetic white dwarfs, back in the 1980s. I then applied this method in the area of star formation, where 2-D radiative transfer proved very useful in interpreting the disk and bipolar structures of Young Stellar Objects (YSOs). Since this time, many scientists have developed new methods to calculate, e.g., the radiative equilibrium solution for dust, gas line and continuum transfer, photoionization, polarization, and relativistic radiative transfer (references for these methods and applications are given later in the text). Now the Monte Carlo method is in widespread use in astronomy and is an exciting area to get into.

This article is designed for readers who are interested in learning the Monte Carlo method for radiative transfer in astrophysics. It starts with the basics needed to write a complete but simple Monte Carlo scattering code (Section 2), and then shows more complicated but common scattering problems (Section 3), dust emission (Section 4.1-4.5), and gas emission (Section 4.6). Not everything is described in detail, e.g., line scattering and gas emission, but numerous references are cited. I have made an attempt to include the most relevant and up-to-date references on methods, but I surely have missed some and I apologize for this.[1]

2. Monte Carlo basics and a simple scattering problem

In the Monte Carlo method for radiative transfer (MCRT), probabilistic methods are used to simulate the transport of individual 'photon packets' (which we will abbreviate as 'photons') through a medium. In this 'random walk', we just have to describe all the radiation sources, trace a path for each photon describing all interactions, and tabulate parameters of interest, such as intensity, flux, angle of exit, position of exit (for imaging), and wavelength. These should converge to a mean and become statistically significant when a large number of photons are processed. Many problems require iteration, and clever methods have been developed to handle this as well as high optical depths efficiently, as will be described later. In this section, we will describe the basic methods needed to solve a simple scattering problem, that of isotropic scattering in a plane-parallel atmosphere (see also Watson & Henney 2001, and Gordon *et al.* 2001 for an overview of the MCRT scattering solution). This is a problem that Chandrasekhar (1946, 1960) calculated analytically. His simplest case was a semi-infinite atmosphere, that is infinite in the x, y, and $-z$ directions and photons emerge from the top of the atmosphere, defined at $z = 0$. This is our most time-consuming case, which can be approximated by a plane parallel atmosphere with a large optical depth ($\tau = 7$ is sufficient) from bottom to top. Coulson, Dave & Sekera (1960) calculated finite thickness atmospheres using Chandrasekhar's method. Our code can be tested by comparing to Coulson *et al.*'s tables, recently updated by Natraj, Li & Yung (2009).

2.1 The Fundamental Principle: sampling probability distributions

The essence of the Monte Carlo Method is sampling from probability distribution functions (PDFs), and this is referred to as the 'Fundamental Principle'. To sample a quantity x_0 from

[1] Please send me any relevant references and I will update the online version of this document.

a PDF $P(x)$, we need to invert the cumulative probability distribution (CPD), $\psi(x_0)$, which is the integral of $P(x)$:

$$\psi(x_0) = \frac{\int_a^{x_0} P(x)dx}{\int_a^b P(x)dx}. \tag{1}$$

As x_0 ranges from a to b, $\psi(x_0)$ ranges from 0 to 1 uniformly (the proof of this can be found in Duderstadt & Martin 1979; see also Kalos & Whitlock 2008 or other standard Monte Carlo texts). Thus, to sample a 'random variate' x_0, we just need to call a random number generator that samples from 0 to 1 uniformly (we call this 'uniform random deviate' ξ), and invert equation (1) to get x_0.

To illustrate, we give the example of sampling the optical depth that a photon travels before being absorbed or scattered. The probability that a photon travels an optical depth τ without interacting is

$$P(\tau)d\tau = e^{-\tau}d\tau. \tag{2}$$

Applying the fundamental principle:

$$\psi(\tau) = \frac{\int_0^{\tau_0} e^{-\tau}d\tau}{\int_0^{\infty} e^{-\tau}d\tau} = 1 - e^{-\tau_0} = \xi. \tag{3}$$

Inverting this gives

$$\tau_0 = -\log(1 - \xi), \tag{4}$$

where ξ is the uniform random deviate returned from the random number generator subroutine. It is worth investigating the algorithm used by your compiler to find out how many numbers it generates before repeating. A good source for a discussion on random number generators and a recommended algorithm is given in Numerical Recipes (Press *et al.* 2007).

Sampling a scattering angle from an isotropic distribution ($P(\mu, \phi)d\mu d\phi = d\mu/2d\phi/(2\pi)$) is also very straightforward, giving

$$\begin{aligned} \mu_0 &= 2\xi_1 - 1 \\ \phi_0 &= 2\pi\xi_2 \end{aligned} \tag{5}$$

where $\mu = \cos\theta$, $d\mu = \sin\theta d\theta$.

We discuss in Section 3 different methods for sampling from more complicated PDFs. Kalos & Whitlock (2008) describe in detail different sampling methods. Carter & Cashwell (1975) describe methods relevant to radiative transfer, such as sampling from a Planck function.

2.2 The random walk

To calculate this problem, we emit photons from the bottom of a plane-parallel atmosphere, defining $\tau_z = 0$, and the top of the atmosphere is $\tau_z = \tau_{atm}$. The initial photon position is $x, y, z = 0, 0, 0$, and the initial direction is $\mu_0, \phi_0 = 0$. Sample optical depth from Eq. (4), and move the photon to a new position: $\tau_{znew} = \tau_{zold} + \mu * \tau$. Check to see if τ_{znew} is greater than τ_{atm}. If not, sample direction from Eq. (5) and continue to randomly walk until the photon exits. When the photon exits the top of the atmosphere, tabulate its angle of exit. Bin the angles uniformly between $\mu = 0 - 1$ and $\phi = 0 - 2\pi$:

$$i = integer(\mu N_\mu) + 1 \tag{6}$$

$$j = integer(\phi * N_\phi + 0.5) + 1; if j > N_\phi, j = 1 \tag{7}$$

where *integer* is a function that converts a real number to an integer (its actual call name depends on the computer language), and N_μ is the number of μ bins. and N_ϕ is the number of ϕ bins.

2.3 Calculating intensity and flux

Next we want to calculate the intensity of the exiting binned photons. From Chandrasekhar (1960; Eq. (1))

$$I_\nu = \frac{dE_\nu}{\cos\theta d\nu d\sigma dA dt} \tag{8}$$

where E_ν is the energy at frequency ν exiting at an angle θ to the normal of a surface with area dA into a solid angle $d\omega$ over time dt. This describes a pencil beam of radiation emitted from the surface of the atmosphere.

If $N_{i,j}$ is the number of photons exiting at μ_i, ϕ_j, and assuming for now monochromatic photons with no time dependence, then the intensity $I_{i,j}$ is given by

$$I_{i,j} = \frac{h\nu N_{i,j}}{\mu_i \Delta\mu \Delta\phi dA} \tag{9}$$

The intensity is usually normalized to flux F. As defined in Chandrasekhar (1960), the net rate of flow of energy across a surface per unit area per unit frequency interval is given by

$$\pi F = \int_{-1}^{1}\int_{0}^{2\pi} I(\mu,\phi)\mu d\mu d\phi \tag{10}$$

A total of N_0 photons are incident at cosine angle μ_0, giving

$$\pi F = \frac{h\nu N_0}{\mu_0 dA} \tag{11}$$

and therefore

$$\frac{I_{i,j}}{F} = \frac{\pi\mu_0 N_{i,j}}{\mu_i N_0 \Delta\mu\Delta\phi} \tag{12}$$

If the incident radiation is isotropic, $I_\nu(\mu,\phi) = I_0$, then Eq. (10) gives $F = I_0$. According to Eq. (8), $dE = I_0 \mu d\mu d\phi dA$. Integrating over solid angle and area gives $E = h\nu N_0 = \pi I_0$, which equals πF. Substituting this definition of F into Eq. (9) gives

$$\frac{I_{i,j}}{F} = \frac{\pi N_{i,j}}{\mu_i N_0 \Delta\mu\Delta\phi}, \tag{13}$$

which is the same as that for parallel incident radiation except there is no factor of μ_0.

By extending this algorithm to include electron scattering (Section 3.1), polarization, and albedo (Section 3.2.2), the code can be compared to Chandrasekhar (1946, 1960) and Code (1950) for large optical depths, and Coulson *et al.* (1960) and Natraj *et al.* (2009) for varying optical depths and incident angles. This is a great way to test out your Monte Carlo code, and learn how to compute intensity and flux. When considering more complicated problems with different boundary conditions, or frequency and time dependence, refer to the original definitions of intensity and flux to properly normalize the results. This is one reason I have referred to Chandrasekhar's (1960) book many times over the last 30 years.

2.4 More complicated geometries

The Monte Carlo Method solves problems in 3-D geometries as easily as 1-D, complicated scattering functions as easily as isotropic, and low optical depth more easily than high; therefore this is where it excels and is very complementary to other methods. All that is needed to solve any scattering problem is to describe where the photons originate from and in what direction, where the scattering material is, how it scatters, and when the photon exits. As described before, at each scatter, a new photon direction is chosen and a new optical depth. In most problems, the density of material varies with position, and the distance a photon travels is related to the optical depth through the exinction opacity (the sum of the absorptive and scattering opacities) of the material:

$$d\tau = \chi_1 \rho ds = \chi_2 n ds = \chi_3 ds \tag{14}$$

reflecting the different units the opacity might have. In this case, the units of χ_1 are cm^{-2} g, the units of χ_2 are cm^2 and the units of χ_3 are cm−1. In the first case multiply by the density ρ (g cm^{-3}), in the second case by the number density n (cm^{-3}), and in the third case, the density has already been factored into the value of χ_3. As the photon propagates, Eq. (14) must be integrated either analytically or numerically. The new photon position is then calculated from

$$\begin{aligned} x &= x_{old} + s \sin\theta \cos\phi \\ y &= y_{old} + s \sin\theta \sin\phi \\ z &= z_{old} + s \cos\theta \end{aligned} \tag{15}$$

In most problems where the density varies with position, we use grids to describe the problem, either spherical-polar, cylindrical, or cartesian. In each grid cell the density is constant across the cell. Given the photon propagation direction, the distance to the nearest wall is calculated, s_{wall} (in a cartesian grid, we find the distance to planes; in a spherical-polar grid, we find the distance to planes (ϕ), cones (θ) and spheres (r)). The photon position is updated using Eq. (15). The optical depth is updated:

$$\tau = \tau_{old} + \chi \rho_{cell} s_{wall} \tag{16}$$

If τ exceeds the sampled value (Eq. (4)), the photon is moved back to where $\tau = \tau_0$; otherwise it continues through the next cell where x, y, z, and τ are updated again. When $\tau = \tau_0$, the photon scatters.

2.5 Producing images

Images are easily computed by tracking the position of the previous interaction. When the photon exits, its position of last interaction (scatter or emission) is projected onto the $x - y$ plane perpendicular to the outgoing direction:

$$\begin{aligned} x_{image} &= z_{old} \sin\theta - y_{old} \cos\theta \sin\phi - x_{old} \cos\theta \cos\phi \\ y_{image} &= y_{old} \cos\phi - x_{old} \sin\phi, \end{aligned} \tag{17}$$

where $(x_{old}, y_{old}, z_{old})$ are the coordinates of the last interaction. Next we bin the photon into a pixel (ix, iy) on the image:

$$\begin{aligned} ix &= integer(nx(x_{image} + x_{max})/(2x_{max})) + 1 \\ iy &= integer(ny(y_{image} + y_{max})/(2y_{max})) + 1, \end{aligned} \tag{18}$$

where (nx, ny) are the number of x and y pixels in the image, and the image size ranges from $[-x_{max} : x_{max}]$ and $[-y_{max} : y_{max}]$.

2.6 Estimating errors

In the simple case of isotropic scattering as described above, the photon energy remains constant as it propagates through the medium, and the fractional error in the intensity is the Poisson statistical error $1/\sqrt{N}$ where N is the number of photons. In more complicated problems as described below, if we sample properly the PDFs for scattering and propagation, then the energy of each photon remains constant and is also given by simple Poisson statistical error. As described below, we could sample from isotropic scattering and then weight the photon by its more complicated phase function for scattering. Then the errors can be estimated from the standard deviation of the summed intensities of the outgoing photons normalized to $\sqrt{N}$. When polarization is included, the other Stokes parameters are estimated in the same way, by the standard deviation of the outgoing Stokes component (Q, U, or V), normalized to $\sqrt{N}$ (Wood *et al.* 1996). The errors are minimized when the PDFs are sampled exactly. Gordon *et al.* (2001) also discuss error estimation.

3. More complicated scattering problems

The kinds of scattering problems usually investigated in astrophysics applications are electron, Compton, resonance line, and dust scattering. In many cases, the scattering phase function (the angular dependence of the scattering function) can be defined or approximated with analytic functions, and in other cases, they are computed numerically and described in tabular form. All of these cases, including the polarization components, can be solved with relative ease with the Monte Carlo method. I summarize one general method here, including polarization (see also Chandrasekhar 1960; Code & Whitney 1995), noting that there are other variations to implement this (Hatcher Tynes *et al.* 2001; Cornet, C-Labonnote, & Szczap 2010; Hillier 1991). We use the Stokes Vector **S** to describe the polarization:

$$\mathbf{S}(\theta, \phi) = [I(\theta, \phi), Q(\theta, \phi), U(\theta, \phi), V(\theta, \phi)] \tag{19}$$

where I is the intensity, Q the linear polarization aligned parallel or perpendicular to the z-axis, U is the linear polarization aligned ±45°to the z-axis and V is the circular polarization. The Stokes vector could also be defined as $[I_{||}(\theta, \phi), I_{\perp}(\theta, \phi), U(\theta, \phi), V(\theta, \phi)]$, where $I_{||}$ is the intensity of light with polarization parallel to the z-axis, $I_{\perp}$ has polarization perpendicular to the z-axis, and $Q = I_{||} - I_{\perp}$. A scattering diagram is shown in Fig. 1 (Chandrasekhar 1960). The photon is originally propagating into direction $\mathbf{P_1}$ and will scatter into direction $\mathbf{P_2}$. In many scattering problems, the phase function can be described analytically dependent only on the angle Θ with respect to $\mathbf{P_1}$. For polarization problems, it is more complicated, because the polarization depends on the frame of reference. We define the polarization in the "observer's" frame (the $x - y - z$ frame in Fig. 1). Thus, we need to rotate into and out of the photon propagation direction to apply the scattering matrix, using Mueller matrices (Chandrasekhar 1960; Code & Whitney 1995). This is not strictly necessary, as the full scattering matrix can be calculated in the observer's frame (e.g., Whitney 1991a). In magnetic problems, it is easier to define the scattering phase function with respect to the magnetic field direction, and rotate in and out of these frames (Whitney & Wolff 2002). The resulting Stokes vector after scattering is:

$$\mathbf{S} = \mathbf{L}(\pi - i_2)\mathbf{R}\mathbf{L}(-i_1)\mathbf{S}', \tag{20}$$

where $\mathbf{S}'$ is the incident Stokes vector and $\mathbf{L}$ is the Mueller matrix that rotates in and out of the photon frame, defined as

$$\mathbf{L}(\psi) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 2\psi & \sin 2\psi & 0 \\ 0 & -sin2\psi & cos2\psi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{21}$$

The scattering matrix $\mathbf{R}(\Theta)$ is

$$\mathbf{R}(\Theta) = a \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \\ P_{41} & P_{42} & P_{43} & P_{44} \end{bmatrix} \tag{22}$$

where Θ is the scattering angle measured from the incident photon direction and a is a normalization factor. Note that if we want to ignore polarization, we can ignore all of the elements except P_{11}.

3.1 Rayleigh scattering

Let us consider the case of Rayleigh scattering, where

$$\begin{aligned} a &= 3/4 \\ P_{11} = P_{22} &= \cos^2\Theta + 1 = M^2 + 1 \\ P_{12} = P_{21} &= \cos^2\Theta - 1 = M^2 - 1 \\ P_{33} = P_{44} &= 2\cos\Theta = 2M \end{aligned} \tag{23}$$

where $M = \cos\Theta$, and the other elements are 0.

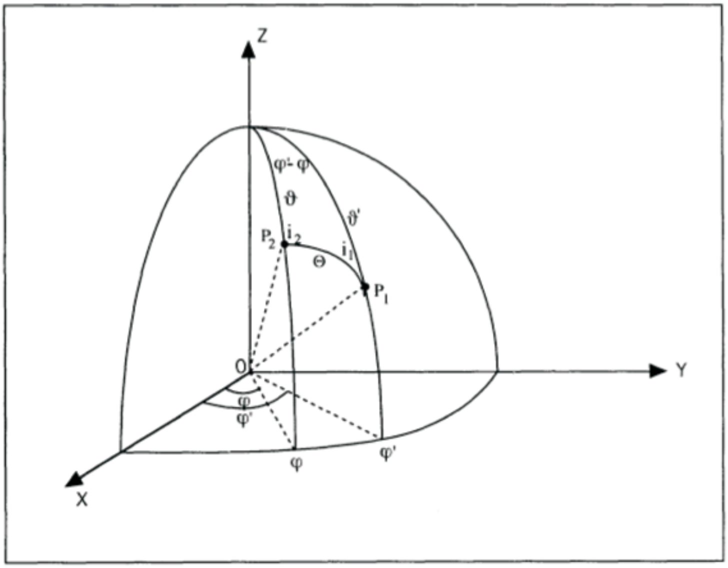


Figure 1. Geometry for scattering. A photon propagating into direction P_1 (θ', ϕ' in the observer's frame) scatters through angle Θ into direction P_2 (θ, ϕ).

Then the I Stokes parameter in the reference frame of the photon is computed:

$$\mathbf{S} = \mathbf{RL}(-i_1)\mathbf{S}', \tag{24}$$

giving:

$$I = (M^2 + 1)I' + (M^2 - 1)\cos 2i_1 Q' - 2M \sin 2i_1 U'. \tag{25}$$

We want to sample the scattering direction (M, i_1) from this function.

3.1.1 *Ignoring polarization and using lookup tables for sampling PDFs*

First consider the case where we ignore the polarization. Then $I = I'(M^2 + 1)$. There are a couple of ways we can sample scattering angle from this PDF. We could sample M from a uniform angular distribution (Eq. (5)), and calculate a new photon intensity at each scatter from $I = I'(M^2 + 1)$. Or we can sample the angle M directly from the PDF $I = I'(M^2 + 1)$. In this case the photon intensity will always be equal to 1 as it propagates through the medium. To do this for Rayleigh scattering, we apply the fundamental principle (Eq. (1)),

$$\xi = \frac{\int_{-1}^{M_0} 1 + M^2 dM}{\int_{-1}^{1} 1 + M^2 dM} = 1/2 + 3/8M_0 + 1/8M_0^2 \tag{26}$$

As described before, ξ is a uniform random number between 0 and 1, obtained from a random number generator. Inverting Eq. (26) to get M_0 for each scatter is not trivial. A fast way to sample M_0 is to make a table of the CPD (Eq. (26)), $1/2 + 3/8M_0 + 1/8M_0^2$, which ranges from 0 to 1 uniformly. Then linearly interpolate this table to get the value of M_0 that corresponds to the value of ξ obtained from the random number generator. Once M_0 is computed, an azimuthal angle i_1 is sampled ($i_1 = 2\pi\xi$), and the new direction in the coordinate frame of the observer is computed (Fig. 1).

3.1.2 *Including polarization and using the rejection method for sampling from PDFs*

If solving the full polarization problem, we will sample M and i_1 from the I Stokes parameter calculated from Eq. (24). As described before, we could sample $M(= \cos^2\Theta)$ and i_1 from uniform angular distribution (Eq. (5)), and calculate a new photon intensity from Eq. (25). Then the intensity of the photon will vary as the photon propagates through the medium. For Rayleigh scattering, where the intensity varies only by a factor of 2 with angle of scatter, it is okay to sample isotropically and weigh the photon intensity. For scattering that has a more peaked function, such as dust scattering, or in strong magnetic fields, this will lead to higher errors and systematic biases (many photons with small intensities and few with large intensities but poor statistics). To prevent this, I generally try to sample from the exact probability distribution. A simple method that samples from complicated probability distributions is the rejection method. All that is needed for this method is to know the peak of the PDF.

In the rejection method, we sample from a rectangle that encloses the curve of $P(x)$ vs x. That is, following Eq. (1), we sample x uniformly from a to b: $x_0 = a + \xi(b - a)$; and we sample y uniformly from 0 to P_{max}, the maximum value of $P(x)$: $y_0 = \xi P_{max}$. We ask if y_0 is less than $P(x_0)$. If so, we accept x_0. If not, we sample again. It is like throwing random darts at the plot and only accepting those that fall below the curve. By throwing enough darts, we accurately sample the different values of x appropriately. That is, in regions of the plot where $P(x)$ is low, we sample those values of x less frequently than regions where $P(x)$ is large. The rejection method is less efficient for highly peaked PDFs; that is, if the rectangle enclosing the PDF has a lot of area above the PDF. However, it is so simple to use that it is still usually much faster and easier to implement than more complicated inversions of the CPD (Eq. (1)). See Kalos & Whitlock (2008) or other Monte Carlo texts for more examples and more sophisticated modifications to this method (such as enveloping highly-peaked functions with simple analytic highly-peaked functions which are sampled from first).

Going back to our scattering problem, as described in Fig. 1, we want to sample scattering angles that change our direction from $\mathbf{P_1}$ to $\mathbf{P_2}$. That is, we want to sample Θ and $\imath_1$, compute the new Stokes parameters and then rotate back into the observer's frame of reference. Using the rejection method, we sample i_1 and $M = \cos\Theta$ from an isotropic distribution (Eq. (5)): $i_1 = 2\pi\xi_1; M = 2\xi_2 - 1$. We calculate $I(M, i_1)$ from Eq. (25). We sample $P(M, i_1) = \xi P_{max}$. If $P(M, i_1)$ is greater than $I(M, i_1)$, we accept M and i_1 as our new scattering angles. Otherwise, we resample until $P(M, i_1)$ is greater than $I(M, i_1)$. As mentioned previously, we need to know the value of P_{max}. This can be determined analytically or numerically (from brute-force calculation over all angles). It is a good idea to verify that $P(M, i_1)$ never exceeds P_{max} during the run.

Now that we have our new scattering angles M and i_1, we compute the new propagation direction and Stokes vectors in the observer's frame. The angles $i_2, \theta, \phi - \phi'$ (Fig. 1) can be calculated from the spherical laws of sines and cosines (Green 1985). The matrices are multiplied through and the Stokes parameters are calculated from Eq. (20). The Stokes vectors are then normalized to the PDF we sampled from $P(M, i_1)$ (Eq. (25)). Then the I Stokes parameter of the photon is equal to 1 as it propagates through.

3.2 Dust scattering

Since dust is ubiquitous throughout the universe, having the capabilities to solve the radiative transfer of dust in multi-dimensional geometries allows us to model everything from planets, extrasolar planets, forming stars, evolved stars, star forming regions, and galaxies throughout the universe. Dust scattering can be approximated with analytic functions or tables produced from numerical models.

3.2.1 *Analytic functions*

The most famous analytic function is the Henyey-Greenstein (H-G) function (Henyey & Greenstein 1941). White (1979) added to this with approximations for the polarization functions. The elements of the scattering matrix $\mathbf{R}(M)$ (where $M = \cos\Theta$) (Eq. (22)) are

$$\begin{aligned} &a = 3/4 \\ &P_{11} = P_{22} = (1-g^2)/(1+g^2-2gM)^{3/2} \\ &P_{12} = P_{21} = -p_1 P_{11}(1-M^2)/(1+M2) \\ &P_{33} = P_{44} = P_{11}(2M)/(1+M^2) \\ &P_{34} = P_{43} = -p_c P_{11}(1-M_f^2)/(1+M_f^2), \end{aligned} \tag{27}$$

where g is the scattering asymmetry parameter, ranging from 0 for isotropic scattering to 1 for fully forward scattering; p_l is the maximum linear polarization; p_c is the peak circular polarization; $M_f = \cos\Theta_f$, $\Theta_f = \Theta(1 + 3.13sexp(-7\Theta/\pi)$, and s is the skew factor which we take to be 1 following White (1979). The other elements in Eq. (22) are 0. Note that this function includes a circular polarization component (P_{34} and P_{43}). This is a second order effect that depends on the linear polarization and is usually small.

Multiplying through Eq. (24) to get the I Stokes parameter in the photon reference frame gives:

$$I = P_{11}I' + P_{12}\cos 2i_1 Q' - P_{12}\sin 2i_1 U' \tag{28}$$

The scattering angles M, i_1 can be sampled using the rejection method (Section 3.1.2).

If you don't care to solve the polarization problem, you just use P_{11} for the scattering phase function. This can be sampled from directly using the following formula (Witt 1977a):

$$M = \frac{1 + g^2 - [(1-g^2)/(1-g+2g\xi)]^2}{2g} \tag{29}$$

Witt (1977a) describes in detail a Monte Carlo dust scattering algorithm using this function, as well as superpositions of H-G functions. He also describes how to force the first scattering in an optically thin nebula to make the code more efficient (see also Gordon *et al.* 2001).

The other parameters that describe the dust properties are the extinction opacity χ (see Eq. (14)) and the albedo ω (Section 3.2.2). These as well as g have been estimated observationally. Theoretical models also match these as well as estimating p_l and p_c which can be tested by comparing scattered light models to polarization observations. All of these quantitites are wavelength-dependent.

3.2.2 *Dust scattering albedo*

The scattering albedo is the ratio of scattered to extinct (scattered + absorbed) flux, and it ranges from 0 to 1. This can be taken into account in one of two ways: either by weighting the photon at each scatter by the albedo, or by casting for a random number ξ to determine if the photon is absorbed or scattered at each interaction. In calculations where we only consider the scattered component of the radiation at a specified wavelength (e.g., dusty sources illuminated by UV, optical, and near-IR radiation), we might think the first solution would be more efficient, that of weighting the scattered photons by albedo. This is often not the case, especially in sources with very high optical depths in some regions, where too much computing time is wasted on photons with little weight and therefore little contribution to the final answer. In those cases, it is much faster to let the photon scatter or absorb by casting for a random number. If ξ is less than the albedo, the photon scatters; otherwise, it is killed, and we proceed to the next photon.

3.2.3 *Tabular functions*

The scattering matrix $\mathbf{R}(M)$ (Eq. (22)) can also be computed numerically. Tables of the 16-element matrix as a function of scattering angle are read at the beginning of the computation. For spherical grains, the matrix is simplified, with only 4 independent elements needed, as above in the analytic approximation. For randomly oriented non-spherical grains, 6 independent elements are needed. For aligned grains, all 16 elements are non-zero.

The rejection method works well at sampling the tabular functions. At the beginning of the code, the peak of the M_{11} element is computed, which we will call I_{peak}. In the cases I have tried, this is also the peak of the I Stokes vector even when the incident radiation is polarized. At each scatter, as described in Section 3.1.2, the angles M and i_1 are sampled uniformly. The values of P_{11}, P_{12}, and P_{13}, P_{14} (if non-zero) are calculated by interpolating the tables (which depend on M). Then the I Stokes parameter in the reference frame of the photon is computed from Eq. (24):

$$I = P_{11}I' + (P_{12}\cos 2i_1 + P_{13}\sin 2i_1)Q' + (P_{13}\cos 2i_1 - P_{12}\sin 2i_1)U' + P_{14}V'. \qquad (30)$$

For spherical grains, only 8 of the scattering matrix elements are filled with 4 unique elements, as in the analytic prescription above: $P_{11} = P_{22}$, $P_{12} = P_{21}$, $P_{33} = P_{44}$, $P_{34} = P_{43}$, and the rest are zero, giving the same form as Eq. (28). As described in Section 3.1.2, a random number ξ is chosen between 0 and the peak of I; if ξ is less than I_{peak}, the angles M and i_1 are accepted, and the rest of the Stokes vectors are calculated from Eq. (20). To verify that we properly calculated the peak of the scattering function, we check at each scattering that the I Stokes parameter does not exceed I_{peak}. if it does, we need to rerun the code with the correct value. Once the scattering angle has been calculated, the other angles are computed, and the Stokes vector in the observer frame are computed (Eq. (20)), as described in Section 3.1.2.

3.2.4 *Aligned grains*

Aligned grains use the full 16-element scattering matrix, calculated as described in the previous section (Section 3.2.3). Instead of rotating in and out of the photon direction frame, we rotate into and out of the frame aligned with the magnetic field along the *z*-axis. The 16-element scattering matrix is defined with respect to field direction rather than photon direction. The additional component here is in the random walk, where the opacities depend on the polarization of the photon. Photons traversing the medium develop Q polarization in the frame of the magnetic field, called dichroism. Photons with some U polarization (w.r.t. magnetic field direction) develop V polarization, called birefringence. Whitney & Wolff (2002) describe how to implement these effects along the photon propagation path.

3.2.5 *Applications of continuum scattering problems*

Most electron scattering applications are in resonance line scattering of stellar winds, as described in the next section. Whitney (1991a) described how to calculate the scattering of electrons in magnetic fields of arbitrary strength, and showed how the magnetic effects can explain the unusual polarization behavior in the polarization of magnetic white dwarfs (Whitney 1991b).

The most widespread applications of Monte Carlo (MC) continuum scattering have been for dust scattering. Witt (1977a,b,c) and Witt & Oshel (1977) pioneered this field describing algorithms for sampling the Henyey-Greenstein function and computing the MC radiative transfer. Witt and collaborators applied these codes to galaxies showing the “blueing” due to scattering partially compensates for reddening by extinction (Witt, Thronson & Capuano 1992) and the effects of clumping on the radiative transfer (Witt & Gordon 1996, 2000). Bianchi *et al.* (2000) also studied the effect of clumping in dusty galaxies. Boisse (1990) studied the effects of clumps in the penetration of UV photons inside molecular clouds. Whitney & Hartmann (1992, 1993), Kenyon *et al.* (1993), and Fischer, Henning & Yorke (1994) calculated dust scattering and polarization in 2-D structures–disks, envelopes, and bipolar cavities — surrounding protostellar envelopes. Several authors have modeled high spatial-resolution images from Young Stellar Objects (YSOs), determining disk/envelope properties and grain size distributions (e.g., Wood & Whitney 1998; Cotera *et al.* 2001; Schneider *et al.* 2003; Wolf, Padgett & Stapelfeldt 2003; Watson & Stapelfeldt 2004, 2007; Duchene *et al.* 2004; Stark *et al.* 2006; Watson *et al.* 2007 and references therein), and polarization maps (Whitney, Kenyon & Gomez 1997; Lucas & Roche 1997, 1998). Whitney & Wolff (2002), Lucas (2003), and Lucas *et al.* (2004) modeled polarization maps of YSOs with aligned grains, to study the magnetic field structures. Jonsson (2006) describes a code for computing scattering in galaxies. The advances of this code are that it follows a spectrum of photons through, rather than a single wavelength; and is designed to work with SPH simulations and on an adaptive grid. This code is widely used in the study of galaxy evolution to visualize galaxy images produced from SPH simulations (such as the GADGET code; Springel, Di Matteo & Hernquist 2005).

3.3 Line scattering problems

3.3.1 *Resonance line scattering and scattering in flows*

Resonance lines are transitions to and from the ground states of bound electrons. The scattering matrix is the sum of a Rayleigh phase function plus an isotropic function. In a flow, such as an expanding atmosphere or universe, we take into account the Doppler shifts of the fluid with respect to the incident photons. Hillier (1991) calculated the electron scattering of lines in Wolf-Rayet stars. He described how to calculate the emission location, that is, where the photon of a given direction and frequency will resonantly interact with the flow, and how to transform the frequency from one frame to the next in the flow. Kurosawa & Hillier (2001) applied these algorithms in a 3-D tree-structured grid, and demonstrated their model on interacting winds in massive binaries (see also Kurosawa, Hillier & Pittard 2002 for an application to the massive binary V444 Cyg). Sundqvist, Puls & Feldmeier (2010) calculated resonance line formation in 2-D wind models, in an ongoing effort to resolve a very interesting new controversy on mass-loss rates from clumpy massive stellar winds (see Puls, Vink & Najarro 2008). They required higher mass loss rates than in the optically thin clump models which they said resolves the controversy. Knigge, Woods & Drew (1995) calculated resonance line scattering in accretion disk winds.

Another useful application for resonance line scattering is the radiative transfer of Lyα photons. This problem can be approximately solved analytically only for a limited number of cases such as a static, extremely opaque and plane-parallel medium. Several authors describe radiative transfer calculations (e.g., Zheng & Miralda-Escude 2002; Verhamme, Schaerer & Maselli 2006; Laursen *et al.* 2009)) and apply them to, e.g., Lyα radiative transfer in a dusty, multiphase medium (Hansen & Oh 2006), Lyα pressure in the neutral intergalactic medium (Dijkstra & Loeb 2008), Lyα escape fractions from simulated high-redshift dusty galaxies (Laursen, Sommer-Larsen & Andersen 2009), cosmological reionization simulations (Zheng *et al.* 2010), and the Lyα forest around high redshift quasars (Partl *et al.* 2010).

3.3.2 *Relativisitic scattering*

In principle, the calculations for relativistic scattering processes are similar, with additional transformations of the photon frequency in and out of the co-moving frame. If gravitational redshift is important, we need to apply this to the photon frequency at each step of the photon path integration. For more information, I refer the reader to other authors who know much more than I: Wang, Wasserman & Salpeter (1988) calculate cyclotron line resonance transfer in neutron star atmospheres; Fernandez & Thompson (2007) also calculate cyclotron resonance scattering in 3-D geometries. Stern *et al.* (1995) describe a large particle (LP) method for simulating non-linear high-energy processes near compact objects. And Dolence *et al.* (2009) describes a general code (grmonty) for relativistic radiative transport.

4. Including emission

Adding emission usually adds wavelength dependence to the problem and allows us to model the spectral dependence of an astrophysical source. The dominant emission processes are from gas and dust. We start with dust, which is the easiest to calculate. Fortunately, a wide variety of astrophysical problems can be addressed with 3-D dust radiative transfer, due to the wealth of infrared data recently available from, e.g., the Spitzer Space Telescope, Herschel Space Observatory, Wide Field Infrared Survey Explorer (WISE), and the upcoming James Webb Space Telescope.

4.1 Dust radiative equilibrium

Due to the nature of its opacity, dust generally scatters and absorbs optical radiation, and emits infrared radiation. For grains larger than about 200 A in radius, we can usually assume that the dust is in thermal equilibrium with the surrounding gas (we will address smaller grains in Section 4.2). The gas-to-dust mass ratio is about 100 in our Galaxy. Even though there is much more gas mass than dust, its opacity is many orders of magnitude larger than gas, so we can usually neglect the gas opacity in dusty nebulae.

We calculate the radiative transfer as described previously, but when a photon is absorbed (see Section 3.2.2), we re-emit a thermal photon. To do that, we need to know the temperature of the dust. This is straightforward to solve under conditions of radiative equilibrium and local thermal equilibrium (LTE). The radiative equilibrium process describes the condition when all of the energy is transported by radiation. Then we can say that the total energy absorbed by a given volume of material is equal to the total energy emitted (Mihalas 1978):

$$4\pi \int_0^\infty \chi_\nu (S_\nu - J_\nu) d\nu, \tag{31}$$

where S_ν is the Source function, or the ratio of the total emissivity to the opacity, J_ν is the average intensity in the same volume, and $\chi_\nu = \kappa_\nu + \sigma_\nu$ is the mass extinction coefficient. In local thermal equilibrium, we can write (Mihalas 1978):

$$S_\nu = (\kappa_\nu B_\nu + \sigma_\nu J_\nu)/(\kappa_\nu + \sigma_\nu) \tag{32}$$

where κ_ν and σ_ν are the mass absorption and scattering coefficients, respectively, and their sum is χ_ν (in units of cm^2/g). The condition of radiative equilibrium is then

$$\int_0^\infty \kappa_\nu B_\nu(T) d\nu = \int_0^\infty \kappa_\nu J_\nu d\nu, \tag{33}$$

This is all the information we need for our Monte Carlo calculation. We will do our calculation on a grid so we can calculate the volume and mass of each cell for the emission properties. This also allows flexibility in including arbitrary density functions and makes optical depth integrations straightforward (Section 2.4).

Bjorkman & Wood (2001, hereafter BW01) describe how to determine the temperature of each grid cell by equating the total absorbed photons with those emitted assuming thermal equilibrium. This gives

$$\sigma T_{cell}^4 = \frac{N_{cell} L}{4N \kappa_P (T_{cell}) m_{cell}}, \tag{34}$$

where N_{cell} is the number of photon packets absorbed in the cell, L is the source luminosity, $\kappa_P(T_{cell})$ is the Planck mean opacity, m_{cell} is the mass of the cell, and N is the total number of photon packets in the simulation. This applies to any continuous opacity source that is independent of temperature. To solve this equation efficiently, we pretabulate the Planck mean opacities and use a simple iterative algorithm.

When a photon is absorbed in a cell we sum its energy into an array for use in computing Eq. (34). We then emit a new photon of equal energy to conserve radiative equilibrium. All that's required is to properly sample its frequency from the emissivity function converted to a PDF:

$$\frac{dP_{cell}}{d\nu} = \frac{j_\nu}{\int_0^\infty j_\nu d_\nu} = \frac{\kappa_\nu B_\nu(T_{cell})}{\int_0^\infty \kappa_\nu B_\nu(T_{cell}) d\nu} \tag{35}$$

where $(dP_{cell}/d\nu)$ is the probability of emitting a photon between frequencies ν and $nu + d\nu$. We precompute the running integral of this function (that is, the cumulative probability distribution or CPD, see Section 3.1.1) for a range of frequencies and temperatures, and interpolate the table based on the sampled random number ξ to get ν.

At the start of our simulation, we do not know the temperature of each cell, so we use an arbitrary value (we start with 3 K), and use the absorbed photons to determine the temperature. We can iterate, i.e., do the calculation several times, and calculate a new temperature for each cell (Eq. (34)) after each iteration, until the cell temperature converges (Lucy 1999a). Alternatively, we can correct the temperature as we go and emit from a corrected emissivity spectrum (BW01). This corrects the emitted spectrum so that the total emitted spectrum at the end of the simulation is appropriate for the temperature of that cell. For example, if the cell starts out cold, the emitted photon frequencies will be lower than the proper spectrum, so as the temperature warms up, we will sample from an overly "hot" spectrum to emit higher frequency photons. This is described graphically in Fig. 1 of BW01. The temperature correction probability distribution is

$$\frac{dP_{cell}}{d\nu} = \frac{\kappa_\nu}{K} \left(\frac{dB_\nu}{dT} \right)_{T=T_{cell}}, \tag{36}$$

where $K = \int_0^\infty \kappa_\nu (dB_\nu/dT) d\nu$ is the normalization constant. Again, we can precompute the CPD and interpolate from this to sample ν based on random number ξ.

Lucy (1999a) derived a much faster way to compute the total absorbed radiation in a grid cell (the right-hand-side of Eq. (33)), using the pathlengths of *all* photons crossing a cell, rather than summing only those absorbed. This gives

$$\int_0^\infty \kappa_\nu J_\nu d\nu = \frac{L}{4\pi N V} \sum \kappa_\nu l, \tag{37}$$

where V is the volume of the cell, l is the pathlength across the cell that a given photon traveled, and the others are as defined in Eq. (34). The pathlengths are summed during the optical depth integration as the photon travels through various cells on its way to an interaction. Following BW01 and equating this with the emitted radiation to solve for temperature, we get:

$$\sigma T_{cell}^4 = \frac{\rho_{cell} L \sum \kappa_\nu l}{4N\kappa_P(T_{cell}) m_{cell}}, \tag{38}$$

Thus, we can call our temperature solver with $\rho_{cell} \sum \kappa_\nu l$ in place N_{cell}. The simplest way to implement this method for calculating temperature is with an iterative scheme. The temperature remains constant during an iteration, we sample frequency from the emissivity (Eq. (35)), and then calculate a new temperature for each cell at the end of the iteration (Eq. (38)). Lucy (1999a) notes that this temperature correction scheme appears identical to the “notorious” lambda-iteration procedures that are known to fail (Mihalas 1978); however it is not the same, because flux is conserved exactly across all surfaces. In fact, this method converges in only a few iterations (3-4).

This method has several advantages over BW01: 1) It is very fast at converging the temperature. Chakrabarti & Whitney (2009) quantified this by running several 3-D simulations and comparing the BW01 and Lucy methods. In the Lucy iteration method, the number of photons required to get an accurate temperature is approximately $N_{temp} \sim 2N_{grid}$, where N_{grid} is the number of grid cells. The BW01 method requires at least $N_{temp} \sim 100N_{grid}$. In the Lucy method, we run the first n iterations using N_{temp} photons, and then run the final iteration using N_{SED}, the number of photons required to produce an SED of our desired signal-to-noise. Usually, N_{SED} is much larger than N_{Temp}. In 2-D problems, the run-time of Lucy and BW01 is similar; in 3-D problems, because there are so many more grid cells, the Lucy method runs much faster. 2) The Lucy method is easily parallelizable. Since the temperature remains constant during an iteration, the photons can be divided up among several processors and run independently. At the end of each iteration, they are summed up and a new temperature is calculated. 3) More complicated physical processes that require iteration can be incorporated in a straightforward way. For example, including temperature dependent opacities (e.g., gas opacity); calculating grain alignment from moments of the radiation intensity; and calculating non-thermal small grain emissivity which requires knowledge of the average intensity in a grid cell.

4.1.1 *High fidelity spectra and images*

A useful technique for computing a high signal-to-noise image and SED is to ‘peel-off’ a photon in a specified (observer’s) direction at every interaction (Yusef-Zadeh, Morris, & White 1984). When a photon is initially emitted, in addition to its sampled direction, we emit an additional photon into one or more specified observer directions, weighted by the PDF, or the probability that it would have gone in this direction. The photon’s intensity is additionally weighted by the extinction it undergoes on its way to the observer $I = I_0 e^{-\tau}$ where τ is the integrated optical depth along its path. At each interaction (scattering or emission), we again peel-off a photon into the observer direction, weighted by the PDF (for scattering or emission), and the extinction. Note that the peeling-off technique does not replace the regular Monte Carlo simulation, but is an added computation. The main ‘trick’ with this is that we have to make sure that the peeled photon is normalized properly. In the regular simulation, this is done at the end of the simulation with the conversion of exiting photons to flux and energy; during the simulation, the PDF’s are normalized to range from 0-1 (to match the random number range). For example, in emitting photon packets from a limb darkened star, we emit each photon with the same energy, but the distribution of emitted photons varies with angle. For the peeled photon, we weight it by the limb-darkening law and need to normalize it properly. Fortunately, this is easy to verify by comparing the peeled images and spectra with the regular Monte Carlo in simulations that test all the emission and scattering processes (i.e., viewing images and SEDs of the star only, then scattering-only simulations, emission-only, high and low optical depths, etc.).

4.2 The diffusion method

In sources with high optical depths, MCRT can become very slow to compute when the photon path length is much shorter than the escape length from a given region. In dust radiative transfer this effect is offset to some extent because the opacity of dust decreases with increasing wavelength: optical photons that are absorbed and re-emitted by the cooler dust get converted to infrared photons that can usually escape. Thus sources with visual optical depths of even 1000 are computed quickly. However, in regions of much higher optical depths, such as protostellar disks, the photons effectively get trapped in the disk midplane, undergoing millions of interactions before escaping. Min *et al.* (2009, hereafter M09) developed a modified random-walk (MRW) that moves photons through optically thick regions, using the diffusion approximation.

In the MRW method, when the optical depth in a grid cell is much larger than 1, we define a sphere whose radius is smaller than the distance to the closest wall, and travel to the edge of the sphere in a single step. The true distance the photon would have traveled in a random walk is calculated using the diffusion approximation. This along with the average mass absorption coefficient are used to compute the total energy deposited and therefore the temperature of the cell. A new photon emerges from the sphere with the frequency sampled from the Planck function at the local dust temperature. If the BW01 temperature correction method is used, the photon frequency is sampled from $dB_\nu(T)/dT$. Robitaille (2010) showed how to compute the local diffusion coefficient D, the average mass absorption coefficient and the dust emission coefficient η_ν without iteration, giving:

$$D = \frac{1}{3\rho\overline{\chi}_R}, \tag{39}$$

$$\overline{\kappa} = \frac{\int_0^\infty \kappa_\nu B_\nu(T)d\nu}{\int_0^\infty B_\nu(T)d\nu} = \overline{\kappa}_P, \tag{40}$$

$$\eta_\nu = \chi_\nu B_\nu(T)\frac{\overline{\kappa}_P}{\overline{\chi}_P}, \tag{41}$$

where χ_P is the Planck mean opacity,

$$\chi_P = \frac{\int_0^\infty \chi_\nu B_\nu(T)d\nu}{\int_0^\infty B_\nu(T)d\nu} \tag{42}$$

and χ_R is the Rosseland mean opacity:

$$\frac{1}{\chi_R} = \frac{\int_0^\infty \chi_\nu B_\nu(T)/\chi_\nu d\nu}{\int_0^\infty B_\nu(T)d\nu}. \tag{43}$$

Robitaille (2010) describes the implementation of the MRW algorithm in his Section 3, so I refer the reader to that.

M09 also describe a Partial Diffusion Approximation (PDA) which can be used to obtain a reliable temperature in regions where few if any photons reach, such as the midplane of an externally illuminated disk with no self-luminosity due to accretion. For computations of images and SEDs, if no photons reach a given region, none are emitted, so PDA is not needed. However, if we want to solve for the vertical hydrostatic density distribution of the disk, the temperature in all regions is required. The PDA assumes that no photons escape the optically thick region without interactions, which simplifies the 3-D radiative diffusion equation (Wehrse, Baschek & von Waldenfels 2000; Rosseland 1924)

$$\nabla \cdot (D \nabla E) = \frac{1}{c} \frac{\partial E}{\partial t} \tag{44}$$

to

$$\nabla \cdot (D \nabla T^4) = 0. \tag{45}$$

This results in a system of linear equations that can be solved knowing the temperature at the boundaries of the optically thick regions. Thus the PDA requires iteration, using the temperature calculated from the MCRT solution. The PDA overestimates the temperature slightly because it does not take into account the few very long-wave photons that can escape from the region and cool it more efficiently.

4.3 Non-equilibrium dust (small grain emission)

Grains smaller than about 200 A, or Very Small Grains (VSGs), as well as large molecules such as Polycyclic Aromatic Hydrocarbons (PAHs) undergo quantum heating from even single photons, which leads to temperature fluctuations. These fluctuations depend on the size of the particle. Given a probability distribution P(T)dT for the temperature of a grain, the emission from an ensemble of VSGs is given by (Misselt *et al.* 2001)

$$L(\nu) = 4\pi \sum_i \int_{a_{min}}^{a_{max}} n_i(a) \sigma_i(a, \lambda) da \int B_\nu(T_{i,a}) P(T_{i,a}) dT \tag{46}$$

where i is the species of the grains (e.g., silicates or carbon), n is the number density of grains (typically units are cm^{-3}) of radius a and σ is the cross section of the grains (in units of cm^2). This can be compared to the left-hand side of Eq. (33), where the grain cross sections are already integrated over size and are all assumed to emit at the same temperature T, which is valid for large grains. Misselt *et al.* (2001) describe how to determine P(T) for VSGs using the continuous cooling approximation developed by Guhathakurta & Draine (1989), which speeds up the calculation significantly. They describe an even more simplied approach to compute the PAH emission:

$$L_{PAH}(a, \nu) = 4\pi\sigma(a, \nu) \overline{B[T(t)]}, \tag{47}$$

where the Planck function is averaged over the mean time between absorptions calculated from

$$\frac{1}{\bar{t}} = \frac{4\pi}{hc} \int_0^{\nu_c} \sigma(a, \nu) J_\nu d\nu, \tag{48}$$

where ν_c is the cutoff frequency in the optical/UV cross section of the PAH molecule (Desert *et al.* 1990).

In their radiative transfer algorithm, Misselt *et al.* (2001) first process the stellar and nebular sources, calculating the transmitted, scattered and absorbed photons in the grid. Then they calculate the dust emission and transfer based on the heating from the absorbed photons. They iterate on the fractional change of energy absorbed by the grid. This method does not conserve energy in a given iteration and may be subject to Lambda iteration issues. The large grain emission is as described in the radiative equilibrium Eq. (33), using the average intensity of each cell computed at the end of an iteration. The PAH and very small grain component is as given in Eqs. (46) and (47). The solution for the very small grains is the most computationally expensive part of the code.

Pontoppidan *et al.* (2007) also use the method of Guhathakurta & Draine (1989) to compute the heating of the very small grains, and do not compute the PAH emission (though they do include PAH absorption opacity). Photons absorbed by these very small grains are lost in the first iteration, to be released in a post-processing step and/or in a second iteration.

Wood *et al.* (2008) bypass the temperature calculations of the VSGs and PAHs altogether, and use look-up tables for the emissivity of these species. The input to the lookup tables is the average intensity J in each grid cell, calculated using the Lucy (1999a) method (Eq. (33), without the opacity). This method requires iteration. In each iteration the photons are emitted from the star and other luminosity sources (e.g., disk accretion) and are processed as described in previous sections. At each interaction, we sample a probability that a photon is absorbed by a thermal grain, a VSG, or a PAH molecule, based on the relative opacities of these material for the frequency of the incoming photon. If a thermal grain, a thermal photon is emitted based on the temperature of the cell (Eq. (35)); if a VSG or PAH, a non-thermal photon is emitted from the pre-computed emissivity spectra based on J in the cell. After each iteration a new temperature and J are computed in each cell. Energy is conserved, and the models converge in 3-4 iterations. This method is as fast as the radiative equilibrium method using the Lucy method. The lookup tables incorporate all the physics of the temperature fluctuations and emission as a function of input radiation field, but are pre-computed so that it does not slow down the radiative transfer calculations. The main approximation to the Wood *et al.* (2008) implementation is that they do not take into account the frequency dependence of the average intensity (J_ν). This assumption is not as egregious as it might seem because the wavelength dependence of the opacity is taken into account, ensuring that PAH and VSG photons are not emitted in regions with high J but low probability of excitation. Future implementations will likely incorporate wavelength dependence to the look-up tables.

4.4 Aligned grain emission

Thermal emission from aligned grains is similar to that of spherical grains except the full Stokes matrix is used in the emission. The dust opacities need to be calculated, along with the degree of alignment. Fiege & Pudritz (2000) describe a method for emitting polarized submillimeter emission in molecular clouds. Bethell *et al.* (2007) and Pelkonen, Juvela & Padoan (2009) show how to calculate the degree of alignment using radiative torques. Hoang & Lazarian (2008, 2009a, 2009b), and Hoang, Draine & Lazarian (2010) present new calculations on the radiative torque mechanism. Because of the low opacities at these wavelengths, the absorption and scattering is ignored in these calculations. In protostellar disks where the grains are larger and the optical depths higher, these approximations are likely not valid. Whitney & Wolff (2002) describe how to include absorption along the photon path and scattering of aligned grains. When emission, scattering, and absorption are included, models can be made at all wavelengths and densities.

4.5 Applications of dust MCRT

Several authors have developed dust MCRT codes that can be applied to a variety of astrophysical objects. Their methods are generally similar to what I described above but there are variations in, for example, conserving energy by re-emitting photons as they are absorbed vs separating the initial emission and re-emission processes; or different coordinate-system rotations for the Stokes vectors (conceptually simple vs computationally efficient). Numerical techniques and codes have been described by Lucy (1999a), Wolf, Henning & Stecklum (1999), Wolf & Henning (2000), Misselt *et al.* (2001), Bjorkman & Wood (2001), Wolf (2003), Stamatellos & Whitworth (2003), Stamatellos, Whitworth, & Ward-Thomson (2004), Whitney *et al.* (2003a,b), Niccolini *et al.* (2003), Goncalves, Galli & Walmsley (2004), Baes *et al.* (2005), Pinte *et al.* (2006), Niccolini & Alcolea (2006), Pontoppidan *et al.* (2007), Bianchi (2008), Wood *et al.* (2008), Min *et al.* (2009), Kama *et al.* (2009), and Robitaille (2010). Adaptive grid techniques have been described by Niccolini & Alcolea (2006). Benchmark tests have been made by Pascucci *et al.* (2004) and Pinte *et al.* (2009).

These codes have been applied widely in the study of protostellar envelopes/disks, and galaxies. In both cases, clumpy structures (e.g., Schartmann *et al.* 2008 and Bianchi 2008 for galaxies, Indebetouw *et al.* 2006 for protostars, Doty, Metzler, & Palotti 2005 for externally heated molecular clouds), and other asymmetric dust distributions (e.g., outflow cavities and disks) require 2-D and 3-D radiative transfer codes to properly interpret the SEDs, images, and polarization.

Grain alignment models have been applied to near-IR polarization maps, to determine magnetic field structures in protostars (Whitney & Wolff 2002; Lucas 2003; Lucas *et al.* 2004); and to submillimeter polarization maps to determine magnetic structures (Fiege & Pudritz 2000), density distributions, grain size distribution (Pelkonen *et al.* 2009), and to test the radiative torque theories for grain alignment, polarization-Intensity relations (Bethell et al. 2007; Pelkonen, Juvela & Padoan 2007), and the Chandrasekhar-Fermi formula (Padoan *et al.* 2001).

The recent explosion of optical and IR data from several observatories and surveys (e.g., Spitzer Space Telescope, Herschel Space Telescope, Hubble Space Telescope, 2MASS, UKIDDS, WISE), combined with advances in dynamical simulations that provide realistic density distributions, has made the development of 3-D dust radiative transfer a very fruitful area of research.

4.6 Gas emission

4.6.1 *Non-LTE MCRT and flows*

As in the scattering and dust emission processes, MCRT is very complementary to other methods. Whereas traditional methods excel in high optical depth LTE 1-D geometries, MCRT can excel in non-LTE, 3-D geometries with complex velocity fields and anisotropic radiation fields. Bernes (1979) outlined a procedure for non-LTE multi-level radiative transfer and demonstrated the method for CO line profiles in a spherical, homogeneous, collapsing dark cloud. Since then, several authors have improved on the Bernes (1979) algorithms to, e.g., extend to 3-D (Park & Hong 1995) allow for very high optical depths (Hartstein & Liseau 1998), treat clumpy structures (Park, Hong & Minh 1996; Juvela 1997; Pagani 1998), accelerate the convergence and include dust emission Hogerheijde & van der Tak (2000), and include multiple molecules (Pavlyuchenkov *et al.* 2007).

The application of MCRT to the computation of expanding gaseous envelopes was described by Abbott & Lucy (1985). Mazzali & Lucy (1993) adapted this code to supernova envelopes, where a single continuum photon can interact with many more spectral lines due to the high velocities of the outflow ($\sim$ 30000 km s^{-1}). The Monte Carlo approach is better suited to this problem than the formal integral type solutions. Mazzali & Lucy (1993) include ionization, electron scattering and line scattering in their code. Lucy (1999b) improves the line formation treatment of this code and the noise in the emergent spectrum by using the formal integral for the emergent intensity. Lucy (2005) removes many of the simplifying assumptions in the earlier codes and solves the time-dependent 3-D NLTE transfer in homologously expanding ejecta of a SN, given the distribution of mass and composition at an initial time t_1. Kasen, Min & Nugent (2006) describe a similarly capable code, which also includes polarization and non-grey opacities, that can provide direct comparison between multidimensional hydrodynamic explosion models and observations. Maeda, Mazzali & Nomoto (2006) and Sim (2007) also developed similar codes based on the Lucy methods.

Long & Knigge (2002) applied the methods of Mazzali & Lucy (1993) to calculate line formation and transfer in accretion disk winds. Sim, Drew & Long (2005) extended this code to include ‘macro atoms’, as devised by Lucy (2002, 2003), allowing energy conservation and radiative equilibrium to be enforced at all times. This allows lines formed by non-resonance scattering or recombination to be modeled.

Carciofi & Bjorkman (2006) employed a 3-D non-LTE code to study the temperature and ionization structure of Keplerian disks around classical Be stars. They devised a method independent of Lucy’s (2002) transition probability method to solve the equations of statistical equilibrium. It is similar in many ways, except that the photon absorption and re-emission mechanisms are uncorrelated, allowing them to dispense with Lucy’s macro atoms, along with their associated internal transitions and Monte Carlo transition probabilities. Their models show that the optically thick regions of the disk are similar to Young Stellar Object (YSO) disks and the optically thin outer parts are like stellar winds. Carciofi & Bjorkman (2008) built on their previous work and solved the steady state nonisothermal viscous diffusion and vertical hydrostatic equilibrium of Keplerian disks. Their solution departs significantly from the analytic isothermal density, affecting the emergent spectrum.

4.6.2 *Photoionization*

Several authors describe algorithms for calculating photoionization, e.g., Och *et al.* (1998), Wood & Loeb (2000), Ciardi *et al.* (2001), Maselli, Ferrara & Ciardi (2003), Ercolana *et al.* (2003), Wood, Mathis & Ercolana (2004), Ercolana *et al.* (2008), and Cantalupo &

Porciana (2011). Some particular features of these codes are Wood *et al.*'s (2004) use of photon packets vs energy packets to more easily match the notation of the recombination coefficients; the x-ray extension to the MOCASSIN code to allow computation detailed high-resolution spectra (Ercolano *et al.* 2008); and photoionization on adaptive mesh refinement grids (Cantalupo & Porciani 2011).

These have been applied to the study of escape of ionizing radiation from high-redshift galaxies (Wood & Loeb 2000), cosmological reionization around the first stars (Ciardi *et al.* 2001), modeling the diffuse ionized gas in the Milky Way and other galaxies (Wood & Mathis 2004, photoevaporating planetary disks (Ercolano & Owen 2010), H II regions (Ercolano, Wesson & Bastian 2010 and references therein), and planetary nebulae (Ercolano *et al.* 2004 and references therein), to name a few.

4.6.3 *Chemistry*

The combinations of dust radiative transfer (Section 4.1) and line radiative transfer (Section 4.6.1) can be used to study the chemistry in clouds. Jorgensen *et al.* (2006) iterated on the dust temperature and molecular line calculations to determine where molecules freeze-out in protostellar envelopes. Spaans (1996) included a chemical network of 44 species to study the effects of clumpiness. Bruderer *et al.* (2009a,b, 2010) demonstrated chemical modeling of Young Stellar Objects in a 3-part series. They pre-calculated a grid of chemical composition as a function of time, for a given gas density, temperature, far-UV irradiation and X-ray flux. The local far-UV flux is calculated by a Monte Carlo radiative transfer code, which includes scattering and temperature calculation. The use of the pre-calculated chemical grid speeds up calculations by several orders of magnitude.

5. Summary

The Monte Carlo method for radiative transfer (MCRT) is complementary to the traditional formal methods. While those excel in 1-D, at high optical-depths, incorporating many gas lines and computing detailed spectra, MCRT excels with 3-D geometries, non-LTE gas processes, anisotropic radiation fields and scattering functions, complex velocity fields, and polarization calculations. Thus MCRT is a great tool to add to the set of well-developed methods for radiative transfer. In fact, it is a necessary tool to interpret the ever-increasing sophistication of our new observatories.

Acknowledgments

I thank the editors for inviting me to write this article in celebration of the Chandra Centennial. It is an honour to be asked.

References

Abbott D. C., Lucy L. B., 1985, ApJ, 288, 679

Baes M., Stamatellos D., Davies J. I., Whitworth A. P., Sabatini S., Roberts S., Linder S. M, Evans R., 2005, NewA, 10, 523

Baes M., Vidal E., Van Winckel H., Deroo P., Gielen C., 2007, BaltA, 16, 92

Bernes C., 1979, A&A, 73, 67

Bethell T., Chepurnov A., Lazarian A., Kim J., 2007, ApJ, 663, 1055

Bianchi S., 2008, A&A, 490, 461

Bianchi S., Ferrara A., Davies J. I., Alton P. B., 2000, MNRAS, 311, 601

Bjorkman J. E., Wood K., 2001, ApJ, 554, 615

Boisse P., 1990, A&A, 228, 483

Bruderer S., Doty S. D., Benz A. O., 2009a, ApJS, 183, 179

Bruderer S., Benz A. O., Doty S. D., van Dishoeck E. F., Bourke T. L., 2009b, ApJ, 700, 872

Bruderer S., Benz A. O., Stauber P., Doty S. D., 2010, ApJ, 720, 1432

Cantalupo S., Porciani C., 2011, MNRAS, 411, 1678

Carciofi A. C., Bjorkman J. E., 2006, ApJ, 639, 1081

Carciofi A. C., Bjorkman J. E., 2008, ApJ, 684, 1374

Carter L. L., Cashwell E. D. 1975, Particle-Transport Simulation with the Monte Carlo Method, Energy Research & Development Administration: Los Alamos

Chandrasekhar S., 1946, ApJ, 103, 351

Chandrasekhar S., 1960, Radiative Transfer, Dover, New York

Chakrabarti S., Whitney B. A., 2009, ApJ, 690, 1432

Ciardi B., Ferrara A., Marri S., Raimondo G., 2001, MNRAS, 324 381

Code A. D., 1950, ApJ, 112, 22

Code A. D., Whitney B. A., 1995, ApJ, 441, 400

Cornet C., C-Labonnote L., Szczap F., 2010, JQSRT, 111, 174

Cotera A. S., *et al.*, 2001, ApJ, 556, 958

Coulson K.L., Dave J.V., Sekera Z., 1960, Tables Related to Radiation Emerging from a Planetary Atmosphere with Rayleigh Scattering, University of California Press, Berkeley

Desert F.-X., Boulanger F., Puget J. L., 1990, A&A, 237, 215

Dijkstra M., Loeb A., 2008, MNRAS, 391, 457

Dolence J. C., Gammie C. F., Moscibrodzka M., Leung P., 2009, ApJS, 184, 387

Doty S. D., Metzler R. A., Palotti M. L., 2005, MNRAS, 362, 737

Duchene G., McCabe C., Ghez A. M., Macintosh B. A., 2004, ApJ, 606, 969

Duderstadt J. J., Martin W. R., 1979, Transport Theory, Wiley, New York

Ercolano B., Owen J. E., 2010, MNRAS, 406 1553

Ercolano B., Wesson R., Bastian N., 2010, MNRAS, 401, 1375

Ercolano B., Barlow M. J., Storey P. J., Liu X.-W., 2003, MNRAS, 340, 1136

Ercolano B., Wesson R., Zhang Y., Barlow M. J., De Marco O., Rauch T., Liu X.-W., 2004, MNRAS, 354, 558

Ercolano B., Young P. R., Drake J. J., Raymond J. C., 2008, ApJS, 175, 534

Fernandez R., Thompson C., 2007, ApJ, 660, 615

Fiege J. D., Pudritz R. E., 2000, ApJ, 544, 830

Fischer O., Henning Th., Yorke H. W., 1994, A&A, 284, 187

Green, R., 1985, Spherical Astronomy, Cambridge University Press, Cambridge

Goncalves J., Galli D., Walmsley M., 2004, A&A, 415, 617

Gordon K. D., Misselt K. A., Witt A. N., Clayton G. C., 2001, ApJ, 551, 269

Guhathakurta P., Draine B. T., 1989, ApJ, 345, 230

Hansen M., Oh S. P., 2006, MNRAS, 367, 979

Hartstein D., Liseau R., 1998, A&A, 332, 703

Hatcher Tynes H., Kattawar G. W., Zege E. P., Katsev I. L., Prikhach A. S., Chaikovskaya L. I., 2001, Applied Optics, 40, 400

Henyey L. G., Greenstein J. L., 1941, ApJ, 93, 70

Hillier D. J., 1991, A&A, 247, 455

Hoang T., Lazarian A., 2008, MNRAS, 388, 117

Hoang T., Lazarian A., 2009a, ApJ, 695, 1457

Hoang T., Lazarian A., 2009b, MNRAS, 697, 1316

Hoang T., Draine B. T., Lazarian A., 2010, ApJ, 715, 1462

Hogerheijde M. R., van der Tak F. F. S., 2000, A&A, 362, 697

Indebetouw R., Whitney B. A., Johnson, K. E., Wood K., 2006, ApJ, 636, 362

Jonsson P., 2006, MNRAS, 372, 2

Jorgensen J. K., Johnstone D., van Dishoeck E. F., Doty S. D., 2006, A&A, 449, 609

Juvela M., 1997, A&A, 322, 943

Kalos M. H., Whitlock P. A., 2008, Monte Carlo Methods: Second Revised and Enlarged Edition, Wiley-VCH Verlag, Wenheim, Germany

Kama M., Min M., Dominik C., 2009, A&A, 506, 1199

Kasen D., Thomas R. C., Nugent P., 2006, ApJ, 651, 366

Kenyon S. J., Whitney B. A., Gomez M., Hartmann, L., 1993, ApJ, 414, 773

Knigge C., Woods J. A., Drew J. E., 1995, MNRAS, 273, 225

Kurosawa R., Hillier D. J., 2001, A&A, 379, 336

Kurosawa R., Hillier D. J., Pittard J. M., 2002, A&A, 388, 957

Laursen P., Sommer-Larsen J., Andersen A. C., 2009, ApJ, 704, 1640

Long K. S., Knigge C., 2002, ApJ, 579, 725

Lucas P. W., 2003, JQSRT, 79, 921

Lucas P. W., Roche P. F., 1997, MNRAS, 286, 895

Lucas P. W., Roche P. F., 1998, MNRAS, 299, 699

Lucas P. W. *et al.*, 2004, MNRAS, 352, 1347

Lucy L. B., 1999a, A&A, 344, 282

Lucy L. B., 1999b, A&A, 345, 211

Lucy L. B., 2002, A&A, 384, 725

Lucy L. B., 2003, A&A, 403, 261

Lucy L. B., 2005, A&A, 429, 19

Maeda K., Mazzali P. A., Nomoto K., 2006, ApJ, 645, 1331

Maselli A., Ferrara A., Ciardi B., 2003, MNRAS, 345, 379

Mazzali P. A., Lucy L. B., 1993, A&A, 279, 447

Mihalas, D., 1978, Radiative Transfer, W. H. Freeman and Co., San Francisco

Min M., Dullemond C. P., Dominik C., de Koter A., Hovenier J. W., 2009, A&A, 497, 155

Misselt K. A., Gordon K. D., Clayton G. C., Wolff M. J., 2001, ApJ, 551, 277

Natraj V., Li K. F., Yung Y. L., 2009, ApJ, 691, 1909

Niccolini G., Alcolea J., 2006, A&A, 456, 1

Niccolini G., Woitke P., Lopez B., 2003, A&A, 399, 703

Och S. R., Lucy L. B., Rosa M. R., 1998, A&A, 336, 301

Padoan P., Goodman A., Draine B. T., Juvela M., Nordlund A., Rognvaldsson O. E., 2001, ApJ, 559, 1005

Pagani L., 1998, A&A, 333, 269

Park Y.-S., Hong S. S., 1995, A&A, 300, 890

Park Y.-S., Hong S. S., Minh Y. C., 1996, A&A, 312, 981

Partl A. M., Dall'Aglio A., Muller V., Hensler G., 2010, A&A, 524, A85

Pascucci I., Wolf S., Steinacker J., Dullemond C. P., Henning Th., Niccolini G., Woitke P., Lopez B., 2004, A&A, 417, 793

Pavlyuchenkov Ya., Semenov D., Henning Th., Guilloteau St., Pietu V., Launhardt R., Dutrey A., 2007, ApJ, 669, 1262

Pelkonen V.-M., Juvela M., Padoan P., 2007, A&A, 461, 551

Pelkonen V.-M., Juvela M., Padoan P., 2009, A&A, 502, 833

Pinte C., Menard F., Duchene G., Bastien P., 2006, A&A, 459, 797

Pinte C., Harries T. J., Min M., Watson A. M., Dullemond C. P., Woitke P., Menard F., Duran-Rojas M. C., 2009, A&A, 498, 967

Pontoppidan K. M., Dullemond C. P., Blake G. A., Evans II N. J., Geers V. C., Harvey P. M., Spiesman W., 2007, ApJ, 656, 991

Press W. H., Teukolsky S.A., Vetterling, W. T., Flannery B. P. 2007, Numerical Recipes 3rd Edition, Cambridge University Press, Cambridge

Puls J., Vink J. S., Najarro F., 2008, A&ARv, 16, 209

Robitaille T. P., 2010, A&A, 520, A70

Rosseland S., 1924, MNRAS, 84, 525

Schartmann M., Meisenheimer K., Camenzind M., Wolf S., Tristram K. R. W., Henning T., 2008, A&A, 482, 67

Schneider G., Wood K., Silverstone M. D., Hines D. C., Koerner D. W., Whitney B. A., Bjorkman J. E., Lowrance P. J., 2003, AJ, 125, 1467

Sim S. A., 2007, MNRAS, 375, 154

Sim S. A., Drew J. E., Long K. S., 2005, MNRAS, 363, 615

Spaans M., 1996, A&A, 307, 271

Springel V., Di Matteo, T., Hernquist L., 2005, ApJ, 620, 79

Stamatellos D., Whitworth A. P., 2003, A&A, 407, 941

Stamatellos D., Whitworth A. P., Ward-Thompson D., 2004, A&A, 420, 1009

Stark D. P., Whitney B. A., Stassun K., Wood K., 2006, ApJ, 649, 900

Stern B. E., Begelman M. C., Sikora M., Svensson R., 1995, MNRAS, 272, 291

Sundqvist J. O., Puls J., Feldmeier A., 2010, A&A, 510, A11

Verhamme A., Schaerer D., Maselli A., 2006, A&A, 460, 397

Wang J. C. L., Wasserman I. M., Salpeter E. E., 1988, ApJS, 68, 735

Watson A. M., Henney W. J., 2001, RMxAA, 37, 221

Watson A. M., Stapelfeldt K. R., 2004, ApJ, 602, 860

Watson A. M., Stapelfeldt K. R., 2007, AJ, 133, 845

Watson A. M., Stapelfeldt K. R., Wood K., Menard, F. 2007, in Reipurth B., Jewitt D., Keil K., eds, Protostars and Planets V, University of Arizona Press, Tucson, p. 523

Wehrse R., Baschek B., von Waldenfels W., 2000, A&A, 359, 780

White R. L., 1979, ApJ, 229, 954

Whitney B. A., 1991a, ApJS, 75, 1293

Whitney B. A., 1991b, ApJ, 369, 451

Whitney B. A., Hartmann L., 1992, ApJ, 395, 529

Whitney B. A., Hartmann L., 1993, ApJ, 402, 605

Whitney B. A., Wolff M. J., 2002, ApJ, 574, 205

Whitney B. A., Kenyon S. J., Gomez M., 1997, ApJ, 485, 703

Whitney B. A., Wood K., Bjorkman J. E., Wolff M. J., 2003a, ApJ, 591, 1049

Whitney B. A., Wood K., Bjorkman J. E., Cohen M., 2003b, ApJ, 598, 1079

Witt A. N., 1977a, ApJS, 35, 1

Witt A. N., 1977b, ApJS, 35, 7

Witt A. N., 1977c, ApJS, 35, 21

Witt A. N., Gordon K. D., 1996, ApJ, 463, 681

Witt A. N., Gordon K. D., 2000, ApJ, 528, 799

Witt A. N., Oshel E. R., 1977, ApJS, 35, 31

Witt A. N., Thronson Jr H. A., Capuano Jr J. M., 1992, ApJ, 393, 611

Wolf S., 2003, CoPhC, 150, 99
Wolf S., Henning Th., 2000, CoPhC, 132, 166
Wolf S., Henning Th., Stecklum B., 1999, A&A, 349, 839
Wolf S., Padgett D. L., Stapelfeldt K. R., 2003, ApJ, 588, 373
Wood K., Loeb A., 2000, ApJ, 545, 86
Wood K., Mathis J. S., 2004, MNRAS, 353, 1126
Wood K., Whitney B. A., 1998, ApJ, 506, 43
Wood K., Bjorkman J. E., Whitney B. A., Code A. D., 1996, ApJ, 461, 828
Wood K., Mathis J. S., Ercolano B., 2004, MNRAS, 348, 1337
Wood K., Whitney B. A., Robitaille T., Draine B. T., 2008, ApJ, 688, 1118
Yusef-Zadeh F., Morris M., White R. L., 1984, ApJ, 278, 186
Zheng Z., Miralda-Escude J., 2002, ApJ, 578, 33
Zheng Z., Cen R., Trac H., Miralda-Escude J., 2010, ApJ, 716, 574

Bull. Astr. Soc. India (2011)

Astrophysical magnetohydrodynamics

James M. Stone[*]

Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08540, USA

Received March 12, 2011; accepted March 25, 2011

Abstract. Over the course of roughly a decade, from the late 1950s through the early 1960s, Chandraskhar made fundamental contributions to basic plasma physics, and the effect of magnetic fields on the dynamics of astrophysical plasmas. This paper reviews recent progress and outstanding problems in *Astrophysical magnetohydrodynamics*, the application of MHD to astrophysical systems, with particular emphasis on the role of Chandra's early contributions to the field. Specific topics discussed include magnetic field amplification by dynamo processes inside stars, the magnetorotational instability and angular momentum transport in accretion disks, MHD turbulence in the interstellar medium of galaxies, and kinetic MHD effects in weakly collisional plasmas. Chandra's contributions in all of these areas endure.

Keywords : MHD – turbulence – accretion disks

1. Introduction

It was a great honour and privilege to speak at the Chandrasekhar Centennial Symposium on the topic of 'Astrophysical Magnetohydrodynamics', especially since there were so many eminent members of the audience whom I would have liked to hear speak on the same topic! The goals of my talk were to provide a summary of recent progress in magnetohydrodynamics (MHD) as applied to a wide variety of astrophysical systems, and to highlight Chandra's early contributions to these topics. The goals of this paper are the same.

Unfortunately, by the time I was a graduate student in the late 1980s, Chandra was no longer working on plasma physics, and therefore I never had the opportunity to meet him personally. However, he still had an enormous impact on me, as on most graduate students, through his books. In particular, his books on radiative transfer (Chandrasekhar 1950), hydrodynamic and hydromagnetic stability (Chandrasekhar 1961), and ellipsoidal figures of equilibrium (Chandrasekhar 1969), all still available as Dover reprints, are as relevant today as they were back then.

Throughout the 1950s and 1960s, Chandra wrote many papers on MHD and plasma physics, following four general themes:

[*] e-mail: jmstone@princeton.edu

1. the statistical properties of turbulence,
2. problems in astrophysical MHD,
3. basic plasma physics, and
4. hydrodynamic and MHD instabilities.

Chandraskhar (1989a,b), volumes 3 and 4 of his selected papers, contain his most important work in these areas. Rather than highlighting individual papers or results, instead I have organized this paper around astrophysical objects of increasing scale. Thus, after a brief introduction to some general concepts in MHD, I will discuss evidence for the importance of magnetic fields first in *stars*, then in *accretion disks*, then in *galaxies*, and finally on the largest scale in *clusters of galaxies*. Each topic will be organized into a separate section.

Finally, it is useful to highlight what Chandra himself wrote about astrophysical MHD back in 1957: “It is clear we are very far from an adequate characterization of cosmic magnetic fields” (Chandrasekhar 1957). Obviously we have come very far since 1957, but in some cases it is clear we still have very far to go.

2. Some elementary MHD

Before discussing results, it is worthwhile to summarize some basic physics of MHD. In a highly collisional plasma with perfect conductivity, the equations of motion are essentially the Euler equations of gas dynamics, supplemented with Maxwell’s equations to describe the evolution of the magnetic field (in particular, Faraday’s Law). The result, usually referred to as the equations of ideal MHD, is

$$\frac{\partial \rho}{\partial t} + \nabla\cdot[\rho\mathbf{v}] = 0, \tag{1}$$

$$\frac{\partial \rho\mathbf{v}}{\partial t} + \nabla\cdot\left[\rho\mathbf{v}\mathbf{v} - \mathbf{B}\mathbf{B} + \mathsf{P}^*\right] = 0, \tag{2}$$

$$\frac{\partial E}{\partial t} + \nabla\cdot\left[(E + P^*)\mathbf{v} - \mathbf{B}(\mathbf{B}\cdot\mathbf{v})\right] = 0, \tag{3}$$

$$\frac{\partial \mathbf{B}}{\partial t} - \nabla\times(\mathbf{v}\times\mathbf{B}) = 0, \tag{4}$$

where P^* is a diagonal tensor with components $P^* = P + B^2/2$ (with P the gas pressure), E is the total energy density

$$E = \frac{P}{\gamma - 1} + \frac{1}{2}\rho v^2 + \frac{B^2}{2}, \tag{5}$$

and $B^2 = \mathbf{B}\cdot\mathbf{B}$. The other symbols have their usual meaning. These equations are written in units such that the magnetic permeability $\mu = 1$. An equation of state appropriate to an ideal gas, $P = (\gamma - 1)e$ (where γ is the ratio of specific heats, and e is the internal energy density), has been assumed in writing Eq. (5). These equations are valid only for non-relativistic flows, and for phenomena at frequencies much less than the plasma frequency. As we shall see in Section 6 there are many interesting frontiers to explore as some of the assumptions underlying the equations of ideal MHD are relaxed, for example in low collisionality plasmas.

Restricting ourselves to one dimensional flow for the moment, it is useful to rewrite the equations of motion in a compact form

$$\frac{\partial \mathbf{U}}{\partial t} = \frac{\partial \mathbf{F}}{\partial \mathbf{U}} \frac{\partial \mathbf{U}}{\partial x} \tag{6}$$

where the components of the vectors **U** and **F** are the conserved variables and their fluxes, respectively, that is

$$\mathbf{U} = \begin{bmatrix} \rho \\ M_x \\ M_y \\ M_z \\ E \\ B_y \\ B_z \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \rho v_x \\ \rho v_x^2 + P + B^2/2 - B_x^2 \\ \rho v_x v_y - B_x B_y \\ \rho v_x v_z - B_x B_z \\ (E + P^*) v_x - (\mathbf{B} \cdot \mathbf{v}) B_x \\ B_y v_x - B_x v_y \\ B_z v_x - B_x v_z \end{bmatrix}. \tag{7}$$

Note that Eq. (6) defines a system of nonlinear hyperbolic partial differential equations (PDEs). The mathematical properties of hyperbolic PDEs are well studied. In particular, the eigenvalues of the Jacobian $\partial \mathbf{F}/\partial \mathbf{U}$ define the characteristic (wave) speeds in MHD.

In fact, probably the most important property of hyperbolic PDEs is that they admit wave-like solutions. Much of the dynamics of magnetized plasmas can be interpreted using the properties of linear and nonlinear wave solutions. The properties of linear waves can be studied using the dispersion relation, derived by looking for solutions for small amplitude disturbances of the form $\exp i(\omega t + \mathbf{k} \cdot \mathbf{x})$, where ω is the frequency and **k** the wavevector, in a stationary, isotropic, homogeneous medium. Inserting this form for the solution into the equations of motion, and keeping only terms which are linear in the disturbance amplitude results in a system of linear equations, which have solutions only if the frequency and wavenumber are related through the following dispersion relation

$$\left[\omega^2 - (\mathbf{k} \cdot \mathbf{V}_A)^2\right]\left[\omega^4 - \omega^2 k^2 \left(V_A^2 + C^2\right) + k^2 C^2 (\mathbf{k} \cdot \mathbf{V}_A)^2\right] = 0 \tag{8}$$

where $\mathbf{V}_A = \mathbf{B}/\sqrt{4\pi\rho}$ is the Alfvén velocity, and $C^2 = \gamma P/\rho$ the adiabatic sound speed. The dispersion relation has three pairs of solutions, which represent right- and left-going waves of three different families. (Note that MHD is immediately different from hydrodynamics, which has only one wave family: sound waves). The MHD wave families are the Alfvén wave (an incompressible transverse wave propagating at speed V_A), and the fast and slow magnetosonic waves (which are both compressible acoustic modes with phase velocity modified by the magnetic pressure). To complicate matters even more, the phase velocity for each mode depends on the angle between the wavevector and the magnetic field, as well as the strength of the magnetic field as measured by the ratio V_A/C. The angular dependence is most easily demonstrated using Friedrichs diagrams, which plot the relative phase velocity of each mode versus the angle between **k** and **B** in a polar diagram (see Section 14.1 in Sturrock 1994 for an example). Such plots clearly demonstrate important properties of MHD waves, for example, for directions parallel to the magnetic field, the Alfvén wave has the same phase velocity as either the fast or the slow magnetosonic wave (which one depends on whether the Alfvén speed is faster or slower than the sound speed). In this case, the modes are degenerate. Mathematically, this reflects the fact that the equations of MHD are not *strictly hyperbolic*, since in some circumstances the eigenvalues of the Jacobian are degenerate. This fact makes finding solutions to the equations of ideal MHD even more complicated.

Another important property of MHD waves in comparison to hydrodynamics is that, because they involve transverse motions, Alfvén waves can be polarized. The sum of two linear polarizations with different phase shifts can lead to circularly polarized Alfvén waves. This means in MHD, all three components of velocity must be kept, even in one dimensional flows, in order to represent all polarizations. Moreover, in non-ideal MHD the left- and right-circularly polarized Alfvén waves can have different phase velocities, these are the whistler waves in the Hall MHD regime (where ions and electrons can drift due to collisions with neutrals). Again, this new behavior is a direct consequence of the complexity of MHD waves, and it is fair to say that the rich dynamics of MHD results in part from this complexity.

Finding analytic solutions to the equations of MHD, beyond those representing linear waves, is very difficult. Usually, very restrictive assumptions are required, such as steady (so that $\partial/\partial t = 0$), and/or one dimensional flow. Today, the most important tools for solving the MHD equations are numerical methods. Grid based methods for MHD are now quite mature, and a variety of public codes are available to study MHD flows in fully three-dimensions, including a rich set of physics beyond ideal MHD. Most grid based methods for MHD adopt the same approach: the conserved variables are discretized on a grid, with volume averaged values stored at cell centers. In order to enforce the divergence-free constraint on the magnetic field, it is better to store area averages of each component of the magnetic field at corresponding cell faces, and evolve these components using electric fields at cell edges, using a technique called “constrained transport”. Figure 1 shows the basic discretization of the variables.

One example of a publicly-available grid code for MHD is Athena (Stone et al. 2008), available at `https://trac.princeton.edu/Athena`. Athena implements a higher-order Godunov scheme based on directionally unsplit integrators, piecewise-parabolic reconstruction, and constrained transport, with a variety of Riemann solvers available to compute the fluxes. With this approach, mass, momentum, energy, and magnetic flux are all conserved to machine precision. Of course, there are many other codes available which implement different algorithms than those used in Athena, and this is a very good thing, because by comparing solutions to the same problem generated by different algorithms, we can gauge whether those solutions are reliable. Throughout the rest of this paper, I will discuss solutions to MHD problems generated by Athena and other codes.

3. Solar magnetoconvection

The best evidence of the importance of magnetic fields to the dynamics of astrophysical plasmas comes from observations of the outer layers of the Sun. Both the presence of

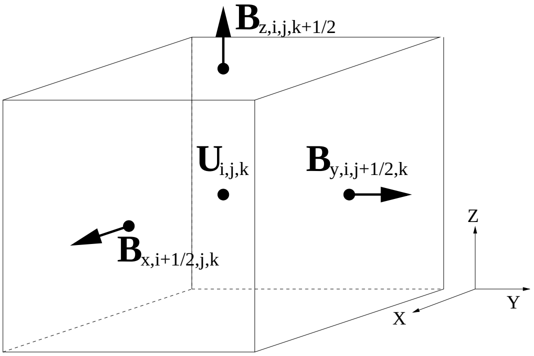

$\mathbf{B}_{z,i,j,k+1/2}$
$\mathbf{U}_{i,j,k}$
$\mathbf{B}_{y,i,j+1/2,k}$
$\mathbf{B}_{x,i+1/2,j,k}$
Z
Y
X

Figure 1. Basic centering of variables for a grid-based numerical method for MHD using constrained transport. Volume averages of conserved variables are stored at cell centers, while area averages of each component of the magnetic field are stored at cell faces.

sunspots in the photosphere, and structures such as filaments, prominences, and flares in the solar corona, demonstrate the key role that magnetic fields play in shaping the dynamics. In fact, the very existence of the hot corona is now interpreted as due to heating by MHD effects. Beautiful images and animations that show magnetic fields in action in the solar corona have been obtained by recent spacecraft missions such as SOHO, TRACE, Yokoh, Hinode, and SDO.

It is thought that most of the magnetic activity of the Sun is driven by the combination of rotation and turbulent flows in the convection zone. In fact, the properties of MHD turbulence driven by convection was one of the problems that first interested Chandra in plasma physics (for examples, see papers in Chandrasekhar 1989a).

Understanding the origin and evolution of the Sun's magnetic field via a dynamo process has been a challenging problem for many decades. In addition to generation of the dipole field due to differential rotation, a process first proposed by Parker (1955), there are also small-scale multipole fields thought to be generated by the convective turbulence that play a role in shaping sunspots and coronal activity. Both the processes that produce sunspots, and the large-scale magnetic field of the Sun, are very active areas of research.

In the case of sunspots, direct numerical simulations of magnetoconvection in the outer layers, including realistic radiative transfer to capture the outer radiative zone, can now reproduce details of observed sunspots, including the penumbral filaments; a beautiful example is given in Rempel *et al.* (2009).

In the case of the solar dynamo, the dipole field is now thought to originate in the *tachocline*, a region of strong shear between the radiative core (which is in solid body rotation, according to results from helioseismology) and the outer convective zone (which is in differential rotation). However, although the sophistication of modern global MHD simulations of magnetoconvection in spherical and rotating stars is impressive, they still fail to explain both the origin of the differential rotation in the convective zone, and the origin of the cyclic dipole field. Solving the solar dynamo problem is important, as we are unlikely to understand magnetic fields in other stars if we cannot first understand the Sun.

4. The MRI in accretion disks

Moving beyond stars, the next set of astrophysical systems where magnetic fields have been identified as being important is accretion disks. Such disks are ubiquitous, occurring in protostellar systems, close binaries undergoing mass transfer, and in active galactic nuclei.

The most basic property of an accretion disk is the angular momentum transport mechanism. This mechanism controls the rate of accretion, which in turn controls the luminosity, variability, and spectrum of the disk. Mass accretion in disks is analogous to nuclear fusion in stars: it is the mechanism that powers the entire system.

It has been known for decades that kinetic viscosity in an astrophysical plasma is too small to explain the angular momentum transport and mass accretion rate, so that some form of "anomalous" viscosity is required (Shakura & Sunyaev 1973). It has also been long suspected that the transport was associated with turbulence in the disk, but disks with Keplerian rotation profiles are linearly stable according to the Rayleigh criterion, that is, so long as the specific angular momentum increases outwards. So the question becomes: what drives turbulence in disks?

The answer seems to be: magnetic fields. Remarkably, disks with Keplerian rotation profiles which contain weak magnetic fields (weak in the sense that the gas pressure is larger than the magnetic pressure) are linearly *unstable* to the magnetorotational instability (MRI), as first recognized by Balbus & Hawley (1991). The MRI can be identified by calculating the linear dispersion relation for MHD waves in a Keplerian shear flow. The simplest analysis which captures the MRI assumes incompressible axisymmetric perturbations, a purely vertical magnetic field, and ideal MHD (all of these assumptions have been relaxed in later analyses, e.g. see Balbus & Hawley 1999 for a review). The resulting dispersion relation is

$$\omega^4 - \omega^2 \left[\kappa^2 + 2\,(\mathbf{k}\cdot\mathbf{V}_A)^2\right] + (\mathbf{k}\cdot\mathbf{V}_A)^2 \left([\mathbf{k}\cdot\mathbf{V}_A]^2 + \frac{d\Omega^2}{d\ln r}\right) = 0 \tag{9}$$

where V_A is the Alfvén speed, and

$$\kappa^2 = \frac{1}{R^3}\frac{d(R^4\Omega^2)}{dR} \tag{10}$$

is the epicyclic frequency (R is the cylindrical radius). Note that the coefficient of the first and second terms in Eq. (9) are positive and negative respectively, therefore solutions with $\omega^2 < 0$ (that is, instability) are possible if the third term is negative. This occurs when

$$(\mathbf{k}\cdot\mathbf{V}_A)^2 < -\frac{d\Omega^2}{d\ln r} \tag{11}$$

Physically, this states that if the rotation frequency in the disk is decreasing outwards (as is true in Keplerian flows), then there are always sufficiently small wavenumbers that will be unstable. Note that this is in direct contradiction to the Rayleigh criterion, which requires the angular *momentum* (not frequency) decrease outward for instability. How small is "sufficiently small" for instability depends on the magnetic field strength (V_A). In practice, if the field is weak ($V_A < C$), there always are unstable modes with wavenumbers large

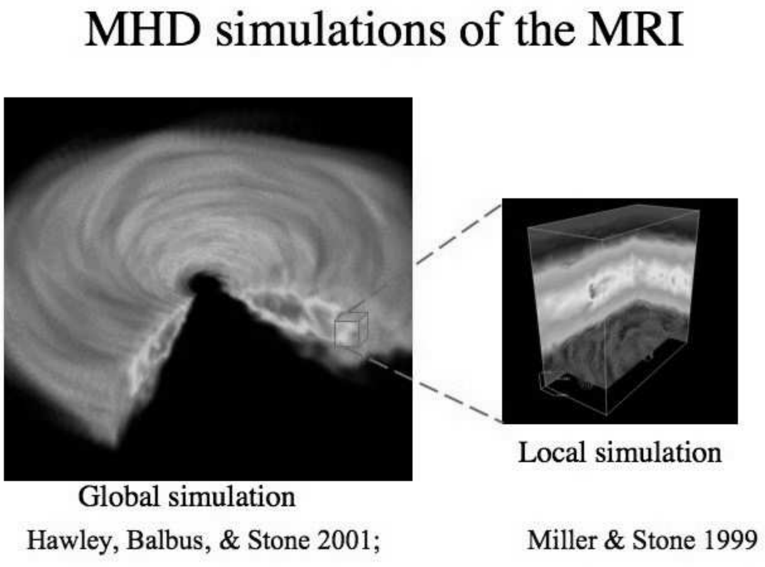


Figure 2. Images of the density from a global simulation of a MRI unstable disk (*left*), and of the density and magnetic field vectors from a local shearing box simulation (*right*).

enough that the corresponding wavelength is less than the vertical scale height (thickness) of the disk.

In fact, studies of the MRI have a long and interesting history. The MRI was first identified by Velikhov (1959) in a study motivated by a rotating plasma experiment. Chandrasekhar (1960) made important contributions, showing the instability was present in a global analysis of magnetized Couette flow. Fricke (1969) found the instability in differentially rotating stars. However, the importance of the MRI to accretion disks was not recognized by any of these authors, in fact Safronov (1972) argued that the inclusion of finite resistivity and viscosity effects would make the MRI unimportant in disks. A key element of confusion seems to be over the lack of recovery of the Rayleigh criterion as the magnetic field strength is decreased to zero. The stability properties of hydrodynamic flows (based on angular momentum gradients) and MHD flows (based on angular velocity gradients) are incompatible, a point discussed in detail by Balbus & Hawley (1991). It was not until their paper that the important role that the MRI plays in disks was identified.

Over the past 20 years, there has been considerable effort to understand the nonlinear regime and saturation of the MRI, mostly using computational methods. Figure 2 shows images from typical simulations of the MRI in both *global* domains, in which the entire disk is evolved over a wide range of radii, and *local* shearing box simulations, in which only a small radial extent of the disk is evolved. The advantage of the shearing box is that by focusing all of the computational resources on a small patch, much higher numerical resolution is possible.

Perhaps the most important result from local shearing box simulations is that in the nonlinear regime, the MRI produces MHD turbulence which has both significant Maxwell and Reynolds stresses that transport angular momentum outward. It is remarkable that the inclusion of a weak field *qualitatively* changes the stability properties of the flow, and results in outward transport at a level required by observations. Numerical simulations of the MRI have also established that turbulence amplifies the magnetic field, and drives an MHD dynamo, and that the power spectrum of the turbulence is anisotropic, with most of the energy on the largest scales (Balbus 2003).

Still, many important questions remain. At the moment, it is not understood how the energy liberated by accretion is dissipated by the turbulence: does most of the energy go into the ions or electrons? It is not understood how MRI unstable disks drive powerful winds and outflows as are observed in many astrophysical systems, and what are the relative contributions of the MRI and winds to angular momentum transport. Finally, calculations which include radiation have only begun to be explored; it is likely many important phenomena may be related to the interaction of the radiation field with the flow field generated by the MRI. All of these questions will undoubtedly be addressed by future efforts.

5. MHD turbulence in the ISM of galaxies

Moving to ever larger scales, the next system in which magnetic fields have been observed to be important is the interstellar medium (ISM) of galaxies. The observation of polarized synchrotron emission from the ISM of the Milky Way and other galaxies, produced by relativistic electrons spiraling around magnetic field lines, is direct proof of the presence of such fields. Moreover, the observations allow the strength and even the direction of the field to be inferred. In most cases, it is found the fields are in equipartition, with the magnetic energy density being about equal to the thermal energy of the gas, and kinetic energy of relativistic particles. Moreover, observations of the kinematics of the ISM in galaxies reveal it is highly turbulent. Thus, interpretation of the dynamics of the ISM requires an understanding of highly compressible MHD turbulence.

In fact, the statistical properties of turbulence were of considerable interest to Chandra. It is revealing to read what he wrote in his Henry Norris Russell Lecture:

> *We cannot construct a rational physical theory without an adequate base of physical knowledge. It would therefore seem to me that we cannot expect to incorporate the concept of turbulence in astrophysical theories in any essential manner without a basic physical theory of the phenomenon of turbulence itself,* (Chandrasekhar 1949).

Fortunately, the theory of energy cascades in strong MHD turbulence has progressed enormously in the last few decades (e.g. Goldreich & Sridhar 1995), so that there now are theories of the power spectrum and statistical properties of MHD turbulence that can be tested and compared to observation. One method to investigate the properties of MHD turbulence is through direct numerical simulation.

Figure 3 shows images from high resolution (1024^3) numerical simulations of highly compressible MHD turbulence with both strong and weak magnetic fields, taken from Lemaster & Stone (2009). The turbulence is driven with a forcing function whose spatial power spectrum is highly peaked at a wavenumber corresponding to about $1/8$ the size of the computational domain. The energy input rate of the driving is held constant, and the turbulence is driven so that the Mach number of RMS velocity fluctuations $M = \sigma_V/C$

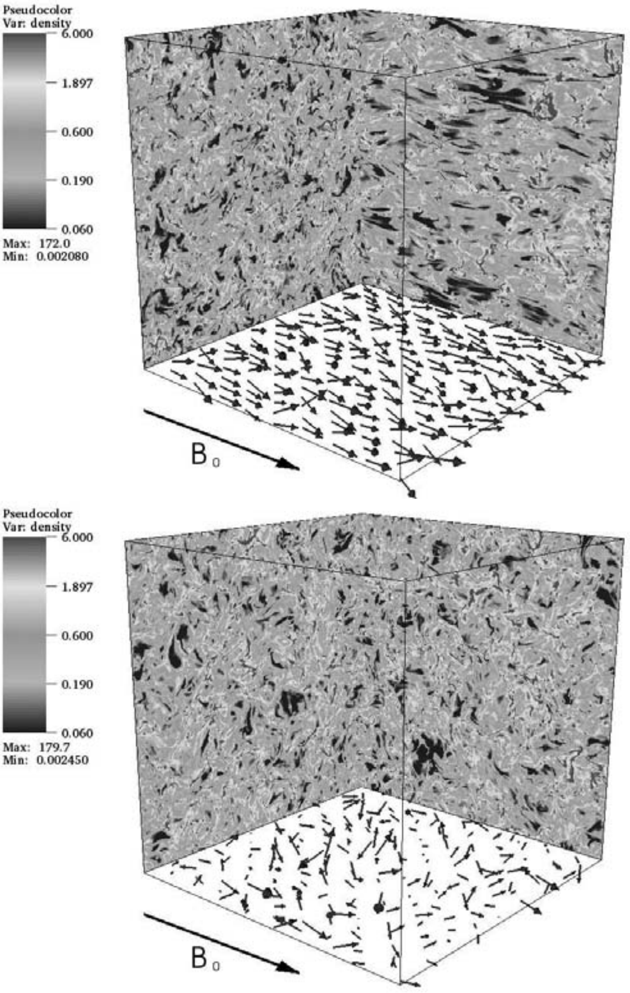


Figure 3. Structure of the density (grayscale) and magnetic field (arrows) in driven supersonic MHD turbulence for strong (top) and weak (bottom) fields.

(where C is the sound speed) is about 7. The magnetic field strength corresponds to a ratio of gas to magnetic pressure $\beta = 8\pi P/B^2$ of 0.01 in the strong field case, and one in the weak field case. This means the Alfvénic Mach number of the turbulence is about one in the strong field case, and 7 in the weak field case.

It is quite clear from the images that in the weak field case, the density fluctuations are isotropic, and the magnetic field is highly tangled. In contrast, in the strong field case the density fluctuations are elongated along the field lines, and the field is more or less ordered. This suggests that the power spectrum of the turbulence will be anisotropic. In fact, this is one of the most basic predictions of the theory (Goldreich & Sridhar 1995).

In addition to investigating the spectrum of fluctuations, such simulations can be used to measure properties such as the decay rate of the turbulence, and how it depends on the magnetic field strength. Early predictions suggested the decay rate of strongly magnetized turbulence would be very low, since it would be dominated by incompressible Alfvén waves. In fact, the simulations (Stone, Ostriker & Gammie 1998; MacLow 1999) found the decay rate of *supersonic* MHD turbulence was very fast, with the decay time about equal to an eddy turn over time on the largest scales, regardless of the field strength. Most of the dissipation was found to occur in shocks. Thus, while Alfvén waves are important to the energetics, the coupling of large amplitude nonlinear Alfvén waves to compressible modes, in particular slow magnetosonic waves, cannot be ignored. This coupling pumps energy into the compressible modes, which then decay in shocks. The result has important implications for the decay of supersonic turbulence in the ISM of galaxies.

Finally, more direct comparison between the simulations and observations is possible using properties such as the polarization angle of background star light. In many regions of the ISM, spinning dust grains become aligned with their long axis perpendicular to the magnetic field. When background stars are viewed through these aligned grains, their light is polarized, with the strength and direction of the polarization vector related to the column density of gas, and the magnetic field strength in the plane of the sky. Using numerical simulations of MHD turbulence, it is possible to compute theoretical maps of the polarization vectors along different viewing angles for background sources viewed through the simulation domain. Figure 4 shows an example for two simulations, both using Mach 10 turbulence with strong ($\beta = 0.01$) and weak ($\beta = 1$) magnetic fields.

It is clear from inspection that in the case of strong fields, the scatter in polarization angle is small, while in the case of weak fields the scatter is large. In fact, this effect was predicted by Chandrasekhar & Fermi (1953), who showed that the scatter in the polarization angle $\delta\phi$ should be related to the plane-of-sky magnetic field strength B_p, gas density ρ, and line-of-sight velocity dispersion δv through

$$B_p = 0.5\frac{(4\pi\rho)^{1/2}\delta v}{\delta\phi} \tag{12}$$

Equation (12) is now known as the “Chandrasekhar-Fermi” formula, and is now routinely used as a technique to measure magnetic field strengths in the ISM.

6. Kinetic MHD effects in clusters of galaxies

Finally, we consider the effect of magnetic fields on the largest structures in the universe, clusters of galaxies. Radio observations of Faraday rotation in background sources indicate that the x-ray emitting plasma trapped in the gravitational potential of clusters is magnetized. Using the x-ray spectra to determine the temperature and density of the plasma shows that the mean free path of charged particles in the plasma is much smaller than the system size, but much larger than the gyroradius, that is the plasma is in the *kinetic MHD* regime.

The most important property of weakly collisional plasmas in the kinetic MHD regime, in comparison to highly collisional plasmas, is that the microscopic transport coefficients become anisotropic. For example, if the electron mean free path is much larger than the

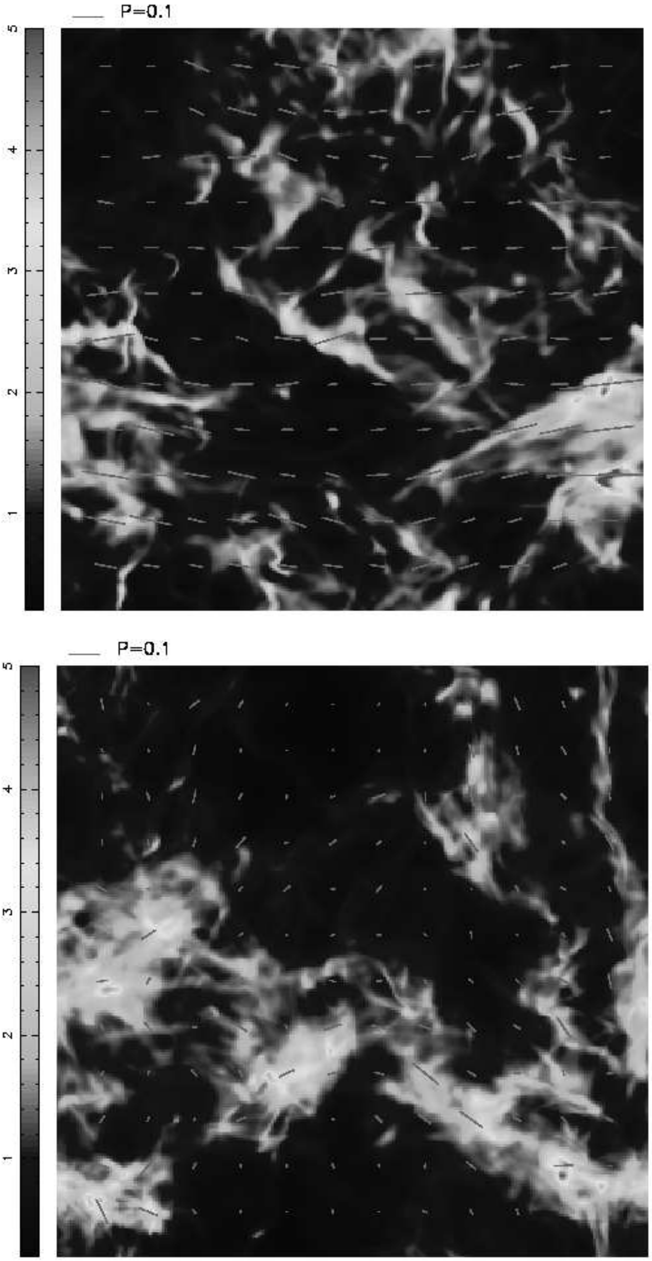


Figure 4. Scatter in polarization angle in supersonic turbulence with a strong field (*top*) and weak field (*bottom*). The grayscale shows the column density, and the line segments show the direction and amplitude of the polarization vector.

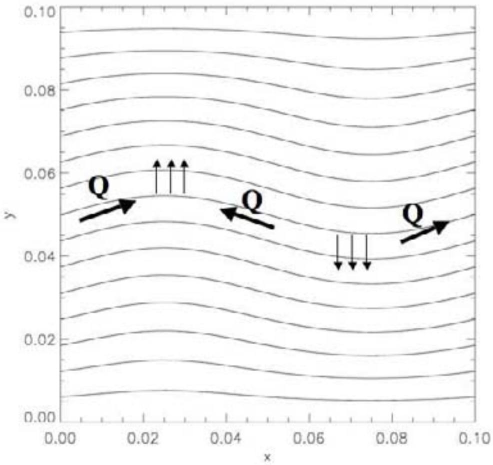


Figure 5. Basic mechanism of the MTI. The structure of the perturbed field lines in a stratified atmosphere are shown (which is hotter on the bottom than top), along with the direction of the heat flux **Q** induced along field lines which results in amplification of the perturbations.

electron gyroradius, thermal conduction is primarily along magnetic field lines. Similarly, when the ion mean free path is much larger than the ion gyroradius, kinematic viscosity is primarily along magnetic field lines. The simplest description of the dynamics is therefore given by the equations of MHD supplemented by anisotropic thermal conduction and viscous transport terms (Braginskii 1965).

Remarkably, the addition of anisotropic transport *qualitatively* changes the dynamics of the plasma. For example, with anisotropic thermal conduction, the convective stability criterion no longer depends on entropy, but only on the temperature gradient (if $dT/dz < 0$, the plasma is *unstable* to convection; Balbus 2000). Convective instability in this regime has been termed the magnetothermal instability (MTI). In fact, other instabilities have also been found in the kinetic MHD regime that might be important in clusters (Quataert 2008) or in diffuse accretion flows (Balbus 2000).

Figure 5, taken directly from a nonlinear simulation (Parrish & Stone 2007), demonstrates the physics of the MTI. Consider a stratified atmosphere in a constant gravitational field. Arrange the vertical profiles of the pressure and density so that the atmosphere is hotter at the bottom than the top, and so that the entropy is constant or increasing upwards. In this case, the atmosphere should be *stable* to convection by the Schwarzschild criterion. Now consider a weak, horizontal magnetic field with anisotropic thermal conduction along field lines. Initially the field lines are parallel to the isotherms, so there is no heat flux in the equilibrium state. Now consider the evolution of vertical perturbations, as shown in the figure. The peaks of the perturbations are at a slightly lower pressure than their equilibrium position, so they expand and cool. The valleys are at a slightly higher pressure, and so contract and heat up. These lead to a temperature gradient, and therefore a heat flux **Q**, along the field lines. The net result is to *increase* the entropy at the peaks (making them more buoyant), and to *decrease* the entropy at the valleys (making them sink). This increases the perturbation, tilts the field line more to the vertical, increases the temperature gradient along the field line and therefore increases the heat flux; and this process runs away as an instability.

The nonlinear regime of the MTI has now been quite well studied using numerical simulations. With non-conducting boundaries at the top and bottom of the domain, the MTI saturates when the temperature profile becomes isothermal. If the top and bottom boundaries are held at fixed temperatures, then vigorous and sustained convection can be driven.

How does the MTI relate to galaxy clusters? Recent work shows that it can play an important role in the temperature profiles of the x-ray emitting gas. When clusters form from gravitational collapse of large-scale structure, the initial temperature profile is centrally peaked. This profile is unstable to the MTI, and simulations of hydrostatic cores with weak magnetic fields show that the MTI causes significant redistribution of the temperature profile of the cluster, along with significant amplification of the magnetic field, in a Hubble time. More recently, the role that externally driven turbulence plays in the plasma dynamics, along with the MTI and other instabilities in the kinetic regime, has been an area of active inquiry (for example, see Parrish, Quataert & Sharma 2009).

7. Summary

I have discussed a very wide range of astrophysical systems where magnetic fields modify or even control the dynamics in order to demonstrate that MHD is now understood to be fundamental to many basic problems in astrophysics. Perhaps the best example is provided by the problem of angular momentum transport in accretion disks. For over thirty years, it was a struggle to understand why such transport occurs. With the identification of the MRI, it became clear that MHD is the key: weakly magnetized Keplerian shear flows are linearly unstable, and subsequent computational studies have shown this instability saturates as MHD turbulence with a significant Maxwell stress. In fact, both Velikhov (1959) and Chandrasekhar (1960) recognized the presence of the instability, although neither realized its importance in accretion disks, perhaps because such disks were not well recognized observationally at the time.

Many frontiers exist in astrophysical MHD, as Section 6 demonstrates. Motivated by the properties of weakly collisional plasmas in the x-ray emitting gas in clusters of galaxies, anisotropic thermal conduction was shown to qualitatively change the dynamics. In particular, it has been found that the stability condition for convection is fundamentally altered when anisotropic conduction is important: stability depends only on the temperature gradient, while the entropy profile is irrelevant. Undoubtedly, many more remarkable results remain to be discovered as ever more realistic descriptions of astrophysical plasmas are adopted.

It is impossible to describe studies of astrophysical MHD without mentioning the important role that numerical methods now play. In fact, computational methods are now the primary tool for the investigation of nonlinear, time-dependent, and multidimensional solutions to the equations of MHD. I wonder what Chandra would think of modern computational methods, and their application to problems in astrophysics?

Finally, I hope this paper has demonstrated that Chandra’s contributions to plasma physics and MHD endure. In particular, his work on the MRI was before its time.

Acknowledgments

I thank the organizers of the symposium for the opportunity to speak, and for running such an interesting and productive meeting. I thank my collaborators who have contributed significantly to the work presented here, in particular S. Balbus, C. Gammie, T. Gardiner, J. Hawley, N. Lemaster, E. Ostriker, and I. Parrish. This work has been supported by grants from the DOE, NSF and NASA ATP programs.

References

Balbus S.A., Hawley J.F., 1991, ApJ, 376, 214
Balbus S.A., Hawley J.F., 1999, Rev. Mod. Phys., 70, 1
Balbus S.A., 2000, ApJ, 534, 420
Balbus S.A., 2003, ARA&A, 41, 555
Braginskii S.I.., 1965, Rev. Pl. Phys., 1, 205
Chandrasekhar S., 1949, ApJ, 110, 329
Chandrasekhar S., 1950, Radiative Transfer, Dover Publications, New York
Chandrasekhar S., Fermi E., 1953, ApJ, 118, 113
Chandrasekhar S., 1957, Proc. Nat. Acad. Sci., 43, 24
Chandrasekhar S., 1960, Proc. Nat. Acad. Sci., 46, 253
Chandrasekhar S., 1961, Hydrodynamic and Hydromagnetic Stability, Dover Publications, New York
Chandrasekhar S., 1969, Ellipsoidal Figures of Equilibrium, Dover Publications, New York
Chandrasekhar S., 1989a, Selected Papers, Volume 3: Stochastic, Statistical, and Hydromagnetic Problems in Physics and Astronomy, University of Chicago Press, Chicago
Chandrasekhar S., 1989b, Selected Papers, Volume 4: Plasma Physics, Hydrodynamic and Hydromagnetic Stability, and Applications of the Tensor-Virial Theorem, University of Chicago Press, Chicago
Fricke K., 1969, A&A, 1, 388
Goldreich P., Sridhar S., 1995, ApJ, 438, 763
Hawley J.F., Balbus S.A., Stone J.M., 2001, ApJ, 554, L49
Lemaster M.N., Stone J.M., 2009. ApJ, 691, 1092
MacLow M.-M., 1999, ApJ, 524, 169
Miller K.A., Stone J.M., 1999, ASSL, 240, 237
Parker E.N., 1955, ApJ, 122, 293
Parrish I.J., Stone J.M., 2007, ApJ, 664, 135
Parrish I.J., Quataert E., Sharma P., 2009, ApJ, 703, 96
Quataert E., 2008, ApJ, 673, 758
Rempel M., Schüssler M., Cameron R.H., Knölker M., 2009, Science, 325, 171
Safronov V.S., 1972, Evolution of the protoplanetary cloud and formation of the earth and planets, Jerusalem (Israel): Israel Program for Scientific Translations, Keter Publishing House
Shakura N.I., Sunyaev R.A., 1973, A&A, 24, 337
Stone J.M., Ostriker E.C., Gammie C.F., 1998, ApJ, 508, L99
Stone J.M., Gardiner T.A., Teuben P., Hawley J.F., Simon J.B., 2008, ApJS, 178, 137
Sturrock P.A., 1994, Plasma Physics, Cambridge University Press, Cambridge
Velikhov E.P., 1959, Sov. Phys. JETP, 36, 995

Bull. Astr. Soc. India (2011)

The formation and evolution of massive black hole seeds in the early Universe

Priyamvada Natarajan*

Department of Astronomy, Yale University, 260 Whitney Avenue, New Haven, CT 06511, USA
Department of Physics, Yale University, P.O. Box New Haven, CT 06520, USA
Institute for Theory and Computation, Harvard University,
60 Garden Street, Cambridge MA 02138, USA

Received February 27, 2011; accepted March 24, 2011

Abstract. Tracking the evolution of high redshift seed black hole masses to late times, we examine the observable signatures today. These massive initial black hole seeds form at extremely high redshifts from the direct collapse of pre-galactic gas discs. Populating dark matter halos with seeds formed in this fashion, we follow the mass assembly history of these black holes to the present time using a Monte-Carlo merger tree approach. Utilizing this formalism, we predict the black hole mass function at high redshifts and at the present time; the integrated mass density of black holes in the Universe; the luminosity function of accreting black holes as a function of redshift and the scatter in observed, local $M_{\rm bh} - \sigma$ relation. Comparing the predictions of the 'light' seed model with these massive seeds we find that significant differences appear predominantly at the low mass end of the present day black hole mass function. However, all our models predict that low surface brightness, bulge-less galaxies with large discs are least likely to be sites for the formation of massive seed black holes at high redshifts. The efficiency of seed formation at high redshifts has a direct influence on the black hole occupation fraction in galaxies at $z = 0$. This effect is more pronounced for low mass galaxies. This is the key discriminant between the models studied here and the Population III remnant 'light' seed model. We find that there exists a population of low mass galaxies that do not host nuclear black holes. Our prediction of the shape of the $M_{\rm bh} - \sigma$ relation at the low mass end and increased scatter has recently been corroborated by observations.

Keywords : black holes – galaxies: evolution – galaxies: high redshift

*e-mail: priyamvada.natarajan@yale.edu

1. Introduction

Demography of local galaxies suggests that most galaxies harbour quiescent super-massive black holes (SMBHs) in their nuclei at the present time and that the mass of the hosted SMBH is correlated with properties of the host bulge. In fact, observational evidence points to the existence of a strong correlation between the mass of the central SMBH and the velocity dispersion of the host spheroid (Tremaine *et al.* 2002; Ferrarese & Merritt 2000, Gebhardt *et al.* 2003; Marconi & Hunt 2003; Häring & Rix 2004; Gültekin *et al.* 2009) and possibly the host halo (Ferrarese 2002) in nearby galaxies. These correlations are strongly suggestive of co-eval growth of the SMBH and the stellar component, likely via regulation of the gas supply in galactic nuclei from the earliest times (Haehnelt, Natarajan, Rees 1998; Silk & Rees 1999; Kauffmann & Haehnelt 2000; Fabian 2002; King 2003; Thompson, Quataert & Murray 2005; Natarajan & Treister 2009).

2. Links between massive SMBH seeds, halo mass and spin

Optically bright quasars powered by accretion onto black holes are now detected out to redshifts of $z > 6$ when the Universe was barely 7% of its current age (Fan *et al.* 2004; 2006). The luminosities of these high redshift quasars imply black hole masses $M_{\rm BH} > 10^9\,M_\odot$. Models that describe the growth and accretion history of supermassive black holes typically use as initial seeds the remnants derived from Pop-III stars (e.g. Haiman & Loeb 1998; Haehnelt, Natarajan & Rees 1998). Assembling these large black hole masses by this early epoch starting from remnants of the first generation of metal free stars has been a challenge for models. Some suggestions to accomplish rapid growth invoke super-Eddington accretion rates for brief periods of time (Volonteri & Rees 2005). Alternatively, it has been suggested that the formation of more massive seeds ab-initio through direct collapse of self-gravitating pre-galactic disks might offer a new channel as proposed by Lodato & Natarajan 2006 [LN06]. This scenario alleviates the problem of building up supermassive black hole masses to the required values by $z = 6$.

We focus on the main features of massive seed models in this review. Most aspects of the evolution and assembly history of this scenario have been explored in detail in Volonteri & Natarajan (2009) and Volonteri, Lodato & Natarajan (2008). In these models, at early times the properties of the assembling SMBH seeds are more tightly coupled to properties of the dark matter halo as their growth is driven by the merger history of halos. However, at later times, when the merger rates are low, the final mass of the SMBH is likely to be more tightly coupled to the small scale local baryonic distribution. The relevant host dark matter halo property at high redshifts in this picture is the spin.

In a physically motivated model for the formation of heavy SMBH seeds (in contrast to the lower mass remnant seeds from Population III stars) as described in LN06, there is a limited range of halo spins and halo masses that are viable sites for the formation of seeds. In this picture, massive seeds with $M \approx 10^5 - 10^6 M_\odot$ can form at high redshift ($z > 15$), when the intergalactic medium has not been significantly enriched by metals (Koushiappas, Bullock & Dekel 2004; Begelman, Volonteri & Rees 2006; LN06; Lodato & Natarajan 2007). As derived in LN06, the development of non-axisymmetric spiral structures drives mass infall and accumulation in a pre-galactic disc with primordial composition. The mass accumulated in the center of the halo (which provides an upper limit to the SMBH seed mass) is given by:

$$M_{\rm BH} = m_{\rm d} M_{\rm halo} \left[1 - \sqrt{\frac{8\lambda}{m_{\rm d} Q_{\rm c}} \left(\frac{j_{\rm d}}{m_{\rm d}} \right) \left(\frac{T_{\rm gas}}{T_{\rm vir}} \right)^{1/2}} \right] \tag{1}$$

for

$$\lambda < \lambda_{\rm max} = m_{\rm d} Q_{\rm c} / 8 (m_{\rm d}/j_{\rm d}) (T_{\rm vir}/T_{\rm gas})^{1/2} \tag{2}$$

and $M_{\rm BH} = 0$ otherwise. Here $\lambda_{\rm max}$ is the maximum halo spin parameter for which the disc is gravitationally unstable, m_d is the gas fraction that participates in the infall and $Q_{\rm c}$ is the Toomre parameter. The efficiency of SMBH formation is strongly dependent on the Toomre parameter $Q_{\rm c}$, which sets the frequency of formation, and consequently the number density of SMBH seeds. The efficiency of the seed assembly process ceases at large halo masses, where the disc undergoes fragmentation instead. This occurs when the virial temperature exceeds a critical value $T_{\rm max}$, given by:

$$\frac{T_{\rm max}}{T_{\rm gas}} = \left(\frac{4\alpha_{\rm c}}{m_{\rm d}} \frac{1}{1 + M_{\rm BH}/m_{\rm d} M_{\rm halo}} \right)^{2/3}, \tag{3}$$

where $\alpha_{\rm c} \approx 0.06$ is a dimensionless parameter measuring the critical gravitational torque above which the disc fragments. The remaining relevant parameters are assumed to have typical values: $m_{\rm d} = j_{\rm d} = 0.05$, $\alpha_{\rm c} = 0.06$ for the $Q_{\rm c} = 2$ case. The gas has a temperature $T_{\rm gas} = 5000$K.

To summarize, every dark matter halo is characterized by its mass M (or virial temperature $T_{\rm vir}$) and by its spin parameter λ. If $\lambda < \lambda_{\rm max}$ (see equation 2) and $T_{\rm vir} < T_{\rm max}$ (equation 3), then a seed SMBH forms in the centre. Hence SMBHs form (i) only in halos within a given range of virial temperatures, and hence, halo masses, and (ii) only within a narrow range of spin parameters, as shown in Figure 1. High values of the spin parameter, leading most likely to disk-dominated galaxies, are strongly disfavored as seed formation sites in this model, and in models that rely on global dynamical instabilities (Volonteri & Begelman 2010).

3. The evolution of seed black holes

We follow the evolution of the MBH population resulting from the seed formation process delineated above in a ΛCDM Universe. Our approach is similar to the one described in Volonteri, Haardt & Madau (2003). We simulate the merger history of present-day halos with masses in the range $10^{11} < M < 10^{15} M_\odot$ starting from $z = 20$, via a Monte Carlo algorithm based on the extended Press-Schechter formalism. Every halo entering the merger tree is assigned a spin parameter drawn from the lognormal $P(\lambda)$ distribution of simulated LCDM halos. Recent work on the fate of halo spins during mergers in cosmological simulations has led to conflicting results: Vitvitska *et al.* (2002) suggest that the spin parameter of a halo increases after a major merger, and the angular momentum decreases after a long series of minor mergers; D'Onghia & Navarro (2007) find instead no significant correlation between spin and merger history. Given the unsettled nature of this matter, we simply assume that the spin parameter of a halo is not modified by its merger history.

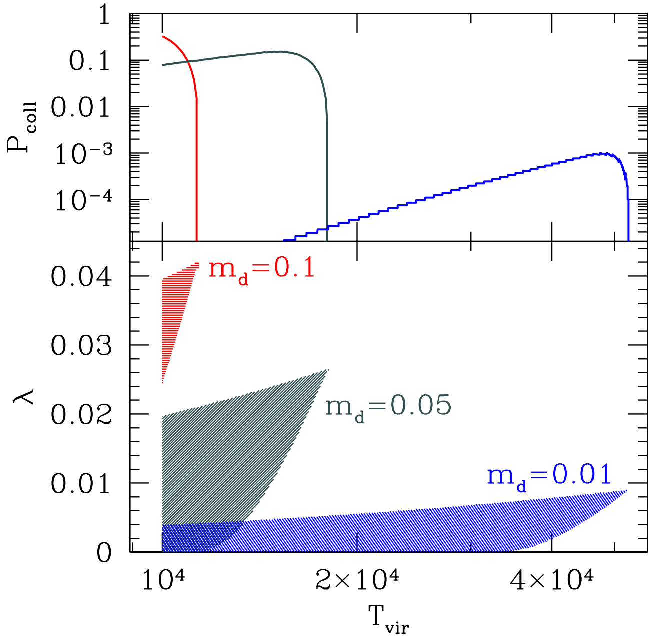


Figure 1. Parameter space (virial temperature, spin parameter) for SMBH formation. Halos with $T_{vir} > 10^4$ K at $z = 15$ are picked to participate in the infall (m_d). The shaded areas in the bottom panel show the range of virial temperatures and spin parameters where discs are Toomre unstable and the joint conditions, $\lambda < \lambda_{\rm max}$ (equation 2) and $T_{\rm vir} < T_{\rm max}$ (equation 3, showing the minimum spin parameter, $\lambda_{\rm min}$ value below which the disc is globally prone to fragmentation) are fulfilled. The top panel shows the probability of SMBH formation and is obtained by integrating the lognormal distribution of spin parameters between $\lambda_{\rm min}$ and $\lambda_{\rm max}$.

When a halo enters the merger tree we assign seed MBHs by determining if the halo meets all the requirements described in Section 2 for the formation of a central mass concentration. As we do not self-consistently trace the metal enrichment of the intergalactic medium, we consider here a sharp transition threshold, and assume that the MBH formation scenario suggested by Lodato & Natarajan ceases at $z \approx 15$ (see also Sesana 2007; Volonteri 2007). At $z > 15$, therefore, whenever a new halo appears in the merger tree (because its mass is larger than the mass resolution), or a pre-existing halo modifies its mass by a merger, we evaluate if the gaseous component meets the conditions for efficient transport of angular momentum to create a large inflow of gas which can either form a MBH seed, or feed one if already present.

The efficiency of MBH formation is strongly dependent on a critical value of the Toomre parameter $Q_{\rm c}$, which sets the frequency of formation, and consequently the number density of MBH seeds. We investigate the influence of this parameter in the determination of the global evolution of the MBH population. Figure 2 shows the number density of seeds formed in three different models with varying efficiency, with $Q_{\rm c} = 1.5$ (low efficiency model A), $Q_{\rm c} = 2$ (intermediate efficiency model B), and $Q_{\rm c} = 3$ (high efficiency

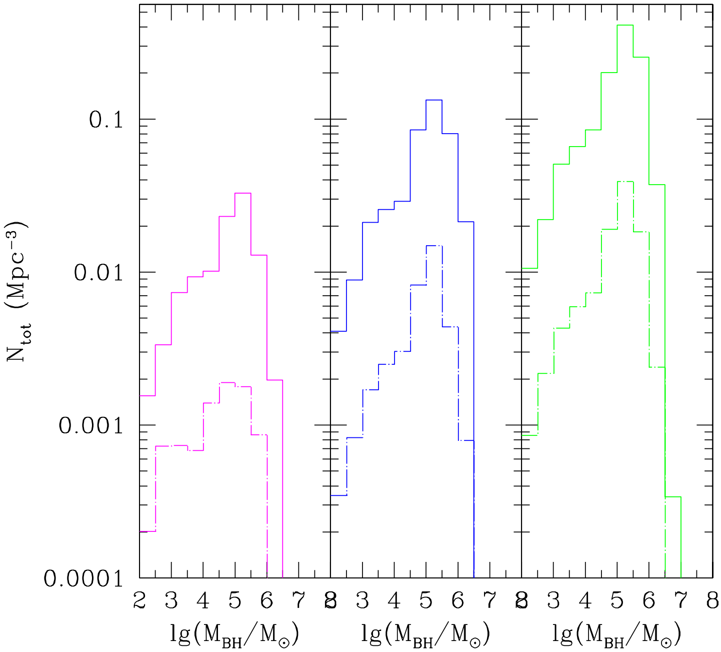


Figure 2. Mass function of MBH seeds in the three Q-models that differ in seed formation efficiency. Left panel: $Q_c = 1.5$ (the least efficient model A), middle panel: $Q_c = 2$ (intermediate efficiency model B), right panel: $Q_c = 3$ (highly efficient model C). Seeds form at $z > 15$ and this channel ceases at $z = 15$. The solid histograms show the total mass function of seeds formed by $z = 15$, while the dashed histograms refer to seeds formed at a specific redshift, $z = 18$.

model C). The solid histograms show the total mass function of seeds formed by $z = 15$ when this formation channel ceases, while the dashed histograms refer to seeds formed in a specific redshift slice at $z = 18$. The number of seeds changes by about one order of magnitude from the least efficient to the most efficient model, consistent with the probabilities shown in Figure 1.

We assume that, after seed formation ceases, the $z < 15$ population of MBHs evolves according to a "merger driven scenario", as described in Volonteri (2006). We assume that during major mergers MBHs accrete gas mass that scales with the fifth power of the circular velocity (or equivalently the velocity dispersion σ_c) of the host halo (Ferrarese 2002). We thus set the final mass of the MBH at the end of the accretion episode to 90% of the mass predicted by the $M_{\rm BH} - \sigma_c$ correlation, assuming that the scaling does not evolve with redshift. Major mergers are defined as mergers between two dark matter halos with mass ratio between 1 and 10. BH mergers contribute to the mass addition of the remaining 10%.

We briefly outline the merger scenario calculation here. The merger rate of halos can be estimated using equation 1 of Fakhouri, Ma & Boylan-Kolchin (2010), where a simple fitting formula is derived from large LCDM simulations. The merger rate per unit redshift and mass ratio (ξ) at fixed halo mass is given by:

$$\frac{dN_m}{d\xi dz}(M_h) = A\left(\frac{M_h}{10^{12}M_0}\right)^{\alpha}\xi^{\beta}\exp\left[\left(\frac{\xi}{\tilde{\xi}}\right)^{\gamma}\right](1+z)^{\eta}. \tag{4}$$

with A = 0.0104, $\alpha = 0.133, \beta = -1.995, \gamma = 0.263, \eta = 0.0993$ and $\tilde{\xi} = 9.72 \times 10^{-3}$. We can integrate the merger rate between $z = 0$ and say, $z = 3$, for major mergers. This gives the number of major mergers a halo of a given mass experiences between $z = 0$ and $z = 3$. Halo mass can be translated into virial circular velocity:

$$V_{\rm c} = 142{\rm km/s}\left[\frac{M_{\rm h}}{10^{12}\,M_{\odot}}\right]^{1/3}\left[\frac{\Omega_m}{\Omega_m^z}\,\frac{\Delta_{\rm c}}{18\pi^2}\right]^{1/6}(1+z)^{1/2}, \tag{5}$$

where $\Delta_{\rm c}$ is the over-density at virialization relative to the critical density. For a WMAP5 cosmology we adopt here the fitting formula $\Delta_{\rm c} = 18\pi^2 + 82d - 39d^2$ (Bryan & Norman 1998), where $d \equiv \Omega_m^z - 1$ is evaluated at the collapse redshift, so that $\Omega_m^z = \Omega_m(1+z)^3/(\Omega_m(1+z)^3 + \Omega_\Lambda + \Omega_k(1+z)^2)$. It is well known that the major merger rate is an increasing function of halo mass or circular velocity. In fact we find that the expected number of mergers between $z = 0$ and $z = 3$ with mass ratio $\xi > 0.3$ is $\simeq 0.4$ for $M_h = 10^8 M_\odot$, $\simeq 0.5$ for $M_h = 10^9 M_\odot$, $\simeq 0.7$ for $M_h = 10^{10} M_\odot$, $\simeq 1.0$ for $M_h = 10^{11} M_\odot$, $\simeq 1.4$ for $M_h = 10^{12} M_\odot$, $\simeq 1.8$ for $M_h = 10^{13} M_\odot$.

In order to calculate the luminosity function of active black holes and to follow the black hole mass growth during each accretion event, we also need to calculate the mass inflow rate. This is assumed to scale with the Eddington rate for the MBH, and is based on the results of merger simulations, which heuristically track accretion onto a central MBH (Di Matteo, Springel & Hernquist 2005; Hopkins *et al.* 2005; Sijacki *et al.* 2007). The time spent by a given simulated AGN at a given bolometric luminosity[1] per logarithmic interval is approximated by Hopkins *et al.* (2005) as:

$$\frac{dt}{dL} = |\alpha| t_Q\, L^{-1}\left(\frac{L}{10^9 L_\odot}\right)^{\alpha}, \tag{6}$$

where $t_Q \simeq 10^9$ yr, and $\alpha = -0.95 + 0.32\log(L_{\rm peak}/10^{12}L_\odot)$. Here $L_{\rm peak}$ is the luminosity of the AGN at the peak of its activity. Hopkins *et al.* (2006) show that approximating $L_{\rm peak}$ by the Eddington luminosity of the MBH at its final mass (i.e., when it sits on the $M_{\rm BH} - \sigma_c$ relation) compared to computing the peak luminosity with equation (6) above gives the same result and in fact, the difference between these two cases is negligible. Volonteri, Salvaterra & Haardt (2006) derive the following simple differential equation to express the instantaneous accretion rate ($f_{\rm Edd}$,in units of the Eddington rate) for a MBH of mass $M_{\rm BH}$

[1] We convert accretion rate into luminosity assuming that the radiative efficiency equals the binding energy per unit mass of a particle in the last stable circular orbit. We associate the location of the last stable circular orbit with the spin of the MBHs, by self-consistently tracking the evolution of black hole spins throughout our calculations (Volonteri 2006). We set 20% as the maximum value of the radiative efficiency, corresponding to a spin slightly below the theoretical limit for thin disc accretion (Thorne 1974).

in a galaxy with velocity dispersion σ_c:

$$\frac{d f_{\rm Edd}(t)}{dt} = \frac{f_{\rm Edd}^{1-\alpha}(t)}{|\alpha| t_Q} \left(\frac{\epsilon \dot{M}_{\rm Edd} c^2}{10^9 L_\odot} \right)^{-\alpha} , \tag{7}$$

where t is the time elapsed from the beginning of the accretion event. Solving this equation provides us with the instantaneous Eddington ratio for a given MBH at a specific time, and therefore we can self-consistently follow the MBH mass. We set the Eddington ratio $f_{\rm Edd} = 10^{-3}$ at $t = 0$. This same type of accretion is assumed to occur, at $z > 15$, following a major merger in which a MBH is not fed by disc instabilities.

4. Results

The repercussions of different initial efficiencies for seed formation for the overall evolution of the MBH population stretch from high-redshift to the local Universe. Detection of gravitational waves from seeds merging at the redshift of formation (Sesana 2007) is probably one of the best ways to discriminate among formation mechanisms. On the other hand, the imprint of different formation scenarios can also be sought in observations at lower redshifts. The various seed formation scenarios have distinct consequences for the properties of the MBH population at $z = 0$.

4.1 Low redshift predictions

4.1.1 *Supermassive black holes in dwarf galaxies*

Obviously, a higher density of MBH seeds implies a more numerous population of MBHs at later times, which can produce observational signatures in statistical samples. More subtly, the formation of seeds in a ΛCDM scenario follows the cosmological bias. As a consequence, the progenitors of massive galaxies (or clusters of galaxies) have a higher probability of hosting MBH seeds (cf. Madau & Rees 2001). In the case of low-bias systems, such as isolated dwarf galaxies, very few of the high-z progenitors have the deep potential wells needed for gas retention and cooling, a prerequisite for MBH formation. In the lowest efficiency model A, for example, a galaxy needs of order 25 massive progenitors (mass above $\sim 10^7 M_\odot$) to ensure a high probability of seeding within the merger tree. In model C, instead, the requirement drops to 4 massive progenitors, increasing the probability of MBH formation in lower bias halos.

The signature of the efficiency of the formation of MBH seeds will consequently be stronger in isolated dwarf galaxies. Figure 3 (bottom panel) shows a comparison between the observed $M_{\rm BH} - \sigma$ relation and the one predicted by our models (shown with circles), and in particular, from left to right, the three models based on the LN06 and Lodato & Natarajan (2007) seed masses with $Q_{\rm c} = 1.5$, 2 and 3, and a fourth model based on lower-mass Population III star seeds. The upper panel of Figure 3 shows the fraction of galaxies that **do not** host any massive black holes for different velocity dispersion bins. This shows that the fraction of galaxies without a MBH increases with decreasing halo masses at $z = 0$. A larger fraction of low mass halos are devoid of central black holes for lower seed formation efficiencies. Note that this is one of the key discriminants between our models

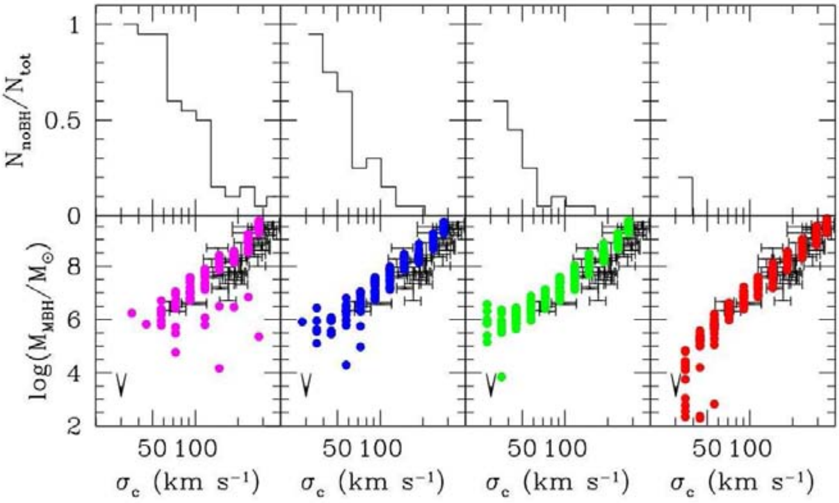


Figure 3. The M_{bh}–velocity dispersion (σ_c) relation at $z = 0$. Every circle represents the central MBH in a halo of given σ_c. Observational data are marked by their quoted errorbars, both in σ_c, and in M_{bh} (Tremaine *et al.* 2002). Left to right panels: $Q_c = 1.5$, $Q_c = 2$, $Q_c = 3$, Population III star seeds. *Top panels:* fraction of galaxies at a given velocity dispersion which **do not** host a central MBH.

and those seeded with Population III remnants. As shown in Figure 3, there are practically no galaxies without central BHs for the Population III seeds.

We can therefore make quantitative predictions for the local occupation fraction of MBHs. Our model A predicts that below $\sigma_c \approx 60$ kms^{-1} the probability of a galaxy hosting a MBH is negligible. With increasing MBH formation efficiencies, the minimum mass for a galaxy that hosts a MBH decreases, and it drops below our simulation limits for model C. On the other hand, models based on lower mass Population III star remnant seeds, predict that massive black holes might be present even in low mass galaxies. Our predictions have been corroborated by recent observations of low mass galaxies (Kormendy & Bender 2011).

Although there are degeneracies in our modeling (e.g., between the minimum redshift for BH formation and the instability criterion), the BH occupation fraction and the masses of the BHs in dwarf galaxies are the key diagnostics. An additional caveat worth mentioning is the possibility that a galaxy is devoid of a central MBH because of dynamical ejections (due to either the gravitational recoil or three-body scattering). The signatures of such dynamical interactions should be more prominent in dwarf galaxies, but ejected MBHs would leave observational signatures on their hosts (Gültekin *et al.* in prep.). On top of that, Schnittman (2007) and Volonteri, Lodato & Natarajan (2008) agree in considering the recoil a minor correction to the overall distribution of the MBH population at low redshift (cf. Figure 4 in Volonteri 2007).

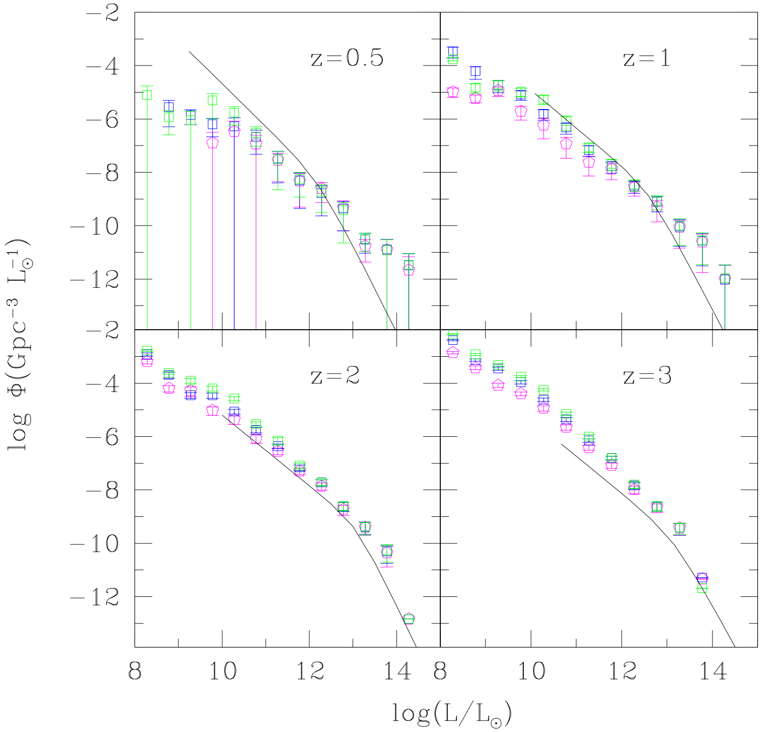


Figure 4. Predicted bolometric luminosity functions at different redshifts with observational data over-plotted. All 3 models match the observed bright end of the LF at high redshifts and predict a steep slope at the faint end down to $z = 1$. The 3 models are not really distinguishable with the LF. However at low redshifts, for instance at $z = 0.5$, all 3 models are significantly flatter at both high and low luminosities and do not adequately match the current data. As discussed in the text, the LF is strongly determined by the accretion prescription, and what we see here is simply a reflection of that fact.

Additionally, as MBH seed formation requires halos with low angular momentum (small spin parameter), we envisage that low surface brightness, bulge-less galaxies with large spin parameters (i.e. large discs) are systems where MBH seed formation is less probable. Furthermore, bulgeless galaxies are believed to have preferentially quieter merger histories and are unlikely to have experienced major mergers that could have brought in a MBH from a companion galaxy.

4.1.2 *Comoving mass density of black holes*

Since during the quasar epoch MBHs increase their mass by a large factor, signatures of the seed formation mechanisms are likely more evident at *earlier epochs*. We compare in Figure 5 the integrated comoving mass density in MBHs to the expectations from Sołtan-type arguments, assuming that quasars are powered by radiatively efficient flows (for details, see

Yu & Tremaine 2002; Elvis, Risaliti & Zamorini 2002; Marconi *et al.* 2004). While during and after the quasar epoch the mass densities in models A, B, and C differ by less than a factor of 2, at $z > 3$ the differences are more pronounced.

A very efficient seed MBH formation scenario can lead to a very large BH density at high redshifts. For instance, in the highest efficiency model C with $Q_c = 3$, the integrated MBH density at $z = 10$ is already $\sim$ 25% of the density at $z = 0$. The plateau at $z > 6$ is due to our choice of scaling the accreted mass with the $z = 0$ $M_{\rm bh} - \sigma$ relation. Since in our models we let MBHs accrete mass that scales with the fifth power of the circular velocity of the halo, the accreted mass is a small fraction of the MBH mass (see the discussion in (Marulli *et al.* 2006)), and the overall growth remains small, as long as the mass of the seed is larger than the accreted mass, which, for our assumed scaling, happens whenever the mass of the halo is below a few times $10^{10} M_\odot$. The comoving mass density, an integral constraint, is reasonably well determined out to $z = 3$ but is poorly known at higher redshifts. All models appear to be satisfactory and consistent with current observational limits (shown as the shaded area).

4.1.3 *Black hole mass function at $z = 0$*

One of the key diagnostics is the comparison of the measured and predicted BH mass function at $z = 0$ for our 3 models. In Figure 6, we show (from left to right, respectively) the mass function predicted by models A, B, C and Population III remnant seeds compared to that obtained from measurements. The histograms show the mass function obtained with our models (where the upper histogram includes all the black holes while the lower one only includes black holes found in central galaxies of halos in the merger-tree approach). The two lines are two different estimates of the observed black hole mass function. In the upper one, the measured velocity dispersion function for nearby late and early-type galaxies from the SDSS survey (Bernardi *et al.* 2003; Sheth *et al.* 2003) has been convolved with the measured $M_{\rm BH} - \sigma$ relation. We note here that the scatter in the $M_{\rm bh} - \sigma$ relation is not explicitly included in this treatment, however the inclusion of the scatter is likely to preferentially affect the high mass end of the BHMF, which provides stronger constraints on the accretion histories than do the seed masses. It has been argued by Tundo *et al.* (2007), Bernardi *et al.* (2007) and Lauer *et al.* (2007) that the BH mass function differs if the bulge mass is used instead of the velocity dispersion in relating the BH mass to the host galaxy. Since our models do not trace the formation and growth of stellar bulges in detail, we are restricted to using the velocity dispersion in our analysis.

The lower dashed curve is an alternate theoretical estimate of the BH mass function derived using the Press-Schechter formalism from Jenkins *et al.* (2001) in conjunction with the observed $M_{\rm BH} - \sigma$ relation. Selecting only the central galaxies of halos in the merger-tree approach adopted here (lower histograms) is shown to be equivalent to this analytical estimate, and this is clearly borne out in the plot. When we include black holes in satellite galaxies (upper histograms, cf. the discussion in Volonteri, Haardt & Madau 2003) the predicted mass function moves towards the estimate based on SDSS galaxies. The higher efficiency models clearly produce more BHs. At higher redshifts, for instance at $z = 6$, the mass functions of active MBHs predicted by all models are in very good agreement, in particular for BH masses larger than $10^6 M_\odot$, as it is the growth by accretion

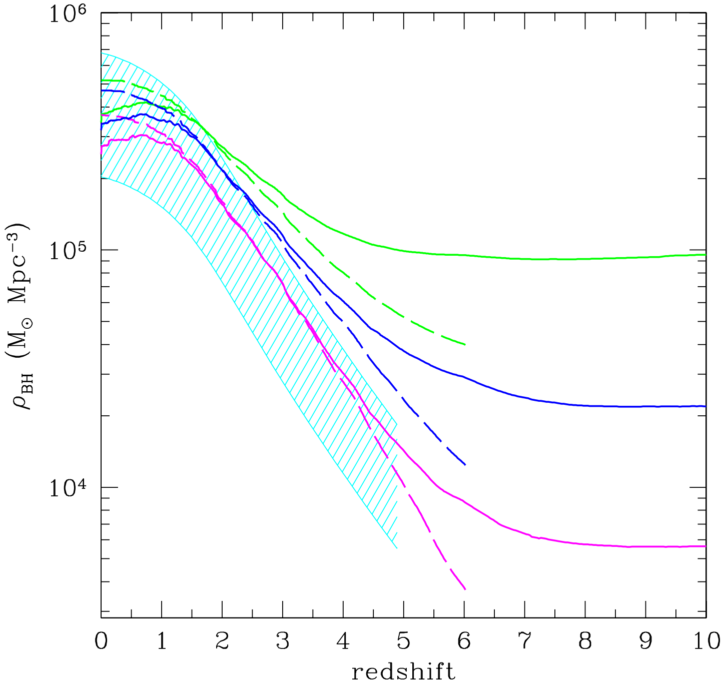


Figure 5. Integrated black hole mass density as a function of redshift. Solid lines: total mass density locked into nuclear black holes. Dashed lines: integrated mass density accreted by black holes. Models based on BH remnants of Population III stars (lowest curve), $Q_c = 1.5$ (middle curve) and $Q_c = 2$ (upper curve). Shaded area: constraints from Sołtan-type arguments, where we have varied the radiative efficiency from a lower limit of 6% (applicable to Schwarzschild MBHs, upper envelope of the shaded area), to about 20%. All 3 massive seed formation models are in comfortable agreement with the mass density obtained from integrating the optical luminosity functions of quasars.

that dominates the evolution of the population. At the highest mass end ($> 10^9\ M_\odot$) model A lags behind models B and C, although we stress once again that our assumptions for the accretion process are very conservative.

The *relative* differences between models A, B, and C at the low-mass end of the mass function, however, are genuinely related to the MBH seeding mechanism (see also Figures 3 and 5). In model A, simply, fewer galaxies host a MBH, hence reducing the overall number density of black holes. Although our simplified treatment does not allow robust quantitative predictions, the presence of a "bump" at $z = 0$ in the MBH mass function at the characteristic mass that marks the peak of the seed mass function (cf. Figure 2) is a sign of highly efficient formation of massive seeds (i.e., much larger mass than, for instance, Population III remnants). The higher the efficiency of seed formation, the more pronounced is the bump (note that the bump is most prominent for model C). Since current measurements of MBH masses extend barely down to $M_{\rm bh} \sim 10^6 M_\odot$, this feature cannot be observationally tested with present data, but future campaigns, with the Giant Magellan

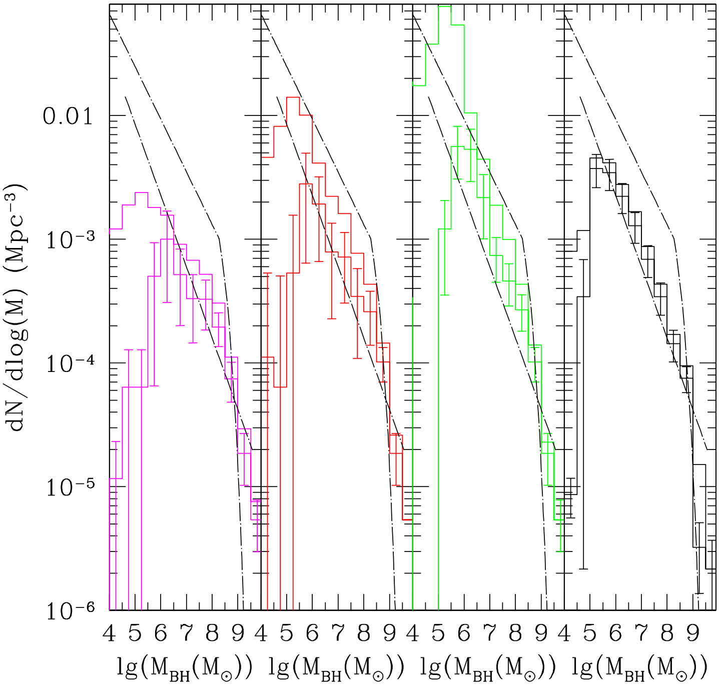


Figure 6. Mass function of black holes at z=0. Histograms represent the results of our models, including central galaxies only (lower histograms with error bars), or including satellites in groups and clusters (upper histograms). Left panel: $Q_c = 1.5$, mid-left panel: $Q_c = 2$, mid-right panel: $Q_c = 3$, right panel: models based on BH remnants of Population III stars. Upper dashed line: mass function derived from combining the velocity dispersion function of Sloan galaxies (Sheth *et al.* 2003, where we have included the late-type galaxies extrapolation), and BH mass-velocity dispersion correlation (e.g., Tremaine et al. 2002). Lower dashed line: mass function derived using the Press-Schechter formalism from Jenkins *et al.* (2001) in conjunction with the $M_{BH} - \sigma$ relation (Ferrarese 2002).

Telescope or JWST, are likely to extend the mass function measurements to much lower black hole masses.

4.2 Predictions at high redshift

4.2.1 *The luminosity function of accreting black holes*

Turning to the global properties of the MBH population, as suggested by Yu & Tremaine (2002), the mass growth of the MBH population at $z < 3$ is dominated by the mass accreted during the bright epoch of quasars, thus washing out most of the imprint of initial conditions. This is evident when we compute the luminosity function of AGN. Clearly the detailed shape of the predicted luminosity function depends most strongly on the accretion prescription used. With our assumption that the gas mass accreted during each merger episode is proportional to V_c^5, we find that distinguishing between the various seed models is difficult. As shown in Figure 4, all 3 models reproduce the bright end of the observed bolometric LF (Hopkins, Richards & Hernquist 2007) at higher redshifts (marked as the solid curve in all the panels), and predict a fairly steep faint end that is as yet undetected. All models fare less well at low redshift, shown in particular at $z = 0.5$. This could be due to the fact that we have used a single accretion prescription to model growth at all times. On the other hand, the decline in the available gas supply at low redshifts (since the bulk of the gas has been consumed before this epoch by star formation activity) likely changes the radiative efficiency of these systems. Besides, observations suggest a sharp decline in the number of actively accreting black holes at low redshifts at different wave-lengths, produced most probably by changes in the accretion flow as a result of changes in the geometry of the nuclear regions of galaxies. In fact, all 3 of our models under-predict the slope at the faint end. There are three other effects that could cause this flattening of the LF at the faint end at low redshift for our models: (i) not having taken into account the result of on-going mergers and the fate of satellite galaxies; (ii) the number of realizations generated and tracked is insufficient for statistics, as evidenced by the systematically larger errorbars and (iii) more importantly, it is unclear if merger-driven accretion is indeed the trigger of BH fueling in the low redshift Universe. We note that the 3 massive seed models and Population III seed model cannot be discriminated by the LF at high redshifts. Models B and C are also in agreement viz-a-viz the predicted BH mass function at $z = 6$ (see Figure 2), even assuming a very high radiative efficiency (up to 20%), while model A might need less severe assumptions, in particular for BH masses larger than $10^7\ M_\odot$.

5. Conclusions

In this review, we outline massive black hole seed formation models and focus on the predictions made by these at high and low redshift. While the errors on mass determinations of local black holes are large at the present time, definite trends with host galaxy properties are observed. The tightest correlation appears to be between the BH mass and the velocity dispersion of the host spheroid. Starting with the ab-initio black hole seed mass function computed in the context of direct formation of central objects from the collapse of pre-galactic discs in high redshift halos, we follow the assembly history to late times using a Monte Carlo merger tree approach. Key to our calculation of the evolution and build-up of mass is the prescription that we adopt for determining the precise mass gain during a merger. Motivated by the phenomenological observation of $M_{\rm BH} \propto V_{\rm c}^5$, we assume that this proportionality carries over to the gas mass accreted in each step. With these prescriptions, a range of predictions can be made for the mass function of black holes at high and low z, and for the integrated mass density of black holes, all of which are observationally determined. We evolve 3 models, designated model A, B and C, which correspond to increasing efficiencies respectively for the formation of seeds at high redshift. These models are compared to one in which the seeds are remnants of Population III stars.

It is important to note here that one major uncertainty prevents us from making more concrete predictions: the unknown metal enrichment history of the Universe. Key to the implementation of our models is the choice of redshift at which massive seed formation is quenched. The direct seed formation channel described here ceases to operate once the

Universe has been enriched by metals that have been synthesized by the first generation of stars. Once metals are available in the Inter-Galactic Medium, gas cooling is much more efficient and hydrogen in either atomic or molecular form is no longer the key player. In this work, we have assumed this transition redshift to be $z = 15$. The efficiency of MBH formation and the transition redshift are somehow degenerate (e.g., a model with $Q = 1.5$ and enrichment redshift $z = 12$ is halfway between model A and model B); if other constraints on this redshift were available we could considerably tighten our predictions.

Below we list our predictions and compare how they fare with respect to current observations. The models investigated here clearly differ in predictions at the low mass end of the black hole mass function. With future observational sensitivity in this domain, these models can be distinguished.

1. Occupation fraction at $z = 0$: Our model for the formation of relatively high-mass black hole seeds in high-z halos has direct influence on the black hole occupation fraction in galaxies at $z = 0$. All our models predict that low surface brightness, bulge-less galaxies with large spin parameters (i.e. large discs) are systems where MBH formation is least probable. We find that a significant fraction of low-mass galaxies might not host a nuclear black hole. This is in very good agreement with the shape of the $M_{\rm bh} - \sigma$ relation determined recently from an observational census (an HST ACS survey) of low mass galaxies in the Virgo cluster reported by Ferrarese *et al.* (2006). While current data in the low mass regime are scant (Barth 2004; Greene & Ho 2007; Kormendy & Bender 2011), future instruments and surveys are likely to probe this region of parameter space with significantly higher sensitivity.
2. High mass end of the local SMBH mass function: While the models studied here (with different black hole seed formation efficiencies) are distinguishable at the low mass end of the BH mass function, at the high mass end the effect of initial seeds appears to be less important. These models cannot be easily distinguished by observations at $z \sim 3$.

One of the key caveats of our picture is that it is unclear whether the differences produced by different seed models on observables at $z = 0$ might be compensated or masked by BH fueling modes at earlier epochs. There could be other channels for BH growth that dominate at low redshifts like minor mergers, dynamical instabilities, accretion of molecular clouds and tidal disruption of stars. The decreased importance of the merger driven scenario is patent from observations of low-redshift AGN, which are for the large majority hosted by undisturbed galaxies (e.g. Pierce *et al.* 2007 and references therein) in low-density environments. However, the feasibility and efficiency of some alternative channels are still to be proven, for example, the efficiency of feeding from large scale instabilities (see discussion in King & Pringle 2007; Shlosman, Frank & Begelman 1989; Goodman 2003; Collin 1999). In any event, while these additional channels for BH *growth* can modify the detailed shape of the mass function of MBHs, or of the luminosity function of quasars, they will not create new MBHs. The occupation fraction of MBHs (see Figure 3) is therefore largely *independent* of the accretion mechanism and a true signature of the formation process.

To date, most theoretical models for the evolution of MBHs in galaxies do not include *how* MBHs form. This work is a first analysis of the observational signatures of massive black hole formation mechanisms in the low redshift Universe, complementary to the investigation by Sesana, Volonteri & Haardt (2007), where the focus was on detection of seeds at the very early times when they form, via gravitational waves emitted during MBH mergers. We focus here on possible dynamical signatures that forming massive black hole seeds carry over to the local Universe. Obviously, the signatures of seed formation mechanisms will be far more clear if considered jointly with the evolution of the spheroids that they host. The mass, and especially the frequency, of the forming MBH seeds is a necessary input when investigating how the feedback from accretion onto MBHs influences the host galaxy, and is generally introduced in numerical models using extremely simplified, *ad hoc* prescriptions (e.g., Springel, Di Matteo & Hernquist 2005; Di Matteo, Springel & Hernquist 2005; Hopkins *et al.* 2006; Croton *et al.* 2005; Cattaneo et al 2006; Bower *et al.* 2006). Adopting more detailed models for black hole seed formation, as outlined here, can in principle strongly affect such results. Incorporating sensible assumptions for the masses and frequency of MBH seeds in models of galaxy formation is necessary if we want to understand the symbiotic growth of MBHs and their hosts.

Acknowledgments

PN would like to acknowledge her collaborators Marta Volonteri and Giuseppe Lodato with whom most of this work was done. She would also like to thank the John Simon Guggenheim Foundation for support from a Guggenheim fellowship and the Institute for Theory and Computation at Harvard University for hosting her.

References

Barth A.J., 2004, IAU Symp., 222, 3
Begelman M.C., Volonteri M., Rees M.J., 2006, MNRAS, 370, 289
Bernardi M., *et al.*, 2003, AJ, 125, 1817
Bernardi M., Sheth R.K., Tundo E., Hyde J.B., 2007, ApJ, 660, 267
Bower R., *et al.*, 2006, MNRAS, 370, 645
Bryan G., Norman M., 1998, ApJ, 495, 80
Cattaneo A., Dekel A., Devriendt J., Guiderdoni B., Blaizot J., 2006, MNRAS, 370, 1651
Collin S., 1999, Phys Rep., 311, 463
Croton D., *et al.*, 2005, MNRAS, 356, 1155
Di Matteo T., Springel V., Hernquist L., 2005, Nature, 433, 604
D'Onghia E., Navarro J., 2007, MNRAS, 380, L58
Elvis M., Risaliti G., Zamorani G., 2002, ApJ, 565, L75
Fabian A.C., 2002, ASPC, 258, 185
Fakhouri O., Ma C.-P., Boylan-Kolchin M., 2010, MNRAS, 406, 2267
Fan X., *et al.*, 2004, AJ, 128, 515
Fan X., *et al.*, 2006, AJ, 132, 117
Ferrarese L., 2002, ApJ, 572, 90
Ferrarese L., Merritt D., 2000, ApJ, 539, L9
Ferrarese L., *et al.*, 2006, ApJS, 164, 334
Gebhardt K., *et al.*, 2003, ApJ, 583, 92
Goodman J., 2003, MNRAS, 337, 937

Greene J., Ho, L., 2007, ApJ, 670, 92
Gültekin K., *et al.*, 2009, ApJ, 698, 198
Haehnelt M., Natarajan P., Rees M. J., 1998, MNRAS, 300, 817
Haiman Z., Loeb A., 1998, ApJ, 503, 505
Häring N., Rix H.-W., 2004, ApJ, 604, L89
Hopkins P.F., Hernquist L., Cox T.J., Di Matteo T., Robertson B., Springel V., 2005, ApJ, 630, 716
Hopkins P.F., Hernquist L., Martini P., Cox T.J., Robertson B., Di Matteo T., Springel V., 2005, ApJ, 625, L71
Hopkins P.F., Richards G.T., Hernquist L., 2007, 654, 731
Jenkins A., Frenk C.S., White S.D.M., Colberg J.M., Cole S., Evrard A.E., Couchman H.M.P., Yoshida N., 2001, MNRAS, 321, 372
Kauffmann G., Haehnelt M., 2000, MNRAS, 311, 576
King A., 2003, ApJ, 596, L27
King A.R., Pringle J.E., 2007, MNRAS, 377, L25
Kormendy J., Bender R., 2011, Nature, 469, 377
Koushiappas S.M., Bullock J.S., Dekel A., 2004, MNRAS, 354, 292
Lauer T., Tremaine S., Richstone D., Faber S. M., 2007, 670, 249
Lodato G., Natarajan P., 2006, MNRAS, 371, 1813 (LN06)
Lodato G., Natarajan P., 2007, MNRAS, 377, L64
Madau P., Rees M. J., 2001, ApJ, 551, L27
Marconi A., Hunt L., 2003, ApJ, 598, L21
Marconi A., Risaliti G., GIlli R., Hunt L.K., Maiolino R., Salvati M., 2004, MNRAS, 351, 169
Marulli F., Crociani D., Volonteri M., Branchini E., Moscardini L., 2006, MNRAS, 368, 1269
Natarajan P., Treister E., 2009, MNRAS, 393, 838
Pierce C.M., *et al.*, 2007, ApJ, 660, L19
Schnittmann J., 2007, ApJ, 667, L133
Sesana A., 2007, MNRAS, 382, L6
Sesana A., Volonteri M., Haardt, F., 2007, MNRAS, 377, 1711
Sheth R., *et al.*, 2003, ApJ, 594, 225
Shlosman I., Frank J., Begelman M.C., 1989, Nature, 338, 45
Sijacki D., Springel V., Di Matteo T., Hernquist, L., 2007, MNRAS, 380, 877
Silk J., Rees M. J., 1998, A&A, 331, L1
Springel V., Di Matteo T., Hernquist, L., 2005, ApJ, 620, L79
Thompson T.A., Quataert E., Murray N., 2005, ApJ, 630, 167
Thorne K., 1974, ApJ, 191, 507
Tremaine S., *et al.*, 2002, ApJ, 574, 740
Tundo E., Bernardi M., Hyde J., Sheth R., Pizzella A., 2007, ApJ, 663, 53
Vivitska M., Klypin A., Kravtsov A., Wechsler R., Primack J., Bullock J., 2002, ApJ 581, 799
Volonteri M., 2006, AIPC, 873, 61
Volonteri M., 2007, ApJ, 663, L5
Volonteri M., Begelman M. C., 2010, MNRAS, 409, 1022
Volonteri M., Natarajan P., 2009, MNRAS, 400, 1911
Volonteri M., Rees M.J., 2005, ApJ, 633, 624
Volonteri M., Gültekin K., Dotti M., 2010, MNRAS, 404, 2143
Volonteri M., Haardt F., Madau P., 2003, ApJ, 582, 559
Volonteri M., Lodato G., Natarajan P., 2008, MNRAS, 383, 1079
Volonteri M., Salvaterra R., Haardt F., 2006, MNRAS, 373, 121
Yu Q., Tremaine S., 2002, MNRAS, 335, 965

Bull. Astr. Soc. India (2011)

Early Universe with CMB polarization

Tarun Souradeep[*]
IUCAA, Post Bag 4, Ganeshkhind, Pune, India

Received April 14, 2011; accepted April 15, 2011

Abstract. The Universe is the grandest conceivable scale on which the human mind can strive to understand nature. The amazing aspect of cosmology, the branch of science that attempts to understand the origin and evolution of the Universe, is that it is largely comprehensible by applying the same basic laws of physics that we use for other branches of physics. The observed cosmic microwave background (CMB) is understood by applying the basic laws of radiative processes and transfer, masterfully covered in the classic text by S. Chandrasekhar, in the cosmological context. In addition to the now widely acclaimed temperature anisotropy, there is also linear polarization information imprinted on the observed Cosmic Microwave background. CMB polarization already has addressed, and promises to do a lot more to unravel the deepest fundamental queries about physics operating close to the origin of the Universe.

Keywords : cosmic microwave background – early Universe – polarization – radiative transfer

1. Introduction

It is an honour to write an invited article commemorating the birth centenary of Nobel laureate, Professor Subrahmanyan Chandrasekhar. The Universe is the grandest conceivable scale on which the human mind can strive to understand nature. Remarkably, even the origin and evolution of the Universe is largely comprehensible by applying the same basic laws of physics that are used in many other branches of physics. Chandrasekhar's research epitomizes this amazing reality, that one can understand complex phenomena in astrophysics by building theories based on the basic laws of physics. This article is devoted to the cosmic microwave background (CMB), in particular, the measured intensity and polarization fluctuations. The physics of this emerging champion among cosmological observables is based on straightforward application of the theory of radiative transfer of the relic radiation from big bang through the cosmic eons – a subject that has been masterfully enshrined in the classic text 'Radiative Transfer' of S. Chandrasekhar (1960). This text is, in fact, cited in the seminal papers on CMB anisotropy and polarization and, subsequent reviews (Peebles & Yu 1970; Bond & Efstathiou 1984, 1987; Bond 1996).

[*] e-mail: tarun@iucaa.ernet.in

Historically, theoretical development always preceded observations in cosmology up until the past couple of decades. However, in sharp contrast, recent developments in cosmology have been largely driven by huge improvements in the quality, quantity and scope of cosmological observations. There are two distinct aspects to modern day cosmology – the background Universe and the perturbed Universe. The 'standard' model of cosmology must not only explain the dynamics of the homogeneous background Universe, but also satisfactorily describe the perturbed Universe – the generation, evolution and finally, the formation of the large-scale structure (LSS) in the Universe observed in the vast galaxy surveys. It is fair to say that cosmology over the past few decades has increasingly seen intense interplay between the theory and observations of the perturbed Universe. Spectacular breakthroughs in various observations have now concretely verified that the present edifice of the standard cosmological models is robust. A set of foundations and pillars of cosmology has emerged, and each is supported by a number of distinct observations, which are listed below.

- Homogeneous, isotropic Universe, expanding from a hot initial phase due to gravitational dynamics described by the Friedman equations derived from laws of General Relativity.
- The basic constituents of the Universe are baryons, photons, neutrinos, dark matter and dark energy (cosmological constant/vacuum energy).
- The homogeneous spatial sections of space-time are nearly geometrically flat (Euclidean space).
- Evolution of density perturbations under gravitational instability has produced the large-scale structure in the distribution of matter starting from the primordial perturbations in the early Universe.
- It has been established that the primordial perturbations have correlation on length scales larger than the causal horizon; this makes a strong case for an epoch of inflation in the very early Universe. The nature of primordial perturbations matches that expected from the generation of primordial perturbations in the simplest models of inflation.

The cosmic microwave background, a nearly uniform, thermal black-body distribution of photons throughout space, at a temperature of 2.7 degrees Kelvin, accounts for almost the entire radiation energy density in the Universe. Tiny variations of temperature and linear polarization of these black-body photons of the cosmic microwave background arriving from different directions in the sky faithfully encode information about the early Universe. Further these photons have travelled unimpeded across the entire observable Universe making them excellent probes of the Universe on the largest observable scales. The much talked about 'dawn of precision era of cosmology' has been ushered in by the study of the perturbed Universe. Measurements of CMB anisotropy and polarization have been by far the most influential of the cosmological observations driving advances in current cosmology in this direction.

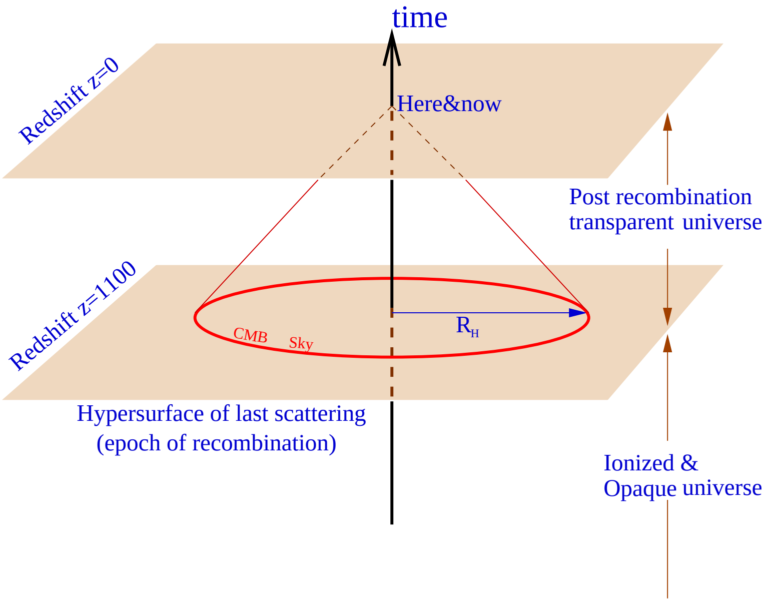


Figure 1. A cartoon explaining the Cosmic Microwave Background (CMB) using a space-(conformal) time diagram. The present Universe is transparent and CMB photons travel to us freely over cosmic distances along our past light cone. In an expanding Universe, the temperature of the Planck black-body CMB is inversely proportional to the expansion factor. When the Universe is about 1100 times smaller, the CMB photons are just hot enough to keep the baryonic matter in the Universe (about 3 quarters Hydrogen, 1 quarter Helium as determined by big bang nucleosynthesis) ionized, and at that epoch there is a sharp transition to an opaque Universe in the past. The CMB photons come to us unimpeded directly from this spherical opaque surface of last scattering at a distance of $R_H = 14$ Gpc that surrounds us – a super IMAX cosmic screen. The red circle depicts the sphere of last scattering in the reduced 2 + 1 dimensional representation of the Universe.

2. CMB anisotropy and polarization

The CMB photons arriving from different directions in the sky show tiny variations in temperature, at a level of ten parts per million, i.e., tens of micro-Kelvin, referred to as the CMB anisotropy, and a net linear polarization pattern at micro-Kelvin to tens of nano-Kelvin level. The tiny variations of temperature and linear polarization of these black-body photons of the cosmic microwave background arriving from different directions in the sky have freely propagated over cosmological distances and carry information about the early Universe. As illustrated in the cartoon in Fig. 1, the cosmic microwave background radiation sky is essentially a *giant, cosmic 'super' IMAX theater screen* surrounding us at a distance of 14 billion parsecs displaying a snapshot of the Universe at a time very close to its origin. Hence the CMB anisotropy and polarization are imprints of the perturbed Universe in the radiation when the Universe was only 0.3 million years old, compared to its present age of about 14 billion years.

It is convenient to express the sky map of CMB temperature anisotropy, $\Delta T(\hat{\mathbf{n}})$ (and polarization, as we shall discuss later), in the direction $\hat{\mathbf{n}}$ in a spherical harmonic expansion:

$$\Delta T(\hat{\mathbf{n}}) = \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{\mathbf{n}})\,. \tag{1}$$

Theory predicts that the primary CMB anisotropy is a statistically isotropic, Gaussian field (of zero mean), and current observations remain fully consistent with this expectation. The anisotropy can then be characterized solely in terms of an angular power spectrum

$$C_\ell = \frac{1}{(2\ell+1)} \sum_{m=-\ell}^{\ell} |a_{\ell m}|^2\,. \tag{2}$$

The C_ℓ spectra for a wide range of parameters within the 'standard' cosmology share a generic set of features neatly related to basic physics, governing the CMB photon distribution function. On the large angular scales (low multipole, ℓ), the CMB anisotropy directly probes the primordial power spectrum of metric fluctuations (scalar gravitational potential and tensor gravitational waves) on scales enormously larger than the 'causal horizon'. On smaller angular scales ($150 < \ell < 1500$), the CMB temperature fluctuations probe the physics of the coupled baryon-photon fluid through the imprint of the acoustic oscillations in the ionized plasma produced by the same primordial fluctuations. At even higher multipoles, the damping tail of the oscillations encodes interesting physics such as the slippage in the baryon-photon coupling, temporal width of the opaque to transparent Universe transition, and weak lensing due to large scale structures in the Universe. Figure 2, which dissects the CMB angular power spectrum, attempts to provide a compact summary of the various kinds of physics involved. Overall, the physics of CMB anisotropy has been very well understood for more than two decades, Furthermore, the predictions of the primary anisotropy and linear polarization and their connection to observables are, by and large, unambiguous (Bond 1996; Hu & Dodelson 2002).

The acoustic peaks occur because the cosmological perturbations excite acoustic waves in the relativistic plasma in the early Universe. The recombination of baryons at redshift $z \approx 1100$ effectively decouples the baryons and photons in the plasma, abruptly switching off the wave propagation. In the time between the excitation of the perturbations and the epoch of recombination, modes of different wavelength can complete different numbers of oscillation periods, or in other words, waves can travel a finite distance and then freeze. This translates the characteristic time scale into a characteristic length scale and leads to a harmonic series of maxima and minima in the CMB anisotropy power spectrum. The acoustic oscillations have a characteristic scale known as the sound horizon, which is the comoving distance that a sound wave could have traveled up to the epoch of recombination. This well-determined physical scale of 150 Mpc is imprinted on the CMB fluctuations at the surface of last scattering. It is the typical scale of the random bright and dull patches on the 'cosmic super-IMAX' screen.

The angle subtended by this physical scale in the CMB sky (IMAX screen) at a known distance of 14 Gpc then allows a sensitive determination of the geometry (Ω_{0K}) of the background Universe. Essentially, the same standard ruler of 150 Mpc placed at 14 Gpc would

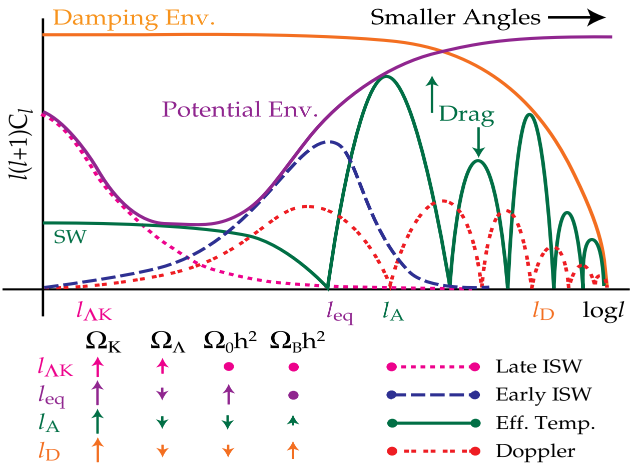


Figure 2. Figure taken from Hu, Sugiyama & Silk (1997) summarizes the different contributions to the primary CMB anisotropy. Also indicated is the dependence of the four length scales that are imprinted on the C_ℓ spectrum by some of the cosmological parameters. The Sachs-Wolfe (SW) plateau at low ℓ is a faithful reproduction of the near scale-invariant spectrum of initial metric perturbations. Integrated Sachs-Wolfe (ISW) effect arises from the evolution of metric perturbations along the path of free streaming CMB photons. Late ISW arises at $\ell < \ell_{\Lambda K}$ if the Universe has significant curvature or cosmological constant. The early ISW contribution at $\ell \sim \ell_{\rm eq}$ is due to the transition from radiation to matter domination. The acoustic and Doppler terms give rise to a harmonic series of oscillatory peaks as a snapshot of the oscillations of a viscous baryon-photon fluid prior to the epoch of recombination. The sound horizon at recombination sets the length scale of the acoustic oscillations. This 'standard ruler' at $z \approx 1100$ then allows an accurate determination of the geometry of the Universe from the location of the first peak, ℓ_A via the angle-distance relationship. High baryon density increases viscous drag leading to suppression of even numbered acoustic peaks relative to odd. Power is exponentially damped at large ℓ due to photon diffusion out of matter over-densities (Silk damping) and finite thickness of the last scattering surface.

subtend different angles in a Universe with different spatial curvature. This determines the location of the series of harmonic peaks of C_ℓ along the multipole ℓ seen in Fig. 2. The amplitude of baryon-photon oscillations can be expected to directly scale with the density of baryons available in the Universe. Consequently, the height of the peaks in the C_ℓ sensitively determine the baryon density, Ω_B. The C_ℓs are sensitive to other important cosmological parameters, such as the relative density of matter, Ω_m, cosmological constant, Ω_Λ, Hubble constant, H_0 and deviation from flatness (curvature), Ω_K. Implicit in C_ℓ is the hypothesized nature of random primordial/initial metric perturbations – (Gaussian) statistics , (nearly scale-invariant) power spectrum, (largely) adiabatic vs. iso-curvature and (largely) scalar vs. tensor component. The 'default' settings in bracket are motivated by inflation (Starobinsky 1982; Guth & Pi 1982; Bardeen, Steinhardt & Turner 1983).

Besides, the entirely theoretical motivation for the paradigm of inflation, the assumption of Gaussian, random, adiabatic scalar perturbations with a nearly scale-invariant power spectrum is arguably also the simplest possible theoretical choice for the initial perturbations. What has been truly remarkable is the extent to which recent cosmological observations have been consistent with and, in certain cases, even vindicated the simplest set of assumptions for the initial conditions for the perturbed Universe discussed below.

The first two decades (~ 1991-2011) of exciting CMB anisotropy measurements have been capped off with the release of 7 years of data from the Wilkinson Microwave Anisotropy Probe (WMAP) of NASA.[1] The first detection of CMB anisotropy by COBE-DMR in 1992 observationally established the origin and mechanism of structure formation in the Universe. Observations were then made at three frequencies, 90, 53 and 31 GHz which allowed a fairly good removal of the 'foreground' contamination of the cosmic signal by the strong emission from our own Galaxy. The 15-years old experimental success story of CMB anisotropy measurements, starting from discovery of CMB anisotropy by the COBE satellite in 1992, has been topped off by the exquisite data from the WMAP. The WMAP satellite was placed at the second Lagrange point of the Sun-Earth system. Measurements from WMAP combine high angular resolution with full sky coverage and high sensitivity due to the stable thermal environment allowed by a space mission. Moreover, observations were made at five frequencies, 94 (W-band), 61 (V-band), 41 (Q-band), 33 (Ka-band) and 23 GHz (K-band) that allowed much better removal of the 'foreground' contamination. Similar to the observational strategy of COBE-DMR, the satellite measures CMB temperature differences between a pair of points in the sky. Each day the satellite covered 30% of the sky, but covers the full sky in 6 months. This massive redundancy in measurements allows the mission to beat down the detector noise from milli-Kelvin to tens of micro-Kelvin level. The WMAP mission acquired data for about nine years up until August 2010 and made that public at regular intervals after a short proprietary period (first year data were released in 2003, three year data in 2006, five year data in 2008, and seven year data in 2010). A final data release of the entire nine years of data is expected in the coming year.

The measured angular power spectrum of the cosmic microwave background temperature fluctuations, C_ℓ, shown in Fig. 3 has become invaluable for constraining cosmological models. The position and amplitude of the peaks and dips of the C_ℓ are sensitive to important cosmological parameters. The most robust constraints obtained are those on the spatial curvature of the Universe and on baryon density. The observations establish that space on cosmic scales is geometrically flat ($\Omega_K = 0$) to within sub-percent precision. The dominant energy content in the present Universe is a mysterious matter with negative pressure dubbed dark energy or the cosmological constant, which contributes about 73% of the total energy budget ($\Omega_\Lambda = 0.73$), followed by cold non-baryonic dark matter about 23% ($\Omega_m = 0.23$) and, most humbly, ordinary matter (baryons) accounts for only about 4% ($\Omega_B = 0.04$) of the matter budget. The current up to date status of cosmological parameter estimates from joint analysis of CMB anisotropy and large-scale structure (LSS) data is usually found in the parameter estimation paper accompanying the most recent results of a major experiment, such as the recent WMAP release of 7-year data (Komatsu *et al.* 2011; Larson *et al.* 2011).

[1] Wilkinson Microwave Anisotropy Probe mission http://wmap.gsfc.nasa.gov/.

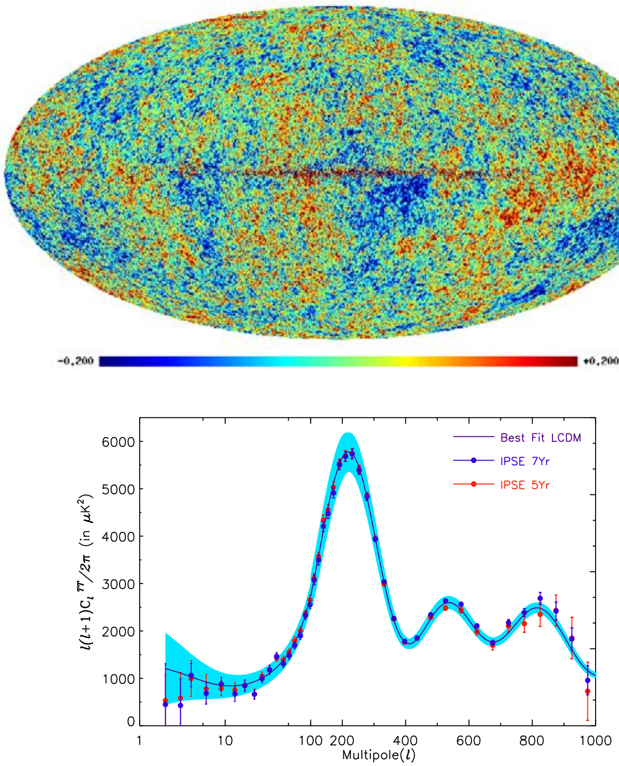


Figure 3. The exquisite temperature anisotropy data from the first three years of data from the WMAP satellite is shown in the figures. *Top:* The top figure shows colour-coded full sky map (in Mollweide projection) of the CMB temperature variations. The temperature variations range between ±200μK with an rms of about 70μK. The angular resolution of features of the map is about a quarter of a degree. For comparison, the first CMB anisotropy measurements in 1992 by the DMR instrument on board the COBE satellite produced the same map at a much coarser resolution of 7 degrees. *Bottom:* The angular power spectrum estimated from the multi-frequency five- and seven-year WMAP data. The result from IPSE, a self-contained model free approach to foreground removal (Saha, Jain & Souradeep 2006; Samal *et al.* 2010) matches that obtained by the WMAP team. The solid curve showing prediction of the best fit power-law, flat, ΛCDM model threads the data points closely [Figure: courtesy Tuhin Ghosh].

More recently, CMB polarization measurements have provided the required complementary information on the nature of initial conditions for the primordial fluctuations. One of the firm predictions of the working ‘standard’ cosmological model is a random pattern of linear polarization (Q and U Stokes parameters) imprinted on the CMB at last scattering surface. Thomson scattering generates CMB polarization anisotropy at decoupling (Bond & Efstathiou 1984; Hu & White 1997). This arises from the polarization dependence of the differential cross section: $d\sigma/d\Omega \propto |\epsilon' \cdot \epsilon|^2$, where ϵ and ϵ' are the incoming and outgoing polarization states involving linear polarization only (Rybicki & Lightman 1979). A local quadrupole temperature anisotropy produces a net polarization, because of the $\cos^2\theta$ dependence of the cross section. A net pattern of linear polarization is retained due to local quadrupole intensity anisotropy of the CMB radiation impinging on the electrons at the last scattering surface. The polarization pattern on the sky can be decomposed into two kinds with different parities. The even parity pattern arises as the gradient of a scalar field called the E-mode. The odd parity pattern arises from the 'curl' of a pseudo-scalar field called the B-mode of polarization. The observed CMB sky map is then characterized by a triplet of random scalar fields: $X(\hat{n}) \equiv \{\Delta T(\hat{n}), E(\hat{n}), B(\hat{n})\}$. It is possible to generalize equation (1) to express both CMB anisotropy and polarization in spherical harmonic space as

$$X(\hat{\mathbf{n}}) = \sum_{\ell=2}^{\infty}\sum_{m=-\ell}^{\ell} a^X_{\ell m} Y_{\ell m}(\hat{\mathbf{n}})\,. \tag{3}$$

and also to define a set of observable angular power spectra analogous to eqn. (2) as

$$C_\ell^{XX'} = \frac{1}{(2\ell+1)}\sum_{m=-\ell}^{\ell} a^X_{\ell m} a^{X'*}_{\ell m}\,. \tag{4}$$

For a statistically isotropic, Gaussian CMB sky, there are a total of 4 power spectra that characterize the CMB signal : $C_\ell^{\rm TT}, C_\ell^{\rm TE}, C_\ell^{\rm EE}, C_\ell^{\rm BB}$. Parity conservation within standard radiative processes eliminates the two other possible power spectra, $C_\ell^{\rm TB}$ & $C_\ell^{\rm EB}$. An important point to note is that the odd-parity B-mode of polarization cannot arise from scalar density perturbations, or potential velocity flow. B mode polarization can arise only due to shear fields acting on photon distribution, such as from gravitational waves and (weak) gravitational lensing deflection of photons.

After the first detection of the CMB polarization spectrum by the Degree Angular Scale Interferometer (DASI) in the intermediate band of angular scales ($l \sim 200 - 440$) in late 2002 (Kovac *et al.* 2002), the field has rapidly grown, with measurements coming in from a host of ground–based and balloon-borne dedicated CMB polarization experiments. The full sky E-mode polarization maps and polarization spectra from WMAP were a new milestone in CMB research (Kogut *et al.* 2003; Page *et al.* 2007). Although the CMB polarization is a clean probe of the early Universe that promises to complement the remarkable successes of CMB anisotropy measurements it is also a much subtler signal than the anisotropy signal. Measurements of polarization by ongoing experiments at sensitivities of μK (E-mode) have had to overcome numerous challenges in the past decade. The tens of nK level B-mode signal pose the ultimate experimental and analysis challenge to this area of observational cosmology. The most current CMB polarization measurement of $C_\ell^{\rm TT}$, $C_\ell^{\rm TE}$ and $C_\ell^{\rm EE}$ and a non-detection of B-modes come from QUaD and BICEP. They also report interesting upper limits on $C_\ell^{\rm TB}$ or $C_\ell^{\rm EB}$, over and above observational artifacts (Wu *et al.* 2009). A non-zero detection of $C_\ell^{\rm TB}$ or $C_\ell^{\rm EB}$, over and above observational artifacts, could be tell-tale signatures of exotic parity violating physics (Lue, Wang & Kamionkowski 1999; Maity, Mazumdar & Sengupta 2004) and the CMB measurements put interesting limits on these possibilities.

The immense dividends of CMB polarization measurements for understanding the physics behind the origin and evolution of our Universe have just started coming in. While

CMB temperature anisotropy can also be generated during the propagation of the radiation from the last scattering surface, the CMB polarization signal must be generated primarily at the last scattering surface, where the optical depth of the Universe transits from large to small values. The polarization information complements the CMB temperature anisotropy by isolating the effect at the last scattering surface from other distinct physical effects acting during the propagation of the photons along the line of sight.

The polarization measurements provide an important test on the adiabatic nature of primordial scalar fluctuations.[2] CMB polarization is produced by the anisotropy of the CMB at recombination, consequently, the angular power spectra of temperature and polarization are closely linked. The power in the CMB polarization signal is produced by the gradient (velocity) term in the same acoustic oscillations of the baryon-photon fluid at last scattering that give rise to temperature (intensity) anisotropy. Hence, clear evidence of adiabatic initial conditions for primordial fluctuations is that the compression and rarefaction peaks in the temperature anisotropy spectrum should be 'out of phase' with the gradient (velocity) driven peaks in the polarization spectra.

Figure 4 taken from Brown *et al.* (2009) reflects the current observational status of CMB E-mode polarization measurements. The recent measurements of the angular power spectrum the E-mode of CMB polarization at large l have confirmed that the peaks in the C_ℓ^{EE} spectra are out of phase with that of the temperature anisotropy spectrum C_ℓ^{TT}.

While the power in the CMB temperature anisotropy at low multipoles ($l \lesssim 60$) first measured by the COBE-DMR (Smoot *et al.* 1992) did point to the existence of correlated cosmological perturbations on super Hubble-radius scales at the epoch of last scattering, it left open the (rather unlikely) 'logical' alternative possibility that all the power at low multipole is generated through the integrated Sachs-Wolfe effect along the line of sight later in the Universe (when the Hubble scale is larger). However, since the polarization anisotropy is generated only at the last scattering surface, the negative trough clearly visible at high significance in the C_ℓ^{TE} spectrum at $l \sim 130$ (which corresponds to a scale larger than the horizon at the epoch of last scattering) sealed this loophole, and provides unambiguous proof of apparently 'acausal' correlations in the cosmological perturbations. This was first first measured by WMAP and later reconfirmed with higher significance by QUaD and BICEP (Kogut *et al.* 2003; Bennett *et al.* 2003; Page *et al.* 2007; Brown *et al.* 2009; Chiang *et al.* 2010).

The B-mode CMB polarization is a very clean and direct probe of the early Universe physics that generated the primordial metric perturbations. Inflationary models necessarily produce tensor perturbations (gravitational waves) that are predicted to evolve independently of the scalar density perturbations, with an uncorrelated power spectrum. The tensor modes on the scales of Hubble-radius along the line of sight to the last scattering distort the photon propagation and generate an additional anisotropy pattern predominantly on the largest scales. (The amplitude of a tensor mode falls off rapidly on sub-Hubble radius scales, hence it is important on angular scales comparable to and larger than the Hubble radius at last scattering). It is common to parametrize the tensor component by the

[2] Another independent observational test comes from the recent measurements of the Baryon Acoustic Oscillations (BAO) in the power spectrum of LSS in the distribution of galaxies. BAO has also observationally established the gravitational instability mechanism for structure formation.

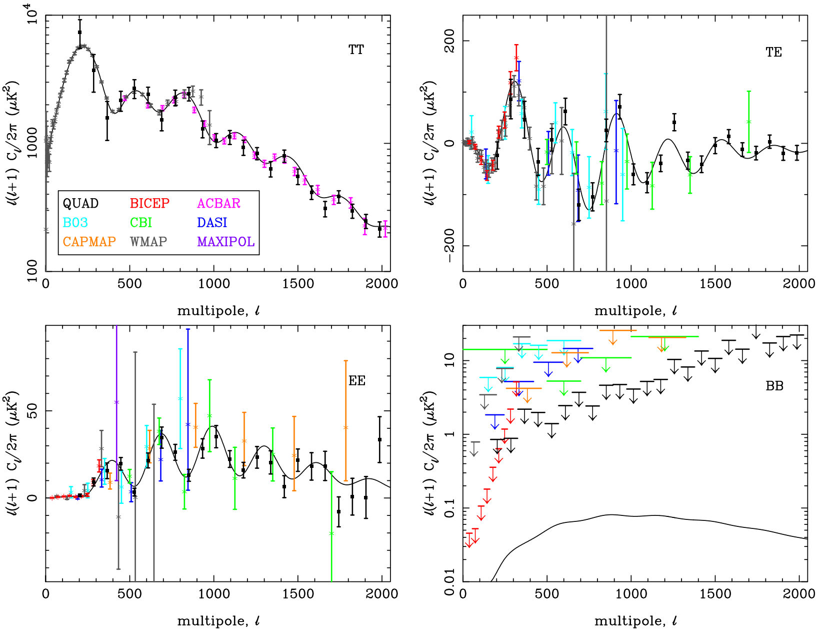


Figure 4. Figure taken from Brown *et al.* (2009) shows a compilation of recent measurements of the angular power spectra of CMB anisotropy and polarization from a number of CMB experiments. The data is good enough to indicate that the peaks in EE and TE are out of phase with that of TT as expected for adiabatic initial conditions. The null BB detection of primary CMB signal from gravity waves is not unexpected (given the ratio of tensor to scalar perturbations expected in the simplest models of inflation). It is important to note that the upper limits on $C_\ell^{\rm BB}$ have improved by almost an order of magnitude in the past two years from experiments such as QUAD and BICEP.

ratio $r_{k_*} = A_{\rm t}/A_{\rm s}$, ratio of $A_{\rm t}$, the primordial power in the transverse traceless part of the metric tensor perturbations, and $A_{\rm s}$, the amplitude scalar perturbation at a comoving wavenumber, k_* (in $\rm Mpc^{-1}$). For power-law models, recent WMAP data alone put an improved upper limit on the tensor to scalar ratio, $r_{0.002} < 0.55$ (95% CL) and the combination of WMAP and the lensing-normalized SDSS galaxy survey implies $r_{0.002} < 0.28$ (95% CL) (MacTavish *et al.* 2006).

On angular scales corresponding to the multipole range $50 < \ell < 150$, the B (curl) component of CMB polarization is a unique signature of tensor perturbations from inflation. The amplitude of tensor perturbations is directly proportional to Hubble parameter during inflation, Loosely speaking, this is related to the Hawking temperature in de-Sitter like space-times. In turn, $H_{\rm inf}$ is related to the energy density $\mathcal{E}_{\rm Inf}$ of the Universe during inflation through the Friedman equation governing cosmological evolution. Hence, the CMB B-polarization is a direct probe of the energy scale of early Universe physics that generated the primordial metric perturbations (scalar & tensor). The relative amplitude of tensor to scalar perturbations, r, sets the energy scale for inflation $\mathcal{E}_{\rm Inf} = 3.4{\times}10^{16}$ GeV $r^{1/4}$. A measurement of B-mode polarization on large scales would give us this amplitude, and hence

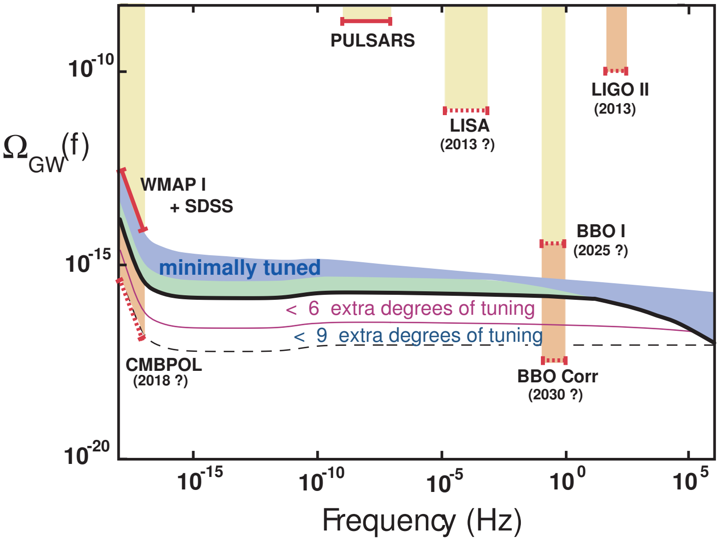


Figure 5. The figure taken from Boyle, Steinhardt & Turok 2006, and NASA/DOE/NSF Task force report on Cosmic Microwave Background research, 2005 (http://www.nsf.gov/mps/ast/tfcr.jsp) shows the theoretical predictions and observational constraints on primordial gravitational waves from inflation. The gravitational wave energy density per logarithmic frequency interval, (in units of the critical density) is plotted versus frequency. The shaded (blue) band labeled 'minimally tuned' represents the range predicted for simple inflation models with the minimal number of parameters and tunings. The dashed curves have lower values of tensor contribution, r, that are possible with more fine tuned inflationary scenarios. The currently existing experimental constraints shown are due to: big bang nucleosynthesis (BBN), binary pulsars, and WMAP-1 (first year) with SDSS. Also shown are the projections for LIGO (both LIGO-I, after one year running, and LIGO-II); LISA; and BBO (both initial sensitivity, BBO-I, and after cross-correlating receivers, BBO-Corr). Also shown is the projected sensitivity of a future space mission for CMB polarization (CMBPol).

a direct determination of the energy scale of inflation. Besides being a generic prediction of inflation, the cosmological gravity wave background from inflation would be also be *a fundamental test of GR on cosmic scales and of the semi-classical behavior of gravity*. Figure 5 summarizes the current theoretical understanding, observational constraints, and future possibilities for the stochastic gravity wave background from inflation. The stochastic gravitational wave background from inflation is expected to exist from cosmological scales down to terrestrial scales. The first CMB normalized GW spectra from inflation using the COBE results was given by Souradeep & Sahni (1992) from IUCAA. This prediction will be targeted by both CMB polarization experiments, as well as by future GW observatories in space, such as Big Bang Observatory (BBO), DECIGO and LISA (Marx *et al.* 2010).

Gravitational lensing of the stronger CMB E-polarization by the ongoing process of structure formation along the line of sight to the last scattering surface also generates a significant B-mode polarization, but on smaller angular scales($\ell > 200$). This prediction is shown as the black curve in bottom left panel for $C_\ell^{\rm BB}$ measurements in Fig. 4. The lensing signal carries important information about the matter power spectrum and its evolution over a range of redshift inaccessible to other observations. This promises to be a powerful probe for constraining the nature of dark energy and, also more excitingly, for determining the neutrino masses. Recent studies indicate that measuring the lensing polarization signal to the cosmic variance limit, can potentially place limits on the total mass of neutrinos at a level comparable to the measured mass differences from neutrinos oscillations.

While there has been no detection of cosmological signal in B-mode polarization, the lack of B-mode power suggests that foreground contamination from polarized emission from our own Galaxy is at a manageable level and is very encouraging news for the prospects of future measurements. The Planck satellite launched in May 2009 will greatly advance our knowledge of CMB polarization by providing foreground/cosmic variance–limited measurements of $C_\ell^{\rm TE}$ and $C_\ell^{\rm EE}$ out beyond $l \sim 1000$. We also expect to detect the weak lensing signal in $C_\ell^{\rm BB}$, although with relatively low precision, that is required for placing ultimate limits on the total neutrino mass. Perhaps, Planck could also detect the stochastic inflationary gravitational wave background if it exists at a level of $r \sim 0.1$. Dedicated future CMB polarization space missions are under study at both NASA and ESA for the time frame 2020+.[3] Lower budget missions would primarily target the low multipole B-mode polarization signature of gravity waves and consequently, identify the viable sectors in the space of inflationary parameters. More ambitious plans, such as COrE,[4] target the entire useful $C_\ell^{\rm BB}$ spectrum and also aim to probe other exciting results from CMB weak lensing measurements.

3. Beyond the angular power spectra of the CMB sky

It is well appreciated that in ‘classical’ big bang model the initial perturbations would have had to be generated ‘acausally’. Besides resolving a number of other problems of classical Big Bang, inflation provides a mechanism for generating these apparently ‘acausally’ correlated primordial perturbations (Starobinsky 1982; Guth & Pi 1982; Bardeen *et al.* 1983). There is increasing effort towards establishing this observationally. There are subtle observations of the CMB sky that could reveal more clearly the mechanism for generations of primordial fluctuation, or, perhaps surprise us by producing insurmountable challenges to the inflation paradigm.

3.1 Statistical isotropy of the Universe

The *Cosmological Principle* that led to the idealized FRW Universe found its strongest support in the discovery of the (nearly) isotropic, Planckian, cosmic microwave background. The isotropy around every observer leads to spatially homogeneous cosmological models.

[3] NASA/DOE/NSF Task force report on Cosmic Microwave Background research, 2005. http://www.nsf.gov/mps/ast/tfcr.jsp (Also available at the Legacy Archive for Microwave Background Data analysis (LAMBDA) site http://lambda.gsfc.nasa.gov).

[4] Cosmic Origins Explorer (COrE) proposal, http://www.core-mission.org.

The large scale structure in the distribution of matter in the Universe (LSS) implies that the symmetries incorporated in FRW cosmological models ought to be interpreted statistically. These are also predicted in the simplest models of inflation.

The CMB anisotropy and its polarization is currently the most promising observational probe of the global spatial structure of the Universe on length scales close to, and even somewhat beyond, the 'horizon' scale ($\sim cH_0^{-1}$). The exquisite measurement of the temperature fluctuations in the CMB provide an excellent test bed for establishing the statistical isotropy (SI) and homogeneity of the Universe. In 'standard' cosmology, the CMB anisotropy signal is expected to be statistically isotropic, i.e., statistical expectation values of the temperature fluctuations $\Delta T(\hat{q})$ are preserved under rotations of the sky. In particular, the angular correlation function $C(\hat{q}, \hat{q}') \equiv \langle \Delta T(\hat{q}) \Delta T(\hat{q}') \rangle$ is rotationally invariant for Gaussian fields. In spherical harmonic space, where $\Delta T(\hat{q}) = \sum_{lm} a_{lm} Y_{lm}(\hat{q})$, the condition of *statistical isotropy* (SI) translates to a diagonal $\langle a_{lm} a^*_{l'm'} \rangle = C_l \delta_{ll'} \delta_{mm'}$ where C_l, is the widely used angular power spectrum of CMB anisotropy. The C_l is a complete description only of a (Gaussian) SI CMB sky and would be (in principle) an inadequate measure for comparing models when SI is violated (Bond, Pogosyan & Souradeep 1998, 2000a,b).

Interestingly enough, the statistical isotropy of CMB has come under a lot of scrutiny after the WMAP results. Tantalizing evidence of SI breakdown (albeit, in very different guises) has mounted in the *WMAP* first year sky maps, using a variety of different statistics. It was pointed out that the suppression of power in the quadrupole and octupole are aligned (Tegmark, de Oliveira-Costa & Hamilton 2004). Further "multipole-vector" directions associated with these multipoles (and some other low multipoles as well) appear to be anomalously correlated (Copi, Huterer & Starkman 2004; Schwartz *et al.* 2004). There are indications of asymmetry in the power spectrum at low multipoles in opposite hemispheres (Eriksen *et al.* 2004). Analysis of the distribution of extrema in *WMAP* sky maps has indicated non-Gaussianity, and to some extent, violation of SI (Larson & Wandelt 2004). The more recent WMAP maps are consistent with the first-year maps up to a small quadrupole difference. The additional years of data and the improvements in analysis have not significantly altered the low multipole structures in the maps (Hinshaw *et al.* 2007). Hence, 'anomalies' persist at the same modest level of significance and are unlikely to be artifacts of noise, systematics, or the analysis in the first year data. The cosmic significance of these 'anomalies' remains debatable also because of the aposteriori statistics employed to ferret them out of the data. The WMAP team has devoted an entire publication to discussing and presenting a detailed analysis of the various anomalies (Bennett *et al.* 2011).

The observed CMB sky is a single realization of the underlying correlation, hence detection of SI violation, or correlation patterns, poses a great observational challenge. It is essential to develop a well defined, mathematical language to quantify SI and the ability to ascribe statistical significance to the anomalies unambiguously. The Bipolar spherical harmonic (BipoSH) representation of CMB correlations has proved to be a promising avenue to characterize and quantify violation of statistical isotropy.

Two point correlations of CMB anisotropy, $C(\hat{n}_1, \hat{n}_2)$, are functions on $S^2 \times S^2$, and hence can be generally expanded as

$$C(\hat{n}_1, \hat{n}_2) = \sum_{l_1, l_2, \ell, M} A^{\ell M}_{l_1 l_2} Y^{l_1 l_2}_{\ell M}(\hat{n}_1, \hat{n}_2)\,. \tag{5}$$

Here $A^{\ell M}_{l_1 l_2}$ are the Bipolar Spherical harmonic (BipoSH) coefficients of the expansion and $Y^{l_1 l_2}_{\ell M}(\hat{n}_1, \hat{n}_2)$ are bipolar spherical harmonics. Bipolar spherical harmonics form an orthonormal basis on $S^2 \times S^2$ and transform in the same manner as the spherical harmonic function with ℓ, M with respect to rotations. Consequently, inverse-transform of $C(\hat{n}_1, \hat{n}_2)$ in eq. (5) to obtain the BipoSH coefficients of expansion is unambiguous.

Most importantly, the Bipolar Spherical Harmonic (BipoSH) coefficients, $A^{\ell M}_{l_1 l_2}$, are linear combinations of *off-diagonal elements* of the harmonic space covariance matrix,

$$A^{\ell M}_{l_1 l_2} = \sum_{m_1 m_2} \langle a_{l_1 m_1} a^*_{l_2 m_2} \rangle (-1)^{m_2} C^{\ell M}_{l_1 m_1 l_2 - m_2} \tag{6}$$

where $C^{\ell M}_{l_1 m_1 l_2 m_2}$ are Clebsch-Gordan coefficients and completely represent the information of the covariance matrix.

Statistical isotropy implies that the covariance matrix is diagonal, $\langle a_{lm} a^*_{l'm'} \rangle = C_l \; \delta_{ll'} \delta_{mm'}$ and hence the angular power spectra carry all the information about the field. When statistical isotropy holds, BipoSH coefficients, $A^{\ell M}_{ll'}$, are zero except those with $\ell = 0, M = 0$ which are equal to the angular power spectra up to a $(-1)^l (2l+1)^{1/2}$ factor. Therefore to test a CMB map for statistical isotropy, one should compute the BipoSH coefficients for the maps and look for nonzero BipoSH coefficients. *Statistically significant deviations of the BipoSH coefficient of map from zero would establish violation of statistical isotropy.*

Since $A^{\ell M}_{l_1 l_2}$ form an equivalent representation of a general two point correlation function, cosmic variance precludes measurement of every individual $A^{\ell M}_{l_1 l_2}$. There are several ways of combining BipoSH coefficients into different observable quantities that serve to highlight different aspects of SI violations. Among the several possible combinations of BipoSH coefficients, the Bipolar Power Spectrum (BiPS) has proved to be a useful tool with interesting features (Hajian & Souradeep 2003; Hajian, Souradeep & Cornish 2005). The BiPS of CMB anisotropy is defined as a convenient contraction of the BipoSH coefficients

$$\kappa_\ell = \sum_{l,l',M} W_l W_{l'} \left| A^{\ell M}_{ll'} \right|^2 \geq 0 \tag{7}$$

where W_l is the window function that corresponds to smoothing the map in real space by a symmetric kernel in order to target specific regions of the multipole space and to isolate the SI violation on corresponding angular scales.

The BipoSH coefficients can be summed over l and l' to reduce the cosmic variance, to obtain reduced BipoSH (rBipoSH) coefficients (Hajian & Souradeep 2006)

$$A_{\ell M} = \sum_{l=0}^{\infty} \sum_{l'=|\ell-l|}^{\ell+l} A^{\ell M}_{ll'}. \tag{8}$$

Reduced bipolar coefficients provide orientation information for the correlation patterns. An interesting way of visualizing these coefficients is to make a *Bipolar map* from $A_{\ell M}$

$$\Theta(\hat{n}) = \sum_{\ell=0}^{\infty} \sum_{M=-\ell}^{\ell} A_{\ell M} Y_{\ell M}(\hat{n}). \tag{9}$$

The symmetry $A_{\ell M} = (-1)^M A^*_{\ell -M}$ of reduced bipolar coefficients guarantees reality of $\Theta(\hat{n})$.

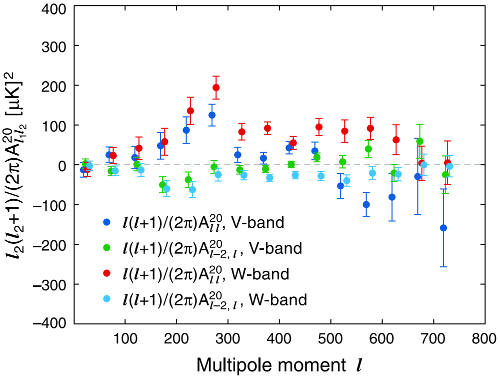


Figure 6. Figure taken from WMAP-7 yr publication on anomalies in the CMB sky (Bennett *et al.* 2011). It shows the measured quadrupolar (bipolar index $L = 2$) bipolar power spectra for V-band and W-band WMAP data, using the KQ75y7 mask. The spherical multipoles have been binned within uniform bands $\delta l = 50$. Only the components of the bipolar power spectra with M = 0 in ecliptic coordinates are shown. A statistically significant quadrupolar effect is seen, even for a single frequency band in a single angular bin.

It is also possible to obtain a measurable band power measure of $A^{\ell M}_{l_1 l_2}$ coefficient by averaging l_1 in bands in multipole space. Recently, the WMAP team has chosen to quantify SI violation in the CMB anisotropy maps by estimating $A^{\ell M}_{ll-i}$ for small value of bipolar multipole, L, band averaged in multipole l. Figure 6 taken from the WMAP-7 release paper (Bennett *et al.* 2011) shows SI violation measured in WMAP CMB maps.

High-resolution CMB polarization maps over large areas of the sky will be delivered by experiments in the near future from Planck. The statistical isotropy of the CMB polarization maps will be an independent probe of the cosmological principle. Since CMB polarization is generated at the surface of last scattering, violations of statistical isotropy are pristine cosmic signatures and more difficult to attribute to the local Universe. The Bipolar Power spectrum has been defined and implemented for CMB polarization and shows great promise (Basak, Hajian & Souradeep 2006; Souradeep, Hajian & Basak 2006).

3.2 Gaussian primordial perturbations

The detection of primordial non-Gaussian fluctuations in the CMB would have a profound impact on our understanding of the physics of the early Universe. The Gaussianity of the CMB anisotropy on large angular scales directly implies Gaussian primordial perturbations (Munshi, Souradeep & Starobinsky 1995; Spergel & Goldberg 1999) as are theoretically motivated by inflation. The simplest inflationary models predict only very mild non-Gaussianity that should be undetectable in the WMAP data.

The CMB anisotropy maps (including the non Gaussianity analysis carried out by the WMAP team data; Komatsu *et al.* 2011) have been found to be consistent with a Gaussian random field. Consistent with the predictions of simple inflationary theories, there are no significant deviations from Gaussianity in the CMB maps using general tests such as Minkowski functionals, the bispectrum, or trispectrum in the three year WMAP data (Spergel *et al.* 2007; Komatsu *et al.* 2011). There have however been numerous claims of anomalies in specific forms of non-Gaussian signals in the CMB data from WMAP at large scales (see discussion in sec. 3.1). Recently, a new class of odd-parity bispectra has been discovered enriching the field significantly (Kamionkowski & Souradeep 2011).

4. Summary

The past few years have seen the emergence of a 'concordant' cosmological model that is consistent both with observational constraints from the background radiation of the Universe as well that from the formation of large scale structures. It is certainly fair to say that the present edifice of the 'standard' cosmological models is robust. A set of foundations, and pillars of cosmology have emerged and are each supported by a number of distinct observations (Ostriker & Souradeep 2004; Souradeep 2011).

Besides precise determination of various parameters of the 'standard' cosmological model, observations have also established some important basic tenets of cosmology and structure formation in the Universe – 'acausally' correlated initial perturbations, adiabatic primordial density perturbations, and gravitational instability as the mechanism for structure formation. The favoured, concordance model inferred is a spatially flat accelerating Universe, where structures have formed by the gravitational evolution of nearly scale invariant, adiabatic perturbations, as expected from inflation. The signature of primordial perturbations observed through the CMB anisotropy and polarization is the most compelling evidence for new, possibly fundamental, physics in the early Universe that underlies the scenario of inflation (or related alternatives). Searches are also on for subtle signals in the CMB maps beyond the angular power spectrum that might violate statistical isotropy (Hajian & Souradeep 2003, 2006; Hajian, Souradeep & Cornish 2005; Basak, Hajian & Souradeep 2006; Souradeep, Hajian & Basak 2006), or Gaussianity (Munshi, Souradeep & Starobinksy 1995; Spergel & Goldberg 1999).

Cosmology is a branch of physics that has seen theoretical enterprise at its best. During the long period of sparse observations in its history, brilliant theoretical ideas (and prejudices) shaped a plausible self-consistent scenario. But current cosmology is passing through a revolution. In the recent past, cosmology has emerged as a data rich field increasingly driven by exquisite and grand observations of unprecedented quality and quantity. These observations have transformed cosmology into an emergent precision science of this century. Further, CMB polarization is arguably emerging as a key observable that can also address fundamental questions related to the origin of the Universe.

Acknowledgments

I would like to thank the editors, D. J. Saikia and Virginia Trimble for inviting me to write this article and for their immense patience in obtaining it from me. It is indeed a privilege to contribute to a commemorative volume honoring an iconic scientist, Prof. S. Chandrasekhar, whom I had the good fortune to meet, and speak with, early in my research career at IUCAA. Finally, I acknowledge my past and present students and collaborators who have contributed to my own research in this exciting area.

References

Bardeen J.M., Steinhardt P.J., Turner M.S., 1983, Phys. Rev. D, 28, 679

Basak S., Hajian A., Souradeep T., 2006, Phys. Rev. D, 74, 021301(R)

Bennett C.L., *et al.*, 2003, ApJS, 148, 1

Bennett, C., *et al.*, 2011, ApJS, 192, 17

Bond J.R. 1996, in Schaeffer R., ed, Cosmology and Large Scale Structure, Les Houches Session LX, August 1993, Elsevier Science Press

Bond J.R., Efstathiou G., 1984, ApJ 285, L45

Bond J.R., Efstathiou G., 1987, MNRAS, 226, 655.

Bond J.R., Pogosyan D., Souradeep T., 1998, Class. Quant. Grav., 15, 2671

Bond J.R., Pogosyan D., Souradeep T., 2000a, Phys. Rev. D, 62, 043005

Bond J.R., Pogosyan D., Souradeep T., 2000b, Phys. Rev. D, 62, 043006

Boyle L.A., Steinhardt P.J., Turok N., 2006, Phys. Rev. Lett., 96, 111301

Brown M.L., *et al.*, 2009, ApJ, 2009, 705, 798

Chandrasekhar S., 1960, Radiative transfer, Dover publications, New York

Chiang H.C., *et al.*, 2010, ApJ, 711, 1123

Copi C.J., Huterer D., Starkman G.D., 2004, Phys. Rev. D, 70, 043515

Eriksen H. K., *et al.*, 2004, ApJ, 605, 14

Guth A.H., Pi S.-Y., 1982, Phys. Rev. Lett., 49, 1110

Hajian A., Souradeep T., 2003, ApJL, 597, L5

Hajian A., Souradeep T., 2006, Phys.Rev., D74, 123521

Hajian A., Souradeep T., Cornish N., 2005, ApJL, 618, L63

Hinshaw G., *et al.*, 2007, ApJS, 170, 288

Hu W., Dodelson S., 2002, ARA&A, 40, 171

Hu W., White M., 1997, New Astron., 2, 323

Hu W., Sugiyama N., Silk J., 1997, Nature, 386, 37

Kamionkowski M., Souradeep T., 2011, Phys. Rev. D, 83, 027301

Kogut A., et. al., 2003, ApJS., 148, 161

Komatsu E., *et al.*, 2011 ApJS, 192, 18

Kovac J. M., *et al.*, 2002 Nature, 420, 772

Larson D., et.al., 2011 ApJS, 192, 16

Larson D.L., Wandelt B.D., 2004, ApJL, 613, L85

Lue A., Wang L., Kamionkowski M., 1999, Phys. Rev. Lett., 83, 1506

MacTavish C.J., *et al.*, 2006, ApJ, 647, 799

Maity D., Majumdar P., Sengupta S., 2004, JCAP, 0406, 005

Marx J., *et al.*, 2010, GWIC Roadmap document, https://gwic.ligo.org/roadmap/

Munshi D., Souradeep T., Starobinsky A., 1995, ApJ, 454, 552

Ostriker J.P., Souradeep T., 2004, Pramana, 63, 817

Page L., *et al.*, 2007, ApJS, 170, 335

Peebles P.J.E., Yu J.T.,1970, ApJ, 162, 815

Rybicki G.B., Lightman A.P., 1979, Radiative processes in astrophysics, New York: Wiley–Interscience

Saha R., Jain P., Souradeep T., 2006, ApJL, 645, L89

Samal P., *et al.*, 2010, ApJ, 714, 840

Schwarz D.J., *et al.*, 2004, Phys. Rev. Lett., 93, 221301

Smoot G.F., *et al.*, 1992, ApJL, 396, L1

Souradeep T., 2011, in Proc. GR-19 July 2010,Mexico, eds, Marolf D., Sudarsky D., Class. Q. Grav., in press

Souradeep T., Sahni V., 1992, Mod. Phys. Lett. A, 7, 3541
Souradeep T., Hajian A.,Basak S., 2006, New Astron.Rev., 50, 889
Spergel D.N., Goldberg D.M., 1999, Phys. Rev. D, 59, 103001
Spergel D., *et al.*, 2007, ApJS, 170, 377
Starobinsky A. A., 1982, Phys. Lett., 117B, 175
Tegmark M., de Oliveira-Costa A., Hamilton A., 2004, Phys. Rev. D, 68, 123523
Wu E.Y.S., *et al.*, 2009, Phys. Rev. Lett., 102, 161302

Bull. Astr. Soc. India (2011)

Gravitational wave astronomy — astronomy of the 21^{st} century

S. V. Dhurandhar[*]

Inter University Centre for Astronomy & Astrophysics, Ganeshkhind, Pune – 411 007, India

Received February 11, 2011; accepted February 27, 2011

Abstract. An enigmatic prediction of Einstein's general theory of relativity is gravitational waves. With the observed decay in the orbit of the Hulse-Taylor binary pulsar agreeing within a fraction of a percent with the theoretically computed decay from Einstein's theory, the existence of gravitational waves was firmly established. Currently there is a worldwide effort to detect gravitational waves with inteferometric gravitational wave observatories or detectors and several such detectors have been built or are being built. The initial detectors have reached their design sensitivities and now the effort is on to construct advanced detectors which are expected to detect gravitational waves from astrophysical sources. The era of gravitational wave astronomy has arrived. This article describes the worldwide effort which includes the effort on the Indian front — the IndIGO project —, the principle underlying interferometric detectors both on ground and in space, the principal noise sources that plague such detectors, the astrophysical sources of gravitational waves that one expects to detect by these detectors and some glimpse of the data analysis methods involved in extracting the very weak gravitational wave signals from detector noise.

Keywords : gravitational waves – black holes – stars: binaries – techniques: interferometric – instrumentation: interferometers

1. Introduction

In the past half a century or so, astronomy has been revolutionised by several unexpected discoveries because of the plethora of windows being opened in various bands of the electromagnetic spectrum. To name a few, the cosmic microwave background and the discovery of pulsars in radio band, gamma-ray bursts, X-ray objects, all go to show that whenever a new window has been opened, startling discoveries have followed. Yet another window to the Universe should soon open up in few years time - the gravitational wave (GW) window. Not only will this window test Einstein's general theory of relativity, but also provide direct evidence for black holes, and more generally test general relativity in the strong field

[*] e-mail: sanjeev@iucaa.ernet.in

regime. Just as the other windows to the Universe have brought in unexpected discoveries, it is not unreasonable to expect the same in this case also — and more so, because this involves changing the physical interaction from electromagnetic to gravitational.

The existence of gravitational waves predicted by the theory of general relativity, has long been verified 'indirectly' through the observations of Hulse and Taylor (Hulse & Taylor 1975; Taylor 1994). The inspiral of the members of the binary pulsar system named after them has been successfully accounted for in terms of the back-reaction due to the radiated gravitational waves — the observational results and the theory agree with each other within a fraction of a percent. However, detecting such waves directly with the help of detectors based either on ground or in space has not been possible so far.

The key to gravitational wave detection is the very precise measurement of small changes in distance. For laser interferometers, this is the distance between pairs of mirrors hanging at either end of two long, mutually perpendicular vacuum chambers. A GW passing through the instrument will shorten one arm while lengthening the other. By using an interferometric design, the relative change in length of the two arms can be measured, thus signalling the passage of a GW at the detector site. GW detectors produce an enormous volume of output consisting mainly of noise from a host of sources both environmental and intrinsic. Buried in this noise will be the GW signature. Sophisticated data analysis techniques are needed to optimally extract the GW signal from the interferometric data. IUCAA has made significant contributions in this area.

2. Interferometric detection of GW

Historically with pioneering efforts of Weber in the 1960s, the detectors were resonant bar detectors which were suspended, seismically isolated, aluminium cylinders. The later versions were cooled to extremely low temperatures — ultracryogenic — to suppress the thermal noise. There are also spherical resonant mass detectors being constructed/operating. However, these ideas although interesting and useful in their niche, have their limitations. In this article we will confine ourselves to the interferometric detectors.

2.1 The principle of interferometric detection

A weak GW is described by a metric perturbation $h_{\mu\nu}$ in general relativity. Typically, for the astrophysical GW sources which are amenable to detection, $h_{\mu\nu} \sim 10^{-22}$. In the transverse-traceless gauge, the $h_{\mu\nu}$ can be expressed in terms of just two amplitudes, h_+ and $h_\times$, called the 'plus' and 'cross' polarisations. If a weak monochromatic gravitational wave of + polarisation is incident on a ring of test-particles, the ring is deformed into an ellipse as shown at the top of Figure 1. Phases, a quarter cycle apart, of the GW are shown in the Figure. For the × polarisation the ellipses are rotated by an angle of 45°. A general wave is a linear combination of the two polarisations.

At the bottom of Figure 1, a schematic of the interferometer is depicted. If the change in the armlength L is δL, then,

$$\delta L \sim hL, \tag{1}$$

where h is a typical component of the metric perturbation.

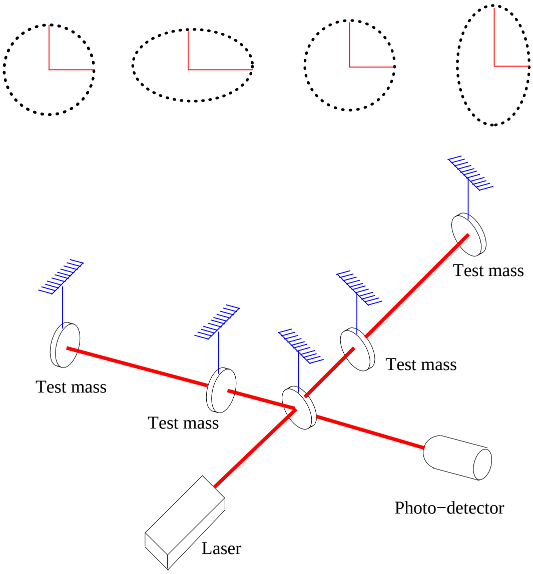


Figure 1. Upper: A circular ring of test particles is deformed into an ellipse by an incident GW. Phases, a quarter of a cycle apart are shown for the + polarisation. The length change in the interferometric arms is also shown schematically. Lower: a schematic diagram of an interferometer is drawn.

For a GW source, h can be estimated from the well-known Landau-Lifschitz quadrupole formula. The GW amplitude h is related to the second time derivative of the quadrupole moment (which has dimensions of energy) of the source:

$$h \sim \frac{4}{r}\,\frac{G}{c^4}\,E^{\rm kinetic}_{\rm nonspherical}, \tag{2}$$

where r is the distance to the source, G is the gravitational constant, c is the speed of light and $E^{\rm kinetic}_{\rm nonspherical}$ is the kinetic energy in the *nonspherical* motion of the source. If we consider $E^{\rm kinetic}_{\rm nonspherical}/c^2$ a fraction of a solar mass and the distance to the source ranging from galactic scale of tens of kpc to cosmological distances of Gpc, then h ranges from 10^{-17} to 10^{-22}. These numbers then set the scale for the sensitivities at which the detectors must operate. The factor of 4 is also important given the weakness of the interaction and the subsequent signal extraction from detector noise.

2.2 Ground-based interferometric detectors

There are a host of noise sources in ground-based interferometric detectors which contaminate the data. At low frequencies there is the seismic noise. The seismic isolation is a sequence of stages consisting of springs/pendulums and heavy masses. Each stage has a low resonant frequency about a fraction of a Hz. The seismic isolation acts as a low pass filter, attenuating high frequencies, but low frequencies get through. This results in a ‘noise wall’ at low frequencies and marks the lower end of the detector bandwidth. It is about 40 Hz for initial detectors but will go down to 10 Hz for advanced detectors increasing the bandwidth. At mid-frequencies up to a few hundred Hz, the thermal noise is important and is due to the thermal excitations both in the test masses — the mirrors — as well as the seismic suspensions. Currently, this seems to be the noise hardest to suppress. The natural modes of the mirrors and the suspension are driven by the thermal excitations. One ‘solution’ is to cool the mirrors/suspensions, but this has its own problems. Nevertheless, the Japanese have planned a detector doing just this — the Large-scale Cryogenic Gravitational-wave Telescope (LCGT) which has been funded recently. At high frequencies the shot noise from the laser dominates. This noise is due to the quantum nature of light. From photon counting statistics and the uncertainty principle, the phase fluctuation is inversely proportional to the square root of the mean number of photons arriving during a period of the wave. So increasing the laser power and hence the mean number of photons during a given period of the wave tends to reduce this noise. Apart from these main noise sources there are other noise sources, an important one among them is gravity gradient noise which cannot be screened and occurs only at low frequencies. The slowly changing gravity gradients are due to natural causes (such as clouds moving in the sky, changes in atmospheric density) or are manmade. Thus long arm lengths, high laser power, and extremely well-controlled laser stability are essential to reach the requisite sensitivity. Figure 2 shows the sensitivity achieved by the initial LIGO detectors (Gonzalez 2005) when the actual noise in the detectors reached theoretical design sensitivity (shown by the bold curve). The sensitivity has continued to improve with time.

2.3 The worldwide network of ground-based interferometric gravitational wave observatories

The USA has been at the forefront in building large scale detectors. The LIGO project (Abramovici *et al.* 1992) has built three detectors, two of armlength 4 km and one of armlength 2 km at two sites about 3000 km apart at Hanford, Washington and at Livingston, Louisiana. The 2 km detector is at Hanford. These initial detectors have had several science runs and the design sensitivity has not only been reached but surpassed. The goal of this initial stage was mainly to vindicate the technologies involved in attaining the design sensitivities. Now the next phase is to build advanced detectors with state of the art technologies which will be capable of observing GW sources and doing GW astronomy. With these future goals a radical decision has been taken by the LIGO project, that of building one of its detectors in Australia — that is LIGO will build two advanced detectors in US and partially fund a full scale detector in Australia with advanced design. This detector is called LIGO-Australia and will be built in collaboration with the Australians who already have an interferometric facility at Gingin near Perth — the AIGO (Australian Interferometric Gravitationalwave Observatory) project. The reason for this decision by the US is clear — it is to increase the baseline and have a detector far removed from other detectors on Earth, which has several advantages, such as improving the localization of the GW source.

In Europe the large-scale project is the VIRGO project (Bradaschia *et al.* 1990) of Italy and France which has built a 3 km armlength detector. After commissioning of the project

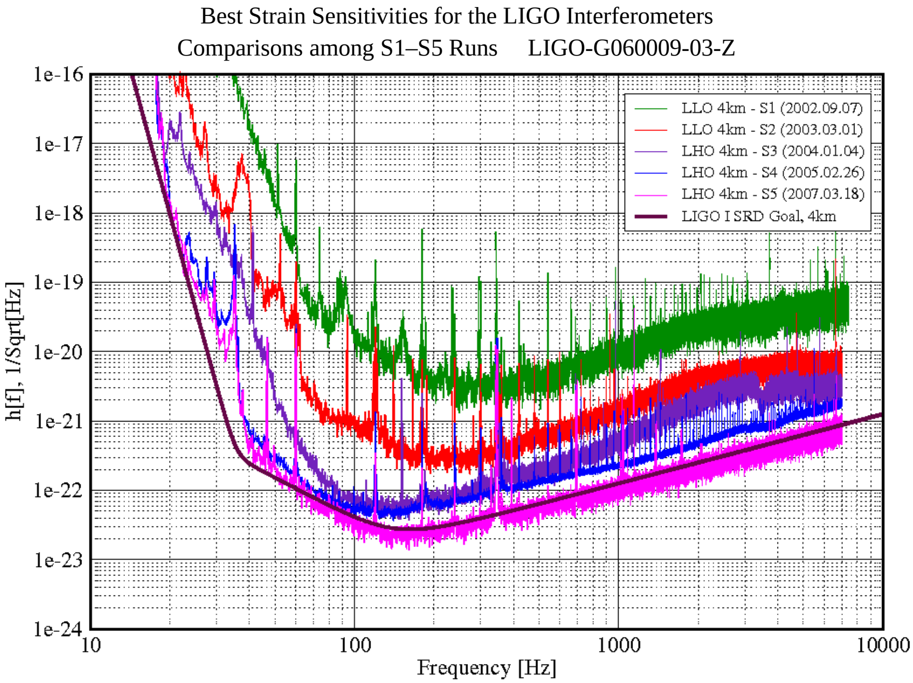


Figure 2. The figure shows the sensitivity achieved by LIGO detectors by March 2007. This sensitivity level has been surpassed in later operations (Gonzalez 2005; see LIGO website).

in 2007, it also had science runs. The GEO600 (Danzmann *et al.* 1995) is a German-British project and whose detector has been built near Hannover, Germany with an armlength of 600 metres. One of the goals of GEO600 is to develop advanced technologies required for the next generation detectors with the aim of achieving higher sensitivity.

Japan was the first (around 2000) to have a large scale detector of 300 m armlength — the TAMA300 detector under the TAMA project (Tsubono 1995) — operating continuously at high sensitivity in the range of $h \sim 10^{-20}$. Now Japan plans to construct a cryogenic inteferometric detector called the LCGT (Large-scale Cryogenic Gravitational wave Telescope; Kuroda 2006) which has been recently funded. The purpose of the cryogenics is to quell the thermal noise. But this technnology is by no means straight forward and will test the skills of the experimenters.

Australia is looking for international partners, because of LIGO-Australia. Given the twenty year old legacy in GW data analysis at IUCAA, Pune and waveform modelling at RRI, Bangalore, Australia would welcome the Indians as partners in this endeavour. Recently, about two years ago, an Indian Initiative in Gravitational Wave Astronomy (IndIGO) has begun whose goal is to promote and foster gravitational wave astronomy in India and join in the worldwide quest to observe gravitational waves. Apart from the data analysis this initiative includes the all important experimental aspect. Accordingly a modest beginning has been made by IndIGO with TIFR, Mumbai approving a 3 metre prototype on which Indian experimenters can get first hand experience and develop expert manpower. This project has already been funded. Concurrently, an MOU with Australia has been signed which purports to ask for funding from Indian agencies in parallel with Australia. An IndIGO consortium has been formed with scientists from leading institutions such as TIFR, RRCAT, RRI, IUCAA, IISERs, Delhi University and CMI, and also including scientists (mainly Indian) working abroad. The current strength of the consortium is about 25 scientists. In order to further this effort the first goal is to muster up sufficient expert and skilled manpower which will launch this activity. It will mean India getting into this worldwide challenging experiment.

Besides the current projects, studies have begun for third generation detectors which will include further advanced technologies to enhance the sensitivities of GW detectors to reach out farther in the sky; the Einstein Telescope (ET) is just such a future goal.

2.4 Space-based detectors: the LISA project

A natural limit occurs on decreasing the lower frequency cut-off beyond ~10 Hz because it is not practical to increase the arm-lengths on ground and also because of the gravity gradient noise which is difficult to eliminate below 10 Hz. Thus, the ground based interferometers will not be sensitive below the limiting frequency of ~10 Hz. But on the other hand, there exist in the cosmos, interesting astrophysical GW sources which emit GW below this frequency such as the galactic binaries, massive and super-massive black hole binaries. If we wish to observe these sources, we need to go to lower frequencies. The solution is to build an interferometer in space, where such noises will be absent and allow the detection of GW in the low frequency regime. **LISA** — *Laser Interferometric Space Antenna* — is a proposed ESA-NASA mission which will use coherent laser beams exchanged between three identical spacecrafts forming a giant (almost) equilateral triangle of side 5×10^6 kilometers to observe and detect low-frequency cosmic GW.[1] The ground-based detectors and LISA complement each other in the observation of GW in an essential way, analogous to the way optical, radio, X-ray, γ-ray observations do for electromagnetic waves. As these detectors begin to operate, a new era of *gravitational astronomy* is on the horizon and a radically different view of the Universe is expected to be revealed. There are also further space projects being considered.

LISA consists of three spacecrafts, flying five million kilometres apart, in an equilateral triangle. The spacecrafts are maintained drag-free by a complex system of accelerometers and micro-propellers. Each spacecraft will carry two optical assemblies that contain the main optics and a free-falling inertial sensor. The light sent out by a laser in one spacecraft is received by the telescope on the distant spacecraft. The incoming light from the distant spacecraft is then mixed with the in-house laser and the differential phase is recorded. This defines one elementary data stream. There are thus six elementary data streams which are formed by going clockwise and anti-clockwise around the LISA triangle. Suitable combinations of these elementary data streams can be used to optimally extract the GW signal from the instrumental noise. In other words, LISA is basically a giant Michelson

[1] http://sci.esa.int/science-e/www/area/index.cfm?fareaid=27; http://lisa.gsfc.nasa.gov

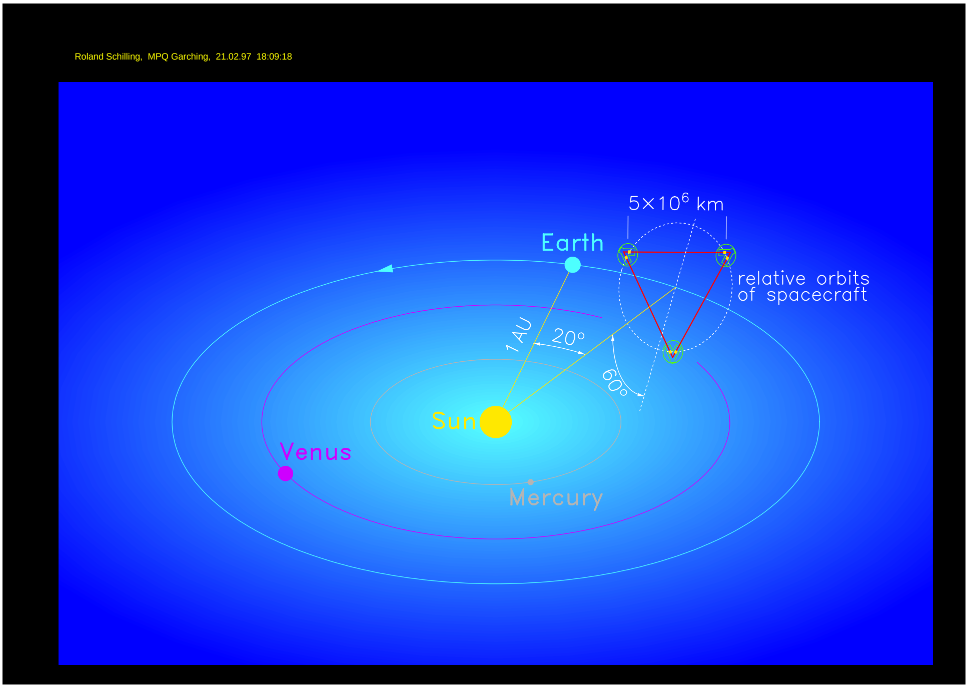


Figure 3. LISA orbital configuration around the Sun, describing a cone with 60° half opening angle. The centroid of the triangle follows an Earth-like orbit trailing 20° behind (Bender *et al.* 1998).

interferometer placed in space, with a third arm added to give independent information on the two gravitational wave polarisations, and for redundancy. The distance between the spacecrafts — the interferometer arm-length — determines the frequency range in which LISA can make observations; it was carefully chosen to allow for the observations of most of the interesting sources of gravitational radiation. Each spacecraft revolves in its own heliocentric orbit. The centre of LISA's triangle will follow Earth's orbit around the Sun, trailing 20° behind. It will maintain a distance of 1 AU (astronomical unit) from the Sun, the average distance between the Earth and the Sun (Figure 3). The spacecrafts rotate in a circle drawn through the vertices of the triangle and the LISA constellation as a whole revolves around the Sun. LISA's operational position was chosen as a compromise between the need to minimise the effects on the spacecrafts of changes in the Earth's gravitational field and the need to be close enough to the Earth for easy communication.

LISA will observe low-frequency GW in the range 0.1 mHz to 0.1 Hz. Since astrophysical systems are generally large and in spite of high velocities do not change their quadrupole moment too quickly, the Universe is richly populated with sources in this frequency band. Also the masses that produce GW in this frequency band are generally large and thus produce stronger GW than those in ground-based detectors, leading to high signal-to-noise ratios (SNR). The signals for LISA arise from a large variety of phenomena, such as merging massive and supermassive black holes, vibrating black holes (quasi-normal modes), stellar mass objects falling into massive and supermassive black holes and GWs of cosmological origin. The high SNRs of these signals imply detailed and accurate information which can test general relativity and its ramifications to unprecedented accuracies. Astrophysics of various objects like compact binaries, stellar remnants can be studied and LISA observations can provide useful clues to events in the early Universe (Bender & Hils 1997; Nelemans, Yungelson & Portegies Zwart 2001; Postnov & Prokhorov 1998; Hills & Bender 2000).

LISA sensitivity is limited by several noise sources. A major noise source is the laser phase (frequency) noise which arises due to phase fluctuations of the master laser. Amongst the important noise sources, laser phase noise is expected to be several orders of magnitude larger than other noises in the instrument. The current stabilisation schemes estimate this noise to about $\Delta\nu/\nu_0 \simeq 3 \times 10^{-14}/\sqrt{\mathrm{Hz}}$, where ν_0 is the frequency of the laser and $\Delta\nu$ the fluctuation in frequency. If the laser frequency noise can be suppressed then the noise floor is determined by the optical-path noise which acts like fluctuations in the lengths of optical paths and the residual acceleration of proof masses resulting from imperfect shielding of the drag-free system. The noise floor is then at an effective GW strain sensitivity $h \sim 10^{-21}$ or 10^{-22}. Thus, cancelling the laser frequency noise is vital if LISA is to reach the requisite sensitivity.

In ground-based detectors the arms are chosen to be of equal length so that the laser light experiences identical delay in each arm of the interferometer. This arrangement precisely cancels the laser frequency/phase noise at the photodetector. However, in LISA it is impossible to achieve equal distances between spacecrafts and also the data are taken at a phasemeter as a beat note between the local oscillator and the incoming beam coming from a spacecraft 5 million km away. In LISA, six data streams arise from the exchange of laser beams between the three spacecrafts — it is not possible to bounce laser beams between different spacecrafts, as is done in ground-based detectors. The technique of time-delay interferometry (TDI) is used (Armstrong, Estabrook & Tinto 1999; Estabrook, Tinto & Armstrong 2000) which combines the recorded data with suitable time-delays corresponding to the three arm-lengths of the giant triangular interferometer. An original approach to this problem was taken by IUCAA. A *systematic method* based on modules over polynomial rings has been successfully formulated which is most appropriate for this problem (Dhurandhar, Nayak & Vinet 2000; 2010). The method uses the redundancy in the data to suppress the laser frequency noise.

3. General discussion of GW sources

3.1 GW sources

Several types of GW sources have been envisaged which could be directly observed by Earth-based detectors: (i) Burst sources — such as binary systems consisting of neutron stars and/or black holes in their inspiral phase or merger phase; supernova explosions — whose signals last for a time between a few milli-seconds and a few minutes, much shorter, than the typical observation time; (ii) stochastic backgrounds of radiation, of either primordial or astrophysical origin, and (iii) continuous wave sources — e.g. rapidly rotating non-axisymmetric neutron stars — where a weak sinusoidal signal is continuously emitted.

As one sees from the discussion that follows, the strengths of these sources are usually well below or even way below, the mean noise level in the detectors either currently operating or even for those planned in the near future — the advanced detectors. This situation makes the expert data analysis all the more vital, firstly in detecting the source, and secondly and more importantly in extracting astrophysical information about it.

Inspiraling binaries have been considered as highly promising sources not only because of the enormous GW energy they emit, but also because they are such 'clean' systems to model; the inspiral waveform can be computed accurately to several post-Newtonian orders adequate for optimal signal extraction and parameter estimation (Blanchet *et al.* 2004). The typical strength of the source is:

$$h \sim 2.5 \times 10^{-23} \left(\frac{\mathcal{M}}{M_\odot}\right)^{5/3} \left(\frac{f}{100\ \mathrm{Hz}}\right)^{2/3} \left(\frac{r}{100\ \mathrm{Mpc}}\right)^{-1}, \tag{3}$$

where $\mathcal{M}$ is the chirp mass equal to $(\mu M^{2/3})^{3/5}$, μ and M being respectively the reduced and the total mass of the system, r is the distance to the source — it is given at the scale of 100 Mpc because such events would be rare and therefore to obtain a reasonable event rate, a sufficient volume of the Universe needs to be covered — and f is the instantaneous fiducial frequency of the source as the source evolves adiabatically during the inspiral stage. Since the phase of the waveform, apart from the amplitude, can be computed accurately by post-Newtonian methods, the optimal extraction technique of matched filtering is used. In the recent past, numerical relativity has been able to make a breakthrough by continuing the waveform to the merger phase and eventually connect it with the ringdown of the final black hole. It is here that Chandrasekhar's contribution stands out because he pointed out that a black hole rings like a bell if it is subjected to a perturbation (Chandrasekhar & Detweiler 1975). In the current context this occurs in the final stages of the merger when a black hole is formed. Quasi-normal modes were first discovered by Vishveshwara (1970) while examining the stability of the Schwarzschild black hole.

Another important burst source of GW is supernovae. It is difficult to reliably compute the waveforms for supernovae, because complex physical processes are involved in the collapse and the resulting GW emission. This limits the data analysis and optimal signal extraction.

Continuous wave sources pose one of the most computationally intensive problems in GW data analysis (Schutz 1989; Brady *et al.* 1998; Cutler, Gholami & Krishnan 2005). A rapidly rotating asymmetrical neutron star is a source of continuous gravitational waves. There are some astrophysical systems known from electromagnetic observations which might be promising sources of continuous GWs. Surveys for continuous GWs have so far not led to a direct detection, but the searches have now become astrophysically interesting. We mention the result for the Crab pulsar in the next subsection. These searches for known systems are not computationally intensive since they target a known sky position, frequency and spindown rate. On the other hand, blind all-sky and broad-band searches for previously unknown neutron stars are a different matter altogether. Long integration times, typically of the order of a few months or years are needed to build up sufficient signal power. The reason for this is that the signal is very weak and lies way below the detector noise level.

We give a typical example:

$$h \sim 10^{-25} \left(\frac{I}{10^{45}\mathrm{gm.cm}^2}\right)\left(\frac{f}{1\mathrm{kHz}}\right)^2 \left(\frac{\epsilon}{10^{-5}}\right)\left(\frac{r}{10\mathrm{kpc}}\right)^{-1}, \tag{4}$$

where I is the moment of inertia of the neutron star, r the distance to the source, f the GW frequency and ϵ is a measure of asymmetry of the neutron star. The asymmetry of a neutron star can occur in various ways such as crustal deformation, intense magnetic fields not aligned with the rotation axis or the Chandrasekhar-Friedman-Schutz instability (Chandrasekhar & Esposito 1970; Friedman & Schutz 1978). This instability is in fact driven by GW emission and consists of strong hydrodynamic waves in the star's surface layers. This phenomenon results in significant gravitational radiation. Earth's motion Doppler modulates the signal, and this Doppler modulation depends on the direction to the GW source. Thus, coherent extraction of the signal whose direction and frequency is unknown is an impossibly computationally expensive task. The parameter space is very large, and a blind survey requires extremely large computational resources.

To detect stochastic background one needs a network of detectors, ideally say two detectors preferably identically oriented and close to one another. The stochastic background arises from a host of unresolved independent GW sources and can be characterised only in terms of its statistical properties. The strength of the source is given by the quantity $\Omega_{\mathrm{GW}}(f)$ which is defined as the energy-density of GW per unit logarithmic frequency interval divided by ρ_{critical}, the energy density required to close the Universe. The typical strength of the Fourier component of the GW strain for the frequency bandwidth $\Delta f = f$ is:

$$\tilde{h}(f) \sim 10^{-26} \left(\frac{\Omega_{\mathrm{GW}}}{10^{-12}}\right)\left(\frac{f}{10\mathrm{Hz}}\right)^{-3/2} \mathrm{Hz}^{-1/2}, \tag{5}$$

The signal is extracted by cross-correlating the outputs. Two kinds of data-analysis methods have been proposed (i) a full-sky search — but this drastically limits the bandwidth (Allen & Romano 1999), (ii) a radiometric search in which the sky is scanned pixel by pixel — since a small part of the sky is searched at a time, it allows for larger bandwidth, and more importantly includes the bandwidth in which the current detectors are most sensitive, thus potentially leading to a large SNR (Mitra *et al.* 2008). Moreover, with this method a detailed map of the sky is obtained.

Apart from these sources, there can be burst sources of GW from mergers or explosions or collapses which may or may not be seen electromagnetically but nevertheless deserve attention. In this case time-frequency methods are the appropriate methods which look for excess power in a given time-frequency box.

In the section on data analysis, we will focus on two of the above mentioned and prominent GW sources, namely, the compact binaries and the continuous wave sources. Before moving on to the data analysis we would like to briefly describe the astrophysically interesting results so far obtained in GW astronomy.

3.2 Astrophysical results from current GW data

Even in this initial stage of the detectors, it is important to note that astrophysically interesting results have been obtained from the data so far taken with the LIGO detectors, more specifically, the data from the S5 run. The data have set astrophysically interesting upper limits on the GW emanating from astrophysical sources. We mention a few of the important results below.

The S5 data have constrained the cosmological GW background in which the upper limit falls below the previous upper limit set by nucleosynthesis (LIGO Science Collaboration and VIRGO Science Collaboration 2009). This result has excluded several string theory motivated big bang models.

The GW data analysis from the S5 run shows that less than 4% of the energy can be radiated away in GW from the Crab pulsar (Abbott *et al.* 2008a) This is because the spindown rate is $\sim 3.7 \times 10^{-10}$ Hz/sec, while no GW signal was observed even as low as $h \sim 2.7 \times 10^{-25}$.

Since no GW signal was detected from the GRB source 0702012, this implies that a compact binary progenitor with masses in the range $1M_\odot < m_1 < 3M_\odot$ and $1M_\odot < m_2 < 40M_\odot$ located in M 31 is excluded as a GW source with 99% confidence. If the binary progenitor was not in M 31, then it rules out a binary star merger progenitor upto a distance of 3.5 Mpc, with 90% confidence and assuming random orientation (Abbott *et al.* 2008b). A search was performed from the LIGO S5 and the Virgo first science runs for the total mass of the component stars ranging from 2 to 35 $\mathrm{M}_\odot$. No GWs were identified. The 90 per cent confidence upper limit on the rate of coalescence of non-spinning binary neutron stars was estimated to be 8.7×10^{-3} yr^{-1} L_{10}^{-1}, where L_{10} is 10^{10} times the blue solar luminosity (Abadie *et al.* 2010).

These are some of the salient astrophysical results which merely serve to indicate the revolutionary scientific impact that GW astronomy can bring to science.

4. Data analysis of GW sources

As can be seen from the foregoing, data analysis of interferometric data is a very important aspect in the quest for detection of gravitational waves. This is because the signal is weak and must be extracted from the noisy data — infact the noise, in general, strongly overwhelms the signal. Thus sophisticated statistical techniques and efficient algorithms based on statistical analysis are vital for extracting the signal from the noise. The data analysis technique of course depends on the nature of the signal. We would like to discuss a couple of sources and their data analysis in more detail. We first describe the matched filtering paradigm for the inspiraling binaries and then describe some current efforts in the so called ‘All sky all frequency search’ for GW from periodic or continuous wave sources, which are based on group theoretic methods. This does not mean that the sources not discussed here are unimportant in any way, but the idea here is to give a flavour of the data analysis methods employed in GW detection. Here we have chosen two such data analysis problems.

In this article we would like to emphasise the importance of the role of symmetries which play a crucial part in increasing the efficiency of an algorithm and in turn reducing the computational burden. The symmetries arise from the physical model of the GW source. The idea is to capture the symmetries in terms of group representation theory and then use the representation theory to develop efficient search algorithms.

4.1 Inspiraling/coalescing binaries

Here we will deal essentially with the inspiral waveform which is the first stage when the stars are relatively far apart, and the stage ends a little before the last stable orbit is reached. The last stable circular orbit for a test particle orbiting a Schwarzschild black hole of mass M is at radial distance of $6MG/c^2$. Here we may take M to be the total mass of the binary components, and then the inspiral stage is the one before the orbit shrinks to around $10M$ or a little less. After the inspiral stage comes the merger stage, and the final stage is that when a single black hole is formed (in the case the masses are two black holes). Just before the final black hole is formed it oscillates, emitting quasi-normal mode radiation finally settling into a stable configuration of a stationary black hole. The merger waveform for black holes can now be computed from numerical relativity; there was a recent breakthrough in 2005 (Pretorius 2005), and this was followed by several groups actually implementing their numerical code (Campanelli *et al.* 2006; Baker *et al.* 2006). The inspiral waveform we will consider also holds for two neutron stars or a neutron star/black hole pair. Here we will restrict ourselves to the binary inspiral and data analysis for it.

4.1.1 *Matched filtering*

The appropriate technique to use, when one has the accurate knowledge of the waveform — especially of the phase — is matched filtering. First, it yields the maximum signal-to-noise (SNR) among all linear filters. Secondly, it is optimal in the Neyman-Pearson sense — in additive Gaussian noise, the matched filter statistic gives the maximum detection probability for a given false alarm rate. The matched filtering operation is defined as follows: if $x(t)$ is the data in the time domain, then the matched filter output $c(\tau)$ at the epoch τ is given by:

$$c(\tau) = \int x(t)q(t+\tau)dt\,, \tag{6}$$

where $q(t)$ is the matched filter. In stationary noise (the noise is independent of absolute time) q has a particularly simple form and is conveniently described in the Fourier domain as:

$$\tilde{q}(f) = \frac{\tilde{h}(f)}{S_h(f)}\,, \tag{7}$$

where $h(t)$ is the expected signal in the detector and $S_h(f)$ is the power spectral density of the noise. An illustration of the matched filtering paradigm is given in Figure 4.

4.1.2 *Searching the parameter space: the spinless case*

In this section we will be considering only the point mass approximation which is valid for black holes and to a large extent for neutron stars if they do not deform. The problem would have been simple if there were a single signal waveform. But the signal depends on several parameters. Thus one is actually searching through a family of signals. The signal has the form:

$$h(t;\mathcal{A}, t_a, \phi_a, \tau_0, \tau_3) = \mathcal{A}a(t-t_a,\tau_0,\tau_3)\cos[\phi(t-t_a,\tau_0,\tau_3)+\phi_a]\,, \tag{8}$$

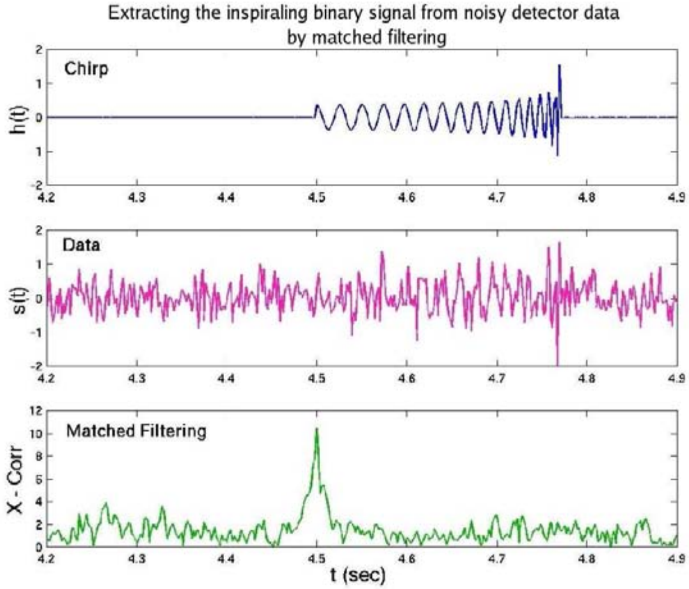


Figure 4. The top part of the figure shows the signal — the inspiral binary waveform usually called the chirp; the middle part shows the signal embedded in the detector noise while the bottom shows the plot of the output of the matched filter $c(\tau)$. By recognising the peak by thresholding, the signal can be detected.

where $\mathcal{A}$ is the amplitude, t_a is the time of arrival of the signal, τ_0 and τ_3 as defined below are functions of the individual masses m_1, m_2 of the binary and ϕ_a the phase at arrival of the wave. The signal described in equation (8) is what is called the restricted post-Newtonian waveform in which the amplitude is of the Newtonian waveform which is slowly varying with time, while the phase is given to as much as accuracy is possible, that is upto the 3.5 post-Newtonian order which is deemed sufficient because it gives a phase accuracy to better than a cycle for the stellar mass objects inspiraling in the bandwidth of the current or even advanced detectors. It is most important for the technique of matched filtering that the phase is known as accurately as possible, because even half a cycle can put the signal waveform out of phase with the template waveform which can lead to substantial decrease in the output of the matched filter. The amplitude $\mathcal{A}$ depends on the chirp mass parameter $\mu M^{2/3}$ and on the fiducial frequency f_a of the wave at the time of arrival t_a. Instead of the masses, it has been found useful to use the chirp times τ_0 and τ_3 as signal parameters — these parameters appear in a simple way in the Fourier transform of the signal, namely, they appear *linearly* in the phase of the Fourier transform in the stationary phase approximation. The final search algorithm becomes simple in terms of these parameters. They are related to M and $\eta = \mu/M$ by the relations:

$$\tau_0 = \frac{5}{256\,\eta f_a}\,(\pi M f_a)^{-5/3}\,, \quad \tau_3 = \frac{1}{8\eta f_a}\,(\pi M f_a)^{-2/3}\,. \tag{9}$$

where τ_0 is the Newtonian time of coalescence and the chirp time τ_3 is related to the 1.5 PN order. As one can see from equation (8), both the amplitude a and phase ϕ depend on these parameters. We do not give the explicit forms of these functions here because they are unimportant to the discussion here, but they can be found in the literature, eg. (Mohanty & Dhurandhar 1996).

We use the maximum likelihood approach, that is, the likelihood ratio must be maximised over the signal parameters, namely, $\{\mathcal{A}, t_a, \phi_a, \tau_0, \tau_3\}$. The maximum likelihood method shows that a matched filter is the simpler surrogate statistic than the likelihood ratio, and it is sufficient to maximise the output of the matched filter over the search parameters. The amplitude $\mathcal{A}$ is readily extracted from the signal by normalising the template waveform. The parameters t_a and ϕ_a are searched for by using the symmetry of the signal family. The signal is *translationally invariant* in time, that is, a signal at another time of arrival is just obtained from translation. This symmetry can be exploited by using the Fast Fourier Transform (FFT), that is, writing equation (6) in the Fourier domain:

$$c(t_a) = \int \frac{\tilde{x}^*(f)\tilde{h}(f)}{S_h(f)} e^{2\pi i f t_a} df + \text{complex conjugate}, \tag{10}$$

where the integral in the Fourier domain is carried out essentially over the bandwidth of the detector and where the signal cuts off at the upper frequency end. This is a family of integrals parametrised by all the signal parameters, in particular, t_a. We have suppressed other parameters to avoid clutter, since we now focus on t_a. But the $c(t_a)$ can be obtained for each t_a by just using the FFT algorithm. This saves enormous computational effort because now the number of operations reduces to order $N \log_2 N$ rather than N^2, where N is the number of samples in the data segment. Typically, for a 500 sec. data train sampled at 2 kHz, $N \sim 10^6$, which implies a saving of computational cost of more than 10^4!

In the phase parameter ϕ_a also, there is a symmetry — changing ϕ_a to say $\phi'_a = \phi_a + \phi_0$ involves just adding a constant phase to the signal and the waveform still remains within the family. This is the so called S^1 symmetry. The search over phase can be carried out by using just two templates say for $\phi_a = 0$ and $\phi_a = \pi/2$. If we call the correlations so obtained c_0 and $c_{\pi/2}$ respectively, where we have computed these correlations from the corresponding templates $h(f; \phi_a = 0)$ and $h(f; \phi_a = \pi/2)$, the $c(\phi_a)$ at arbitrary ϕ_a is then given by,

$$c(\phi_a) = c_0 \cos\phi_a + c_{\pi/2} \sin\phi_a\,, \tag{11}$$

where we have suppressed other parameters to avoid clutter. Moreover, the maximisation of $c(\phi_a)$, the surrogate statistic, over ϕ_a can be done analytically. Thus,

$$\max_{\phi_a} c(\phi_a) = \left(c_0^2 + c_{\pi/2}^2\right)^{1/2}\,. \tag{12}$$

Thus the kinematical parameters t_a and ϕ_a in the signal waveform are efficiently dealt with; the search over the masses, which are the dynamical parameters, now remains. There does not seem to be any efficient way, for example of using symmetries, to search over these parameters. Figure 5 shows the parameter space for $1M_\odot \leq m_1, m_2 \leq 30M_\odot$ in the parameters τ_0 and τ_3. Since the waveform is symmetric in m_1, m_2, one needs to only search the space $m_1 \leq m_2$. This gives roughly a triangular shape to the search region of

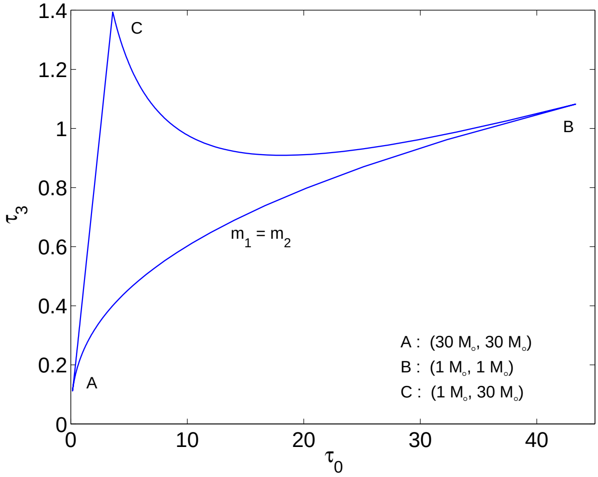


Figure 5. Parameter space in terms of the parameters τ_0, τ_3 for the mass range $1M_\odot \le m_1, m_2 \le 30M_\odot$ and $f_a = 40$ Hz.

the parameter space which is topologically equivalent to the triangle in (m_1, m_2) space. One now spans the parameter space densely with a bank of templates. The templates are arranged so that the maximum mismatch between a signal and a template never exceeds a small fixed amount. The usual number is taken to be 3% which corresponds to a maximum loss of 10% in the event rate of the signals. With this criterion, in the parameters τ_0, τ_3 the templates are approximately uniformly spaced. The idea is to *tile* the parameter space so that (i) there are no ‘holes’ and (ii) there is minimal overlap, so that the number of templates is reduced to a minimum which then in turn reduces the computational burden. The best such scheme happens to be hexagonal tiling as shown in Figure 6. One does this template placement elegantly by defining a metric (Balasubramanian, Sathyaprakash & Dhurandhar 1996; Owen 1996) over the parameter space. Then one finds that the metric coefficients are nearly constant when the parameters τ_0, τ_3 are used; the parameters in fact play the role of coordinates on a signal manifold, and the above statement can be reworded as saying that τ_0, τ_3 are like Cartesian coordinates, while m_1, m_2 are curvilinear. For this level of mismatch the number of templates required is $\sim 10^4$ for the noise PSD of the initial LIGO. If the signal is cut-off at a little less than 1 kHz, so that the sampling rate is 2 kHz, then a simple computation shows that the online search for these signals is little more than 3 GFlops.

One now takes the maximum of the matched filter output over all the parameters and compares this maximum with a threshold. The threshold is set by the noise statistics and the false alarm rate that one is prepared to tolerate. Clearly, the false alarm rate must be much less than the expected event rate. If the noise is Gaussian, the c maximised over ϕ_a has a Rayleigh distribution in the absence of the signal. Assuming a false alarm rate of 1 per year, gives a false alarm probability of $\sim 10^{-14}$ for one year observation period, which in turn sets the threshold at 8.2 (this is in units of the standard deviation of the

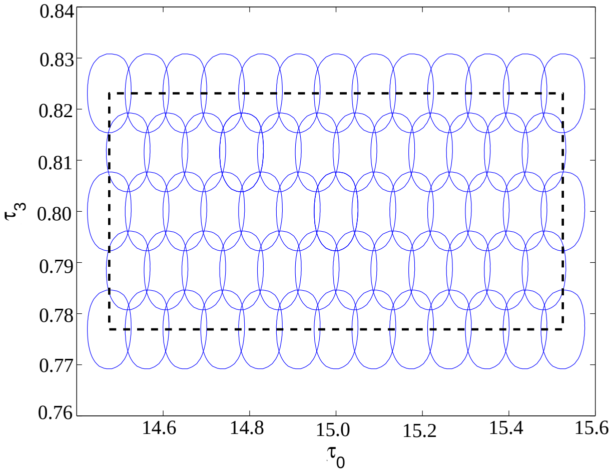


Figure 6. Hexagonal tiling of the parameter space.

Gaussian noise). Detection is announced if the c maximised over the parameters exceeds the threshold. However, in order to achieve good detection probability one must have c well over the threshold — thus $c > 8.9$ gives a detection probability better than about 95%.

The foregoing describes the general matched filtering paradigm. It clearly holds for a larger mass range or a larger bandwidth. If one lowers the lower limit of the band to 10 Hz as will be the case for advanced detectors and increases the mass range to begin from say $0.2M_\odot$, the online search requirement increases by a hundred times. Also if one includes spins, the waveform now must depend on 6 more parameters, namely, the spin vectors $\vec{S}_1, \vec{S}_2$, and the computational burden increases roughly by three orders of magnitude.

In order to deal with the rather large computational cost, hierarchical schemes have been proposed (Mohanty & Dhurandhar 1996; Sengupta, Dhurandhar & Lazzarini 2003) which can reduce this cost. The idea is to look for triggers with a low threshold with a coarse bank of templates, and then follow only the trigger events by a fine search and then use the high threshold as in the regular search described above. This method can reduce the cost considerably and theoretical factors of few tens in reducing the cost have been obtained in stationary Gaussian noise. These reduction factors of course will be significantly less in the presence of real detector noise. Recently Cannon *et al.* (2010) have shown that singular value decomposition can be used to reduce the computational cost. They have shown that for neutron star/neutron star inspiral where the parameter space is smaller, the computational gain can be almost an order of magnitude.

4.1.3 *Merger and ringdown*

For a black hole merger, one must solve Einstein's equations with complex initial and boundary conditions. Due to the nonlinearity of Einstein's equations, the problem is highly complex, and work had been going on for few decades before Pretorius (2005) made a breakthrough. This was followed by other successful numerical solutions (Campanelli *et al.*

2006; Baker *et al.* 2006). Clearly, the first such solutions corresponded to nonrotating black holes which now are not too difficult to obtain. The rather surprising fact was that the merger phase is a smooth continuation of the inspiral phase, contrary to what had been expected. There are results also for spinning black holes. Work is in progress to obtain numerical relativistic solutions spanning the entire parameter space for different mass ratios and spins. There are also interesting effects such as 'kicks' in the general case; the final black hole has a residual linear momentum.

The final ring-down phase consists of a superposition of quasi-normal modes (QNM) with their amplitude depending on the details of the perturbation. But each QNM is uniquely given by the black hole mass and the angular momentum parameter. The 'no hair' theorem for black holes in general relativity states that a black hole is completely characterised by its mass and angular momentum. The above mentioned property of QNMs is a consequence of this theorem. Thus observing QNMs would unambiguously show that the source is a black hole and also confirm the no hair theorem of general relativity.

The important point for data analysis is that the inspiral waveform can be continued into the merger phase and to the ring-down phase of the final black hole to obtain a single stitched waveform, thus yielding a higher SNR. The mass range can now go upto $100M_\odot$ and the distance by about a factor of 2 which would then correspondingly increase the event rate by about an order of magnitude. These searches are currently being performed by the Ligo Science Collaboration.

4.2 The all sky, all frequency search for GW from rotating neutron stars

We will consider the simple model of an isolated rapidly rotating neutron star and ignore spindown. We will show here, how the group theory and other algebraic methods can be used to elegantly formulate the problem by exploiting the symmetries in the physical model. In this endeavour, we will make use of the *stepping around the sky method* — a method proposed by Schutz more than twenty years ago (Schutz 1989), which gives an apt framework for this approach. There have been a host of methods proposed, notably the Hough transform, the stack and slide, and resampling methods (Schutz & Papa 1999; Patel *et al.* 2010) which reduce the computational cost over the straight-forward search over the sky direction, frequency, and spindown parameters. Although these methods significantly reduce the computational burden, it is not reduced to the point where the search can be performed with the current computer resources available in reasonable time. Therefore, it becomes necessary to explore novel approaches which address this problem.

Consider a *barycentric frame* in which the isolated neutron star is at rest or moving with uniform velocity. Ignoring spindowns the signal in this frame is assumed to be a pure sinusoid — monochromatic of constant frequency say f. The detector however, takes part in an accelerated motion — in general a superposition of simple harmonic motions of varying amplitudes and phases. Thus the signal in the detector is not a pure sinusoid but is modulated by Doppler effects — the Doppler correction depending on the direction to the source, relative to the motion of the detector. Since the detector moves in a complicated way relative to the barycentre, a complex Doppler profile is generated which depends on the direction to the source. If the direction to the source and the frequency are unknown, the Doppler profile is unknown and then one must face the problem of scanning over all directions in the sky and also over frequency. From astrophysical considerations usually the maximum frequency $f_{\rm max}$ is taken to be 1 kHz. The *stepping* method gives a direct way for obtaining the Fourier transform in the barycentric frame of the demodulated signals connecting two different directions say $\hat{\mathbf{n}}$ and $\hat{\mathbf{n}}'$.

The signal expected is so weak that one typically needs to integrate the signal for several months or a year before one can build up significant SNR. So if the observation time needed is $T \sim 10^7$ sec or more and if the maximum frequency $f_{\rm max}$ to be scanned is taken about a kHz, then the number of samples in a data train are $N \sim 2 f_{\rm max} T \sim 10^{10}$. Since the detector orbits the Sun in this time, the 'aperture' of the 'telescope' is the diameter D of the Earth's orbit; $D \sim 3 \times 10^8$ km while the minimum GW wavelength is $\lambda_{\rm GW} \sim 300$ km corresponding to a frequency of 1 kHz. Thus the resolution is $\Delta\theta = \lambda_{\rm GW}/D \sim 10^{-6}$ radian, a fraction of an arc second — the Fourier transform of such a signal spreads into a million Fourier bins, and consequently the signal is lost in the noise of the detector. One therefore needs to demodulate the signal first and then take its Fourier transform in order to collect the signal power in a single frequency bin. This means one needs to scan or demodulate over $N_{\rm patches} \simeq 4\pi/\Delta\theta^2 \sim 10^{13}$ directions or patches in the sky. So even this naive calculation gives the number of operations for the search to be $N_{\rm ops} \sim 3 N_{\rm patches} N \log_2 N \sim 10^{25}$ if one were to perform the FFT of the data train after demodulating in each direction. A machine with a speed of few teraflops would need several thousand years to perform the analysis! Moreover, this estimate excludes overheads, and ignores spindown parameters. Including these would increase the cost of the search by several orders of magnitude. Thus the search is highly computationally expensive, and novel and original ideas should be explored, if this search has to be brought within the capabilities of current resources or those envisaged in the near future. The approach outlined here is based on group theory and is one such attempt towards finding a solution to this problem.

Moreover, there exists also the possibility of using this approach in tandem with the previous approaches which have been aimed at reducing the computational cost. It is envisaged that a judicious combination of several approaches may go towards alleviating the computational burden.

4.2.1 *The formulation of the problem*

Let the motion of the detector be described in general by $\mathbf{R}(t)$ in the barycentric frame (X, Y, Z); for circular motion $\mathbf{R}(t) = R(\cos\Omega t\,\hat{\mathbf{X}} + \sin\Omega t\,\hat{\mathbf{Y}})$ — we take the detector motion in the (X, Y) plane — where $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ are unit vectors along the X and Y axes respectively, and Ω is the angular velocity of the detector in the barycentric frame. We will treat $\mathbf{R}(t)$ generally for now until later when we specialise to circular motion. The key defining equation which describes the transformation between barycentric time t and detector time t' is:

$$s'(t') = s(t)\,. \tag{13}$$

The detector time coordinate t', which is in fact a retarded or advanced time, is given by $t' = t - \mathbf{R}(t)\cdot\hat{\mathbf{n}}/c$ and is related to the barycentric time coordinate t. From our assumptions, the signal in the barycentric frame can be taken to be monochromatic. So after demodulation a Fourier transform is all that is necessary to extract the signal from the detector noise. It is in fact the matched filter!

We write the Doppler modulation in an abstract form in terms of an operator. The signal at the detector coming from the direction $\hat{\mathbf{n}}$ is related to the signal in the barycentric frame by the equation:

$$\mathbf{s}'(\hat{\mathbf{n}}) = M(\hat{\mathbf{n}})\,\mathbf{s}\,, \tag{14}$$

where $M(\hat{\mathbf{n}})$ is the modulation operator which is defined via equation (13). This operator has explicit representations in the time as well as in the Fourier domain (Dhurandhar & Krishnan 2011). Note that in the all sky, all frequency search we do not know $\hat{\mathbf{n}}$. Therefore we need to scan over the directions. A trial demodulation is performed for some general direction $\hat{\mathbf{n}}'$ given by $\hat{\mathbf{n}}' = (\sin\theta'\cos\phi', \sin\theta'\sin\phi', \cos\theta')$, which is not necessarily $\hat{\mathbf{n}}$. Thus we try the direction $\hat{\mathbf{n}}'$ and have a trial demodulated signal,

$$\mathbf{s}_{\rm trial}(\hat{\mathbf{n}}';\hat{\mathbf{n}}) = M^{-1}(\hat{\mathbf{n}}')\,\mathbf{s}'(\hat{\mathbf{n}})\,. \tag{15}$$

If $\hat{\mathbf{n}}' \neq \hat{\mathbf{n}}$, then the demodulation is incorrect and we must try again with a different $\hat{\mathbf{n}}'$ until we get to $\hat{\mathbf{n}}$ or atleast get close enough. If $\hat{\mathbf{n}}' \simeq \hat{\mathbf{n}}$, we must observe a peak in Fourier domain. Using these formulae we can now *step directly* to a direction $\hat{\mathbf{n}}'$ as follows:

$$\mathbf{s}_{\rm trial}(\hat{\mathbf{n}}';\hat{\mathbf{n}}) = Q(\hat{\mathbf{n}}',\hat{\mathbf{n}})\,\mathbf{s}\,, \tag{16}$$

where the *stepping* operator is defined by:

$$Q(\hat{\mathbf{n}}',\hat{\mathbf{n}}) = M^{-1}(\hat{\mathbf{n}}')\,M(\hat{\mathbf{n}})\,. \tag{17}$$

This was the approach suggested by Schutz, now expressed in our formulation, so that one may directly *step* from the direction $\hat{\mathbf{n}}$ to the direction $\hat{\mathbf{n}}'$ in the space of demodulated waveforms. This formulation was expected to enhance the efficiency of the search, for example, by using the sparseness of the matrices. The approach here builds upon this formulation. Apart from the sparseness of matrices, the idea is to use symmetries in the problem for stepping efficiently in the sky. The symmetry is made manifest via the language of group theory.

In order to get a group structure and go beyond the method advocated by Schutz, it is necessary to expand the scope of the direction vectors $\hat{\mathbf{n}}$ to the full three dimensional Euclidean space $\mathcal{R}^3$. It is clear that this is required because even a step in the sky namely, $\hat{\mathbf{n}}' - \hat{\mathbf{n}}$ will not be of unit length. Thus it is necessary as also convenient to ‘unwrap’ the space of directions, which is a projective space, to its universal covering space $\mathcal{R}^3$. We then define the operators $M(\mathbf{a})$, where $\mathbf{a}$ is an arbitrary vector in $\mathcal{R}^3$, and $\mathbf{a}$ is used in the ‘retarded time’ instead of $\hat{\mathbf{n}}$. We can then show that these operators now form a group, atleast approximately, well within the physical requirements (Dhurandhar & Krishnan 2011). It is important to note that these operators M act on functions, namely signals, and map them to other signals — the signals are Doppler shifted. Such groups are called transformation groups in the literature (Vilenkin 1988).

4.2.2 *Circular motion of the detector*

For concreteness, we give an example of circular motion of the detector. This is a very simplified case because in reality the detector partakes of a complicated superposition of simple harmonic motions which have complex set of phases. This simple case is taken to see how the group theory helps. The group now is reduced to Euclidean group in 2 dimensions, usually denoted by $E(2)$. We consider the motion as above and consider the situation when the motion consists of exactly one orbit, i.e. $0 \leq t \leq T$ and $\Omega T = 2\pi$. Then in the Fourier space, where $n = f/T$ and $n' = f'/T$, we look at the action of $M(\mathbf{a})$ on the complete orthonormal basis of the Hilbert space of square integrable functions over $[0, T]$, namely, the set of functions $e^{2\pi i n t/T}$. The natural scalar product on this Hilbert space for the two functions g_1 and g_2 is defined by:

$$(g_1, g_2) = \frac{1}{T} \int_0^T g_1(t)\, g_2^*(t)\, dt\,. \tag{18}$$

In this basis, the matrix representation for M (we have chosen $\mathbf{a} = \hat{\mathbf{n}}$ a unit vector) is readily given:

$$\begin{aligned} M(\hat{\mathbf{n}}; n', n) &= (M(\hat{\mathbf{n}})\, e^{2\pi i n \frac{t'}{T}},\, e^{2\pi i n' \frac{t'}{T}}) \\ &= (e^{2\pi i n \frac{t}{T}},\, e^{2\pi i n' \frac{t'}{T}}) \\ &= \frac{1}{T} \int_0^T dt'\, e^{2\pi i n \frac{t}{T} - 2\pi i n' \frac{t'}{T}}\,. \end{aligned} \tag{19}$$

where we have used the definition $M(\hat{\mathbf{n}}) s(t') = s(t)$. An explicit expression for $M(\hat{\mathbf{n}}; n', n)$ can be obtained for the direction $\hat{\mathbf{n}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. Writing $\psi = \Omega t'$ and $\beta = R\Omega/c$ we obtain:

$$\begin{aligned} M(\hat{\mathbf{n}}; n', n) &= \frac{1}{2\pi} \int_0^{2\pi} d\psi\, e^{i(n-n')\psi + in\beta\sin\theta\cos(\psi-\phi)} \\ &\equiv e^{i\chi(n-n')}\, J_{n-n'}\left(n\beta\sin\theta\right)\,, \end{aligned} \tag{20}$$

where $\chi = \phi + \pi/2$ is the translated azimuthal angle.

From the form of $M(\hat{\mathbf{n}}; n', n)$ it is evident that when applied to the data vector x_n, the search in χ can be performed by a fast Fourier transform; the stepping in the azimuthal parameter is done in an efficient way. If there are B samples of the χ parameter, then the search over χ for a given θ and frequency n' can be performed in order of $B \log_2 B$ number of operations. It may be further possible to reduce the number of operations by similar methods, but this example underlines the role of symmetry and the group theory in developing efficient data processing algorithms.

5. Concluding remarks

The era of gravitational wave astronomy has arrived. The initial detectors have not only reached their promised sensitivities but have surpassed them. The advanced detectors will start operating in few years time and the era of gravitational wave astronomy would then have truly begun. From the astrophysical knowledge that we possess as of now, one should expect a fair rate of gravitational wave events that one should be able to observe. An important recent development has been LIGO-Australia where LIGO is planning to build one of its detectors in Australia with partial funding from Australia. A detector far away and out of the plane of other detectors in US and Europe would greatly benefit the search of gravitational waves. India is also thinking of chipping in, so that India also has a stake in this exciting world project. Already, India has a 20 year old legacy in gravitational wave data analysis at IUCAA, Pune and wave form modelling at RRI, Bengaluru, and recently a three metre prototype detector at T.I.F.R., Mumbai has been funded. Apart from the groundbased detectors, there is also the prospect of the space-based ESA-NASA detector LISA which will bring in important astrophysical information at low frequencies complementing the ground-based detectors. The future looks bright for GW astronomy.

References

Abadie J., *et al.*, 2010, Phys. Rev. D, 82, 102001

Abbott B., *et al.*, 2008a, ApJL, 683, L45

Abbott B., *et al.*, 2008b, ApJ, 681, 1419

Abramovici A., *et al.*, 1992, Science, 256, 325

Allen B., Romano J., 1999, Phys. Rev. D, 59, 102001

Armstrong J.W., Estabrook F.B., Tinto M., 1999, ApJ, 527, 814

Baker J.G., Centrella J.M., Choi D.I., Koppitz M., van Meter J., 2006, PRL, 96, 111102

Balasubramanian R., Sathyaprakash B.S., Dhurandhar S.V., 1996, Phys. Rev. D, 53, 3033

Bender P.L. Hils D., 1997, Class. Quantum Grav., 14, 1439

Bender P., *et al.*, 1998, Laser Interferometer Space Antenna for the detection and observation of gravitational waves, Pre-Phase A Report

Blanchet L., Damour T., Esposito-Farése G., Iyer B. R., 2004, Phys. Rev. Lett., 93, 091101

Bradaschia C., *et al.*, 1990, Nucl. Instum. Methods Phys. Res. A, 289, 518

Brady P.R., Creighton T., Cutler C., Schutz B.F., 1998, Phys. Rev. D, 57, 2101

Cannon K., Chapman A., Hanna C., Keppel D., Searle A.C., Weinstein A.J., 2010, Phys. Rev. D, 82, 044025

Chandrasekhar S., Detweiler S., 1975, Proc. Roy. Soc. (London) A, 344, 441

Chandrasekhar S. Esposito F.P., 1970, ApJ, 160, 153

Campanelli M., Lousto C.O., Marronetti P., Zlochower Y., 2006, Phys. Rev. Lett., 96, 111101

Cutler C., Gholami I., Krishnan B., Phys. Rev. D, 72, 042004

Danzmann K., *et al.*, 1995, in First Edoardo Amaldi Conference on Gravitational Wave Experiments, Ed. E. Coccia, G. Pizzella, F. Ronga, World Scientific, Singapore

Dhurandhar S.V., Krishnan B., 2011, Mathematics Today, 26, 64

Dhurandhar S.V., Nayak K.R., Vinet J.-Y., 2002, Phys. Rev. D, 65, 102002

Dhurandhar S.V., Nayak K.R., Vinet J.-Y., 2010, Class. Quant. Grav., 27, 135013

Estabrook F.B., Tinto M., Armstrong J.W., 2000, Phys. Rev. D, 62, 042002

Gonzalez G., 2005, in the Proceedings of the IVth Mexican School of Astrophysics, July 18-25, 2005

Friedman J.L., Schutz B.F., 1978, ApJ, 222, 281

Hils D., Bender P.L., 2000, Ap.J, 537, 334

Hulse R.A., Taylor J.H., 1975, ApJ, 195, L51

Kuroda K., the LCGT Collaboration, 2006, Class. Quantum Grav., 23, S215

Ligo Science Collaboration and Virgo Science Collaboration, 2009, Nature, 460, 990

Mitra S., Dhurandhar S., Souradeep T., Lazzarini A., Mandic V., Bose S., Ballmer S., 2008, Phys. Rev. D, 77, 042002

Mohanty S., Dhurandhar S.V., 1996, Phys. Rev. D, 54, 7108

Nelemans G., Yungelson L.R., Portegies Zwart S.F., 2001, A&A, 375, 890

Owen B., 1996, Phys. Rev. D, 54, 2421

Patel P., Siemens X., Dupuis R., Betzwieser J., 2010, Phys. Rev. D, 81, 084032

Postnov K.A., Prokhorov M.E., 1998, Ap.J, 494, 674

Pretorius F., 2005, Phys. Rev. Lett., 95, 121101

Schutz B.F., 1989, in D. Blair, ed, The Detection of Gravitational Waves, Cambridge University Press, Cambridge, p. 406

Schutz B.F., Alessandra Papa M., 1999, in Gravitational waves and experimental gravity, Proceedings of Modiond, Editions Frontieres, Orsay

Sengupta A., Dhurandhar S.V., Lazzarini A., 2003, Phys. Rev. D, 67, 082004

Taylor J.H., 1994, Rev. Mod. Phys., 66, 711

Tsubono, K., 1995, in Gravitational Wave Experiments, Ed., Coccia, E. Pizzella, G., Ronga, F., World Scientific, Singapore, p. 112

Vilenkin N.J., 1988, Special functions and the theory of group representations, American Mathematical Society

Vishveshwara C.V., 1970, Phys. Rev. D, 1, 2870

Gravitational waves from perturbed stars

V. Ferrari[*]

Dipartimento di Fisica G. Marconi, Sapienza Università di Roma and
Sezione INFN ROMA1, Piazzale Aldo Moro 5, 00185 Roma, Italy

Received February 11, 2011; accepted February 28, 2011

Abstract. Non radial oscillations of neutron stars are associated with the emission of gravitational waves. The characteristic frequencies of these oscillations can be computed using the theory of stellar perturbations, and they are shown to carry detailed information on the internal structure of the emitting source. Moreover, they appear to be encoded in various radiative processes, as for instance, in the tail of the giant flares of Soft Gamma Repeaters. Thus, their determination is central to the theory of stellar perturbation. A viable approach to the problem consists in formulating this theory as a problem of resonant scattering of gravitational waves incident on the potential barrier generated by the spacetime curvature. This approach discloses some unexpected correspondences between the theory of stellar perturbations and the theory of quantum mechanics, and allows us to predict new relativistic effects.

Keywords : gravitational waves – black hole physics – stars: oscillations – stars: neutron – stars: rotation

1. Introduction

The theory of stellar perturbations is a very powerful tool to investigate the features of gravitational signals emitted when a star is set in non-radial oscillations by any external or internal cause. The characteristic frequencies at which waves are emitted are of great interest in these days, since gravitational wave detectors Virgo and LIGO are approaching the sensitivity needed to detect gravitational waves emitted by pulsating stars. These frequencies carry information on the internal structure of a star, and appear to be encoded in various radiative processes; thus, their study is a central problem in perturbation theory and in astrophysics. In this paper I will illustrate the theory of perturbations of a non-rotating star, describing in particular the formulation that Chandrasekhar and I developed, its motivation and outcomes. Furthermore, I will briefly describe how the theory has been applied to study the oscillation frequencies of neutron stars.

[*] e-mail: valeria.ferrari@roma1.infn.it

In order to frame the problem in an appropriate historical perspective, it is instructive to remind ourselves how the study of stellar perturbations was treated in the framework of Newtonian gravity. In that case, the adiabatic perturbations of a spherical star are described by a fourth-order, linear, differential system which couples the perturbation of the Newtonian potential to those of the stellar fluid. All perturbed quantities are Fourier-expanded and, after a suitable expansion in spherical harmonics, which allows for the separation of variables, the relevant equations are manipulated in such a way that the quantity which is singled out to describe the perturbed star is the Lagrangian displacement $\vec{\xi}$ experienced by a generic fluid element; indeed, the changes in density, pressure and gravitational potential induced by the perturbation, can all be expressed uniquely in terms of $\vec{\xi}$. The equations for $\vec{\xi}$ have to be solved by imposing appropriate boundary conditions at the centre of the star, where all physical quantities must be regular, and on its boundary, where the perturbation of the pressure must vanish. These conditions are satisfied only for a discrete set of *real* values of the frequency, $\{\omega_n\}$, which are the frequencies of the star's *normal modes*. Thus, the linearized version of the Poisson and of the hydro equations are reduced to a characteristic value problem for the frequency ω.

An adequate base for a rigorous treatment of stellar pulsations of a spherical star in general relativity was provided by K.S. Thorne and collaborators in a series of papers published in the late sixties–early seventies of the last century (Thorne & Campolattaro 1967, 1968; Campolattaro & Thorne 1970; Thorne 1969a,b; Ipser & Thorne 1973). The theory was developed in analogy with the Newtonian approach, and was later completed by Lindblom & Detweiler (1983), who brought the analytic framework to a form suitable for the numerical integration of the equations, thus allowing for the determination of the real and imaginary parts of the characteristic frequencies of the $\ell = 2$, *quasi-normal modes*. Indeed, a main difference between the Newtonian and the relativistic theory is that in general relativity the oscillations are damped by the emission of gravitational waves, and consequently the mode eigenfrequencies are complex. Higher order ($\ell > 2$) mode frequencies were subsequently computed by Cutler & Lindblom (1987).

In 1990, Professor Chandrasekhar and I started to work on stellar perturbations, and we decided to derive ab initio the equations of stellar perturbations following a different approach, having as a guide the theory of black hole perturbations rather than the Newtonian theory of stellar perturbations.

1.1 Black hole perturbations: wave equations and conservation laws

In 1957 T. Regge and J.A. Wheeler set the basis of the theory of black hole perturbations showing that, by expanding the metric perturbations of a Schwarzschild black hole in tensorial spherical harmonics, Einstein's equations can be separated (Regge & Wheeler 1957). Spherical harmonics belong to two different classes, depending on the way they transform under the parity transformation $\theta \to \pi - \theta$ and $\varphi \to \pi + \varphi$; those which transform like $(-1)^{(\ell+1)}$ are named *odd*, or *axial*, those that transform like $(-1)^{\ell}$ are named *even*, or *polar*. The perturbed equations decouple in two distinct sets belonging to the two parities. Regge & Wheeler further showed that by Fourier-transforming the time dependent variables, the equations describing the radial part of the *axial* perturbations can easily be reduced to a single Schroedinger-like equation, and 13 years later Frank Zerilli showed that this can also be done for the much more complicated set of *polar* equations (Zerilli 1970a,b); thus, the axial and polar perturbations of a Schwarzschild black hole are described by the wave equation

$$\frac{d^2 Z_\ell^\pm}{dr_*^2} + \left[\omega^2 - V_\ell^\pm(r)\right] Z_\ell^\pm = 0 \,, \tag{1}$$

$$V_\ell^-(r) = \frac{1}{r^3}\left(1 - \frac{2M}{r}\right)[\ell(\ell+1)r - 6M] \tag{2}$$

$$V_\ell^+(r) = \frac{2(r-2M)}{r^4(nr+3M)^2}[n^2(n+1)r^3 + 3Mn^2r^2 + 9M^2nr + 9M^3] \,. \tag{3}$$

where $r_* = r + 2M\log\left(\frac{r}{2M} - 1\right)$, $n = \frac{1}{2}(\ell+1)(\ell-2)$, and M is the black hole mass. The superscript $-$ and $+$ indicate, respectively, the Regge-Wheeler equation for the axial perturbations, and the Zerilli equation for the polar perturbations, and the corresponding potentials. The Regge-Wheeler potential for $\ell = 2$ is shown in Figure 1.

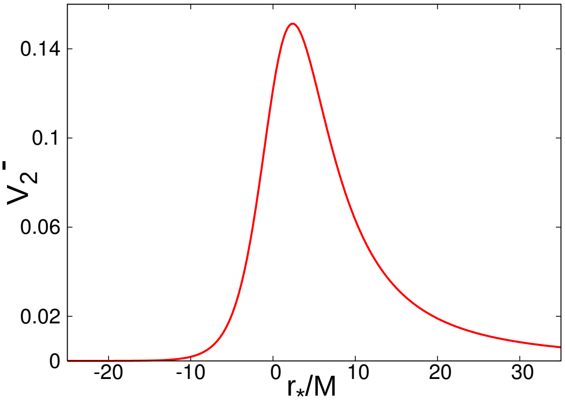


Figure 1. The potential barrier generated by the axial perturbations of a Schwarzschild black hole for $\ell = 2$. The potential for the polar perturbations has a similar form.

An alternative approach to studying black hole perturbations considers the perturbations of the Weyl and Maxwell scalars within the Newman-Penrose formalism. Using this approach in 1972, S. Teukolsky was able to decouple and separate the equations governing the perturbations of a Kerr black hole (Teukolsky 1972, 1973), and to reduce them to a single master equation for the radial part of the perturbation R_{lm}:

$$\Delta R_{lm,rr} + 2(s+1)(r-M)R_{lm,r} + V(\omega,r)R_{lm} = 0 \,, \qquad \Delta = r^2 - 2Mr + a^2. \tag{4}$$

Variable separation was achieved in terms of oblate spheroidal harmonics, and the potential $V(\omega, r)$ is given by

$$V(\omega,r) = \frac{1}{\Delta}\left[(r^2+a^2)^2\omega^2 - 4aMrm\omega + a^2m^2 + 2is(am(r-M) - M\omega(r^2-a^2))\right] \tag{5}$$

$$+\left[2is\omega r - a^2\omega^2 - A_{lm}\right] \,, \tag{6}$$

where a is the black hole angular momentum, A_{lm} is a separation constant, and s, the spin-weight parameter, takes the values $s = 0, \pm 1, \pm 2$, respectively for scalar, electromagnetic and gravitational perturbations. It may be noted that for the Schwarzschild perturbations, due to the background spherical symmetry, non-axisymmetric modes with a $e^{im\phi}$ dependence can be deduced from axisymmetric, $m = 0$ modes by suitable rotations of the polar axes. As a consequence, the potentials (2) and (3) do not depend on the harmonic index m. Conversely, since Kerr's background is axisymmetric, this degeneracy is removed and the potential (6) depends on m. Moreover, while the Schwarzschild potentials (2) and (3) are real and independent of frequency, the potential barrier of a Kerr black hole is complex, and depends on frequency.

The wave equations governing black hole perturbations show that the curvature generated by a black hole appears in the perturbed equations as a one-dimensional potential barrier; consequently, the response of a black hole to a generic perturbation can be studied by investigating the manner in which a gravitational wave incident on that barrier is transmitted, absorbed and reflected. Thus, the theory of black hole perturbations can be formulated as a scattering theory, and the methods traditionally applied in quantum mechanics to investigate the behaviour of physical systems described by a Schroedinger equation can be adapted and used to study the behaviour of perturbed black holes. For instance, it is known that in quantum mechanics, given a one-dimensional potential barrier associated with a Schroedinger equation, the singularities in the scattering cross-section correspond to complex eigenvalues of the energy and to the so-called quasi-stationary states. Since the perturbations of a Schwarzschild black hole are described by the Schroedinger-like equation (1) with the one-dimensional potential barriers (2) and (3), in which the energy is replaced by the frequency, the singularities in the scattering cross-section will provide the complex values of the black hole eigenfrequencies, and the corresponding eigenstates will be the black hole Quasi Nomal Modes (QNM). These modes satisfy the boundary conditions of a pure outgoing gravitational wave emerging at radial infinity ($r_* \to +\infty$), and a pure ingoing wave impinging at the black hole horizon ($r_* \to -\infty$). That these solutions should exist had been suggested by C.V. Vishveshwara in 1970 (Vishveshwara 1970), and the next year W.H. Press confirmed this idea by numerically integrating the axial wave equation (1), and by showing that an arbitrary initial perturbation ends in a ringing tail, which indicates that black holes possess some proper modes of vibration (Press 1971). However, it was only in 1975 that S. Chandrasekhar and S. Detweiler computed the complex eigenfrequencies of the quasi-normal modes of a Schwarzschild black hole, by integrating the Riccati equation associated with the axial equation (1) (Chandrasekhar & Detweiler 1975). In addition, they also showed that the transmission and the reflection coefficients associated respectively with the polar and with the axial potential barriers are equal. As a consequence, the polar and the axial perturbations are isospectral, i.e. the polar and axial QNM eigenfrequencies are equal. This equality can be explained in terms of a transformation theory which clarifies the relations that exist between potential barriers admitting the same reflection and absorption coefficients. This is an example of how the scattering approach has been effective not only to determine the QNM frequencies, but also to investigate the inner relations existing among the axial and polar potential barriers and to gain a deeper insight in the mathematical theory of black holes, which was illustrated by S. Chandrasekhar in his book on the subject (Chandrasekhar 1984). Following this approach, a variety of methods developed in the context of quantum mechanics have been used to determine the QNM spectra of rotating and non-rotating black holes, such as the WKB and higher order WKB method, phase-integral methods and the theory of Regge poles, just to mention some of them (Schutz & Will 1985; Ferrari & Mashhoon 1984a,b; Andersson, Araujo & Schutz 1993a,b,c; Andersson 1994; Andersson & Thylwe 1994).

1.2 A conservation law for black hole perturbations and its generalization to perturbed stars

In quantum mechanics the equation

$$|R|^2 + |T|^2 = 1, \tag{7}$$

where R and T are the reflection and transmission coefficients associated with a potential barrier, expresses the symmetry and unitarity of the scattering matrix; it says that, if a wave of unitary amplitude is incident on one side of the potential barrier, it gives rise to a reflected and a transmitted wave such that the sum of the square of their amplitudes is still one. Therefore, Eq. (7) is an energy conservation law for the scattering problem described by the Schroedinger equation with a potential barrier. This conservation law is a consequence of the constancy of the Wronskian of pairs of independent solutions of the Schroedinger equation. Similarly, the constancy of the Wronskian of two independent solutions of the black hole wave equations allows us to write the same relation between the reflection and transmission coefficients associated with the potential barrier, and therefore it shows that such an energy conservation law also governs the scattering of gravitational waves by a perturbed black hole. It should be stressed that such energy conservation law *does not* exist in the framework of the exact non-linear theory; however, it can be derived in perturbation theory both for Schwarzschild, Kerr and Reissner-Nordstrom black holes.

This possibility led Chandrasekhar to the following consideration. Since in general relativity, any distribution of matter (or more generally energy of any sort) induces a curvature of the spacetime – a potential well – instead of picturing the non-radial oscillations of a star as caused by some unspecified external perturbation, we can picture them as excited by incident gravitational radiation. Viewed in this manner, the reflection and absorption of incident gravitational waves by black holes and the non-radial oscillations of stars, become different aspects of the same basic theory. However, this idea needed to be substantiated by facts, and our starting point was to show that also for perturbed stars it is possible to write an energy conservation law in terms of Wronskians of independent solutions of the perturbation equations of a spherical star. This is easy if we consider the axial perturbations of a non-rotating star, because in that case, as we shall show in Section 2, the perturbed equations can be reduced to a wave equation with a one dimensional potential barrier as for black holes. However, to derive the conservation law for the polar perturbations was not easy, because the corresponding equations are a fourth order linear differential system, in which the perturbed metric functions couple to the fluid perturbations, and it was not clear how to define the conserved current. Anyway, working hard on the equations, we were able to derive a vector $\vec{\mathbf{E}}$ in terms of metric and fluid perturbations, which satisfies the following equation (Chandrasekhar & Ferrari 1990a):

$$\frac{\partial}{\partial x^{\alpha}} E^{\alpha} = 0, \qquad \alpha = (x^2 = r, x^3 = \vartheta). \tag{8}$$

(It is worth reminding ourselves that, due to the spherical symmetry, it is not restrictive to consider axisymmetric perturbations $m = 0$). The vanishing of the ordinary divergence implies that, by Gauss's theorem, the flux of $\vec{\mathbf{E}}$ across a closed surface surrounding the star is a constant. When the fluid variables are switched off, this conservation law reduces to that derived for a Schwarzschild black hole, and therefore we thought we were on the right track. However, there was still a question to answer: are we entitled to say that the vector $\vec{\mathbf{E}}$ actually represents the flux of gravitational energy which develops through the stars and propagates outside? If so, Eq. (8) should reduce to the second variation of the time component of the well known equation

$$\frac{\partial}{\partial x^{\nu}} [\sqrt{-g}(T^{\mu\nu} + t^{\mu\nu})] = 0. \tag{9}$$

where $t^{\mu\nu}$ is the stress-energy pseudotensor of the gravitational field. The problem is that $t^{\mu\nu}$ is not uniquely defined; indeed Eq. (9) shows that it is defined up to a divergenceless term. A possible definition is that given by Landau & Lifschitz (1975), which has the advantage of being symmetric. However, the second variation of the time component of Eq. (9) assuming $t^{\mu\nu} = t^{\mu\nu}_{LL}$, does not give the divergenceless equation satisfied by our vector $\vec{\mathbf{E}}$, neither for the Einstein-Maxwell case, nor in the case of a star. Then, Raphael Sorkin suggested that the pseudo-tensor whose second variation should reproduce our conserved current is the Einstein pseudo-tensor, because its second variation retains its divergence-free property, provided only the equations governing the static spacetime and its linear perturbations are satisfied.[1] This property is a consequence of the Einstein pseudo-tensor being a Noether operator for the gravitational field; the Landau-Lifshitz pseudotensor failed to reproduce the conserved current because it does not satisfy the foregoing requirements. In addition, Sorkin pointed out that the contribution of the source should be introduced not by adding the second variation of the source stress-energy tensor $T^{\mu\nu}$, as one might naively have thought, but through a suitably defined Noether operator, whose form he derived for an electromagnetic field (Sorkin 1991). Though this operator does not coincide with $T^{\mu\nu}$, it gives the same conserved quantities. Thus, the flux integral which we had obtained, I would say, by brute force, working directly on the perturbed hydrodynamical equations, could be obtained from a suitable expansion of the Einstein pseudo-tensor showing that, as for black holes, energy conservation also governs phenomena involving gravitational waves emitted by perturbed stars (Chandrasekhar & Ferrari 1991a). We therefore decided to derive ab initio the equations of perturbations of a spherical star in the same gauge used when studying the perturbations of a Schwarzschild black hole, and to study the problem as a scattering problem. In the next sections I shall briefly illustrate the main results we obtained by using this approach (Chandrasekhar & Ferrari 1990b, 1991b,c, 1992; Chandrasekhar, Ferrari & Winston 1991).

[1] It should be mentioned that the first variation of the Einstein pseudo-tensor vanishes identically.

2. Perturbations of a non-rotating star

As for a Schwarzschild black hole, when the equations describing the perturbations of a spherical star are perturbed and expanded in spherical tensor harmonics they decouple in two distinct sets, one for the polar and one for the axial perturbations. The axial equations do not involve fluid motion except for a stationary rotation, while the polar equations couple fluid and metric perturbations. The axial equations are therefore much simpler and we showed that, after separating the variables and Fourier-expanding the perturbed functions, they can be combined as in the Schwarzschild case, and reduced to a single Schroedinger-like equation with a one-dimensional potential barrier (Chandrasekhar & Ferrari 1990b, 1991c):

$$\frac{d^2 Z_\ell^-}{dr_*^2} + [\omega^2 - V_\ell^-(r)] Z_\ell^- = 0, \tag{10}$$

where

$$r_* = \int_0^r e^{-\nu+\mu_2} dr, \tag{11}$$

and

$$V_\ell^-(r) = \frac{e^{2\nu}}{r^3}[\ell(\ell+1)r + r^3(\epsilon - p) - 6m(r)]. \tag{12}$$

The functions $\nu(r)$ and $\mu_2(r)$, which appear in the definition of the radial variable r_*, are two metric functions which are found by solving the equations of stellar structure for an assigned equation of state (EOS). $\epsilon(r)$ and $p(r)$ are the energy density and the pressure in the unperturbed star; outside the star they vanish and Eq. (12) reduces to the Regge-Wheeler potential (2) of a Schwarzschild black hole. Thus, the axial potential barrier generated by the curvature of the star depends on how the energy-density and the pressure are distributed inside the star in the equilibrium configuration, and therefore it depends on the equation of state of matter inside the star. As an example, in Figure 2 we show the $\ell = 2$ potential barrier for an ideal, constant density star with $R/M = 2.8$ (left panel) and $R/M = 2.4$ (right panel). If we compare the potential shown in Figure 2 with the Regge-Wheeler potential of a Schwarzschild black hole shown in Figure 1, we notice an important difference. The Schwarzschild potential vanishes at the black hole horizon, and has a maximum at $r_{max} \sim 3M$, whereas the potential barrier of a perturbed star tends to infinity at $r = 0$. Thus, for a Schwarzschild black hole waves are scattered by a one-dimensional potential barrier, whereas in the case of a star they are scattered by a central potential.

Since the axial perturbations do not excite any motion in the fluid, for a long time they have been considered as trivial. But this is not true if we adopt the scattering approach: the absence of fluid motion simply means that the incident axial wave experiences a potential scattering, and this scattering can, in some extreme conditions, be resonant. Indeed, if we look for solutions that are regular at $r = 0$ and behave as pure outgoing waves at infinity, we find modes which do not exist in Newtonian theory; if the star is extremely compact, the potential in the interior is a well, and if this well is deep enough there can exist one or more more slowly damped quasi-normal modes, or s-modes (Chandrasekhar & Ferrari 1991c). For example, if the mass of the star is, say, $M = 1.4\ M_\odot$ and $R/M = 2.4$, i.e. the stellar compactness is $M/R = 0.42$, as shown in Figure 2 the well inside the star is deep enough to allow one quasi-normal mode. The number of s-modes increases with the depth of the well,

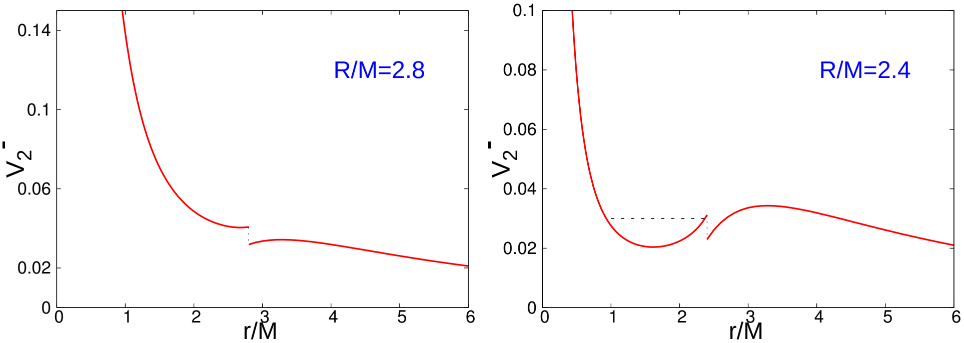


Figure 2. The $\ell = 2$ potential barrier for a constant density star of mass $M = 1.4\,M_\odot$ and $R/M = 2.8$ (left panel) and $R/M = 2.4$ (right panel). The horizontal, dashed line in the right panel corresponds to the value of frequency ω^2, which corresponds to a solution regular at $r = 0$ and behaving as a pure outgoing wave at radial infinity, i.e. to a quasi-normal mode.

which corresponds to a larger stellar compactness. However, it should be mentioned that neutron stars are not expected to have such a large compactness, unless one invokes some exotic equation of state. The *s*-modes are also named *trapped modes* because, due to the slow damping, they are effectively trapped by the potential barrier, and not much radiation can leak out of the star when these modes are excited. Axial modes on a second branch are named *w*-modes and are highly damped (Kokkotas 1994). The *w*-mode frequency also depends on the stellar compactness, as we shall show in Section 3.1. Therefore they carry interesting information on the internal structure of the star.

It should be stressed that the axial modes do not have a Newtonian counterpart.

Our approach to the polar perturbations, which couple the perturbations of the gravitational field to those of the metric, is different from the Newtonian approach briefly described in the introduction. Rather than focusing on the fluid behaviour, we focus on the variables which describe the spacetime perturbations, assuming that, as in the case of black holes, they are excited by the incidence of polar gravitational waves belonging to a particular angular harmonic. A careful scrutiny of the structure of the polar equations shows that it is possible to decouple the equations describing the metric from those describing the fluid perturbations. This decoupling allows us to solve the equations for the spacetime perturbations with no reference to the motion that can be induced in the fluid, and this is possible in general. Once the solution for the metric perturbations is found, the fluid variables can be determined in terms of them by simple algebraic relations without further ado (Chandrasekhar & Ferrari 1990b). The final set of equations to solve is described in Section 2.1.

2.1 The equations for the polar perturbations

Assuming that the metric which describes the unperturbed star has the form

$$ds^2 = e^{2\nu}(dt)^2 - e^{2\psi}d\varphi - e^{2\mu_2}(dr)^2 - e^{2\mu_3}(d\theta)^2, \tag{13}$$

the functions that describe the polar perturbations, expanded in spherical tensor harmonics and Fourier-expanded are

$$\begin{array}{rl} \delta\nu = N_\ell(r)P_\ell(\cos\theta)e^{i\omega t} & \delta\mu_2 = L_\ell(r)P_\ell(\cos\theta)e^{i\omega t} \\ \delta\mu_3 = [T_\ell(r)P_\ell + V_\ell(r)P_{\ell,\theta,\theta}]e^{i\omega t} & \delta\psi = [T_\ell(r)P_\ell + V_\ell(r)P_{\ell,\theta}\cot\theta]e^{i\omega t}, \\ \delta p = \Pi_\ell(r)P_\ell(\cos\theta)e^{i\omega t} & 2(\epsilon + p)e^{\nu+\mu_2}\xi_r(r,\theta)e^{i\omega t} = U_\ell(r)P_\ell e^{i\omega t} \\ \delta\epsilon = E_\ell(r)P_\ell(\cos\theta)e^{i\omega t} & 2(\epsilon + p)e^{\nu+\mu_3}\xi_\theta(r,\theta)e^{i\omega t} = W_\ell(r)P_{\ell,\theta}e^{i\omega t}, \end{array} \tag{14}$$

where $P_\ell(\cos\theta)$ are Legendre's polynomials, ω is the frequency, δp and $\delta\epsilon$ are perturbations of the pressure and of the energy density, and ξ_r, ξ_θ are the relevant components of the Lagrangian displacement of the generic fluid element. Note that (N, L, T, V) and (Π, E, U, W) are, respectively, the radial part of the metric and of the fluid perturbations. After separating the variables the relevant Einstein's equations for the metric functions become

$$\begin{cases} X_{\ell,r,r} + \left(\frac{2}{r} + \nu_{,r} - \mu_{2,r}\right)X_{\ell,r} + \frac{n}{r^2}e^{2\mu_2}(N_\ell + L_\ell) + \omega^2 e^{2(\mu_2-\nu)}X_\ell = 0, \\ (r^2G_\ell)_{,r} = n\nu_{,r}(N_\ell - L_\ell) + \frac{n}{r}(e^{2\mu_2} - 1)(N_\ell + L_\ell) + r(\nu_{,r} - \mu_{2,r})X_{\ell,r} + \omega^2 e^{2(\mu_2-\nu)}rX_\ell, \\ -\nu_{,r}N_{\ell,r} = -G_\ell + \nu_{,r}[X_{\ell,r} + \nu_{,r}(N_\ell - L_\ell)] + \frac{1}{r^2}(e^{2\mu_2} - 1)(N_\ell - rX_{\ell,r} - r^2G_\ell) \\ -e^{2\mu_2}(\epsilon + p)N_\ell + \frac{1}{2}\omega^2 e^{2(\mu_2-\nu)}\left\{N_\ell + L_\ell + \frac{r^2}{n}G_\ell + \frac{1}{n}[rX_{\ell,r} + (2n+1)X_\ell]\right\}, \\ L_{\ell,r}(1 - D) + L_\ell\left[\left(\frac{2}{r} - \nu_{,r}\right) - \left(\frac{1}{r} + \nu_{,r}\right)D\right] + X_{\ell,r} + X_\ell\left(\frac{1}{r} - \nu_{,r}\right) + DN_{\ell,r} + \\ N_\ell\left(D\nu_{,r} - \frac{D}{r} - F\right) + \left(\frac{1}{r} + E\nu_{,r}\right)\left[N_\ell - L_\ell + \frac{r^2}{n}G_\ell + \frac{1}{n}\left(rX_{\ell,r} + X_\ell\right)\right] = 0, \end{cases} \tag{15}$$

where

$$\begin{cases} A = \frac{1}{2}\omega^2 e^{-2\nu}, \qquad Q = \frac{(\epsilon+p)}{\gamma p}, \\ \gamma = \frac{(\epsilon+p)}{p}\left(\frac{\partial p}{\partial\epsilon}\right)_{entropy=const}, \qquad B = \frac{e^{-2\mu_2}\nu_{,r}}{2(\epsilon+p)}(\epsilon_{,r} - Qp_{,r}), \\ D = 1 - \frac{A}{2(A+B)} = 1 - \frac{\omega^2 e^{-2\nu}(\epsilon+p)}{\omega^2 e^{-2\nu}(\epsilon+p) + e^{-2\mu_2}\nu_{,r}(\epsilon_{,r} - Qp_{,r})}, \\ E = D(Q-1) - Q, \\ F = \frac{\epsilon_{,r} - Qp_{,r}}{2(A+B)} = \frac{2[\epsilon_{,r} - Qp_{,r}](\epsilon+p)}{2\omega^2 e^{-2\nu}(\epsilon+p) + e^{-2\mu_2}\nu_{,r}(\epsilon_{,r} - Qp_{,r})}, \end{cases} \tag{16}$$

and V_ℓ and T_ℓ have been replaced by X_ℓ and G_ℓ defined as

$$\begin{cases} X_\ell = nV_\ell \\ G_\ell = \nu_{,r}[\frac{n+1}{n}X_\ell - T_\ell]_{,r} + \frac{1}{r^2}(e^{2\mu_2} - 1)[n(N_\ell + T_\ell) + N_\ell] \\ +\frac{\nu_{,r}}{r}(N_\ell + L_\ell) - e^{2\mu_2}(\epsilon + p)N_\ell + \frac{1}{2}\omega^2 e^{2(\mu_2-\nu)}[L_\ell - T_\ell + \frac{2n+1}{n}X_\ell]. \end{cases} \tag{17}$$

These equations are valid in general, also for non-barotropic equations of state. It should be stressed that Eqs. (15) govern the variables (X, G, N, L) which are *metric perturbations*; however, since the motion of the fluid is excited by the polar perturbation, we may want to determine the fluid variables, (Π, E, U, W); they can be obtained in terms of the metric functions using the following algebraic relations

$$W_\ell = T_\ell - V_\ell + L_\ell,$$

$$\Pi_\ell = -\frac{1}{2}\omega^2 e^{-2\nu}W_\ell - (\epsilon + p)N_\ell, \qquad E_\ell = Q\Pi_\ell + \frac{e^{-2\mu_2}}{2(\epsilon + p)}(\epsilon_{,r} - Qp_{,r})U_\ell,$$

$$U_\ell = \frac{[(\omega^2 e^{-2\nu}W_\ell)_{,r} + (Q+1)\nu_{,r}(\omega^2 e^{-2\nu}W_\ell) + 2(\epsilon_{,r} - Qp_{,r})N_\ell](\epsilon + p)}{[\omega^2 e^{-2\nu}(\epsilon + p) + e^{-2\mu_2}\nu_{,r}(\epsilon_{,r} - Qp_{,r})]}.$$

Outside the star the fluid variables vanish, and the polar equations reduce to the wave equation (1) with the Zerilli potential (3).

As discussed above, for the axial perturbations, the frequencies of the quasi-normal modes were found by solving a problem of scattering by a central potential; for the polar perturbations it is not so simple, because a Schroedinger equation holds only in the exterior of the star, whereas a higher order system must be solved in the interior. It is still a scattering problem, but of a more complex nature since the incident polar gravitational waves, which excite the perturbations, drive the fluid pulsations, which in turn emit the scattered component of the wave. This approach was very fruitful in many respects. First of all, given the equilibrium configuration for any assigned equation of state, it was very easy to evaluate the QNM-frequency by integrating the equations for the metric perturbations inside and outside the star, looking for the solutions which, being regular at $r = 0$, behave as pure outgoing waves at infinity. Furthermore, we generalized the perturbed equations to slowly rotating stars, and derived the equations which describe how the axial perturbations couple to the polar (Chandrasekhar & Ferrari 1991b).

2.2 Perturbed equations for a slowly rotating star

Very briefly, the coupling mechanism is the following. Let Z_ℓ^{0-} be the axial radial function, solution of Eq. (10), which describes the perturbation of a non-rotating star; let $\epsilon(\Omega)Z_\ell^{1-}$ be the perturbation to first order in the star angular velocity Ω. The axial perturbation is the sum of the two:

$$Z_\ell^- = Z_\ell^{0-} + \epsilon(\Omega) Z_\ell^{1-}.$$

As Z_ℓ^{0-}, the function Z_ℓ^{1-} satisfies the wave equation (10) with the same potential (12), but with a forcing term:

$$\sum_{\ell=2}^{\infty} \left\{ \frac{d^2 Z_\ell^{-1}}{dr_*^2} + \left[\omega^2 - V_\ell^-\right] Z_\ell^{-1} \right\} C_{\ell+2}^{-\frac{3}{2}}(\mu) = r e^{2\nu - 2\mu_2} (1-\mu^2)^2 \sum_{\ell=2}^{\infty} S_\ell^0(r,\mu), \tag{18}$$

where $\mu = \cos\vartheta$ and $C_{\ell+2}^{-\frac{3}{2}}(\mu)$ are the Gegenbauer polynomials. The source term S_ℓ^0 is

$$S_\ell^0 = \varpi_{,r}[(2W_\ell^0 + N_\ell^0 + 5L_\ell^0 + 2nV_\ell^0 P_{\ell,\mu} + 2\mu V_\ell^0 P_{\ell,\mu,\mu}] + 2\varpi W_\ell^0 (Q-1)\nu_{,r} P_{\ell,\mu};$$

it is a combination of the functions which describe the *polar* perturbations on the *non-rotating* star, found by solving the equations given in Section 2.1. It should be stressed that the coupling function ϖ is the function responsible for the Lense-Thirring effect. Thus a rotating star exerts a dragging not only of the bodies, but also of the waves, and consequently an incoming polar gravitational wave can convert, through the fluid oscillations it excites, some of its energy into outgoing axial waves. This is a purely relativistic effect, and it is due to the dragging of inertial frames. It is interesting to note that the coupling between axial and polar perturbations satisfies rules that are similar to those known in the theory of atomic transitions: a Laporte rule and a selection rule, according to which the polar modes belonging to *even* ℓ can couple only with the axial modes belonging to *odd* ℓ, and conversely, and that it must be

$$l = m + 1, \qquad \text{or} \qquad l = m - 1.$$

Furthermore, the coupling satisfies a propensity rule (Fano 1985): the transition $\ell \to \ell + 1$ is strongly favoured over the transition $\ell \to \ell - 1$.

At the time Chandrasekhar and I wrote the series of papers on stellar perturbations, there was a growing interest in the subject, also motivated by the fact that the construction of ground based interferometric detectors, LIGO in the US and Virgo in Italy, had just started. Many studies addressed the problem of finding the frequencies of the QNMs, to establish what kind of information they carry on the internal structure of the emitting source. The collective effort developed using essentially two different perturbative approaches: one in the frequency domain, as for the theory developed by Thorne and collaborators or by Chandrasekhar and myself, another in the time domain. The time domain approach basically consists in separating the equations of stellar perturbations as usual in terms of spherical harmonics, and in solving the resulting equations in terms of two independent variables, radial distance and time. The equations are excited using some numerical input, like for instance a Gaussian impulse, and then the QNM frequencies are found by looking at the peaks of the Fourier transform of the signal obtained by evolving the time-dependent equations numerically. A disadvantage of this evolution scheme is that one cannot get the complete spectrum of the QNMs either for a star or for a black hole. The reason is that, although any perturbation is the sum of the harmonics involved, in practice only a few of them can be clearly identified; thus, to find some more modes one has to proceed empirically by changing the initial conditions. However, the evolution of the time dependent equations is, still today, the only viable perturbative method to find the QNM frequencies and waveforms emitted by rapidly rotating relativistic stars. To describe the problems which emerge when dealing with the perturbations of a rapidly rotating star is beyond the scope of this paper; I will discuss some related issues in the concluding remarks.

3. Neutron star oscillations

We shall now show how the theory of perturbations of non-rotating stars can be applied to gain some insight into the internal structure of the emitting source. Different classes of modes probe different aspects of the physics of neutron stars. For instance the fundamental mode (f-mode), which has been shown to be the most efficient GW emitter by most numerical simulations, depends on the average density, the pressure modes (p-modes) probe the sound speed throughout the star, the gravity modes (g-modes) are associated with thermal/composition gradients and the w-modes are spacetime oscillations. Furthermore, crustal modes, superfluid modes, magnetic field modes can, if present, add to the complexity of stellar dynamics. The sensitivity of ground based gravitational detectors has steadily improved over the years in a broad frequency window; the advanced version of LIGO and Virgo, and especially third generation detectors like ET, promise to be powerful instruments to detect signals emitted by oscillating stars. The frequencies of quasi normal modes are encoded in these signals; therefore, as the Sun oscillation frequencies are used in helioseismology to probe its internal structure, we hope that in the future it will be possible to use gravitational waves to probe the physics of neutron stars. One of the issues which is interesting to address concerns the equation of state of matter in a neutron star core, which is actually unknown. This problem is of particular interest, because the energies prevailing in the inner core of a neutron star are much larger than those accessible to high energy experiments on Earth. In the core, densities typically exceed the equilibrium density of nuclear matter, $\rho_0 = 2.67 \times 10^{14}$ g/cm^3; at these densities neutrons cannot be treated as non-interacting particles, and the main contribution to pressure, which comes from neutrons, cannot be derived only from Pauli's exclusion principle. Indeed, with only this contribution, we would find that the maximum mass of a neutron star is 0.7 $M_\odot$ which, as observations show, is far too low. This clearly shows that NS equilibrium requires a pressure other than the degeneracy pressure, the origin of which has to be traced back to the nature of hadronic interactions. Due to the complexity of the fundamental theory of strong interactions, the equations of state appropriate to describe a NS core have been obtained within models, which are constrained, as much as possible, by empirical data. They are derived within two main, different approaches: the nonrelativistic nuclear many-body theory, NMBT, and the relativistic mean field theory, RMFT. In NMBT, nuclear matter is viewed as a collection of pointlike protons and neutrons, whose dynamics is described by the nonrelativistic Hamiltonian:

$$H = \sum_i \frac{p_i^2}{2m} + \sum_{j>i} v_{ij} + \sum_{k>j>i} V_{ijk} \,, \tag{19}$$

where m and p_i denote the nucleon mass and momentum, respectively, whereas v_{ij} and V_{ijk} describe two- and three-nucleon interactions. These potentials are obtained from fits of existing scattering data (Wiringa, Stoks & Schiavilla 1995), (Pudliner *et al.* 1995). The ground state energy is calculated using either variational techniques or G-matrix perturbation theory. The RMFT is based on the formalism of relativistic quantum field theory, nucleons are described as Dirac particles interacting through meson exchange. In the simplest implementation of this approach the dynamics is modeled in terms of a scalar and a vector field (Walecka 1974). The equations of motion are solved in the mean field approximation, i.e. replacing the meson fields with their vacuum expectation values, and the parameters of the Lagrangian density, i.e. the meson masses and coupling constants, can be determined by fitting the empirical properties of nuclear matter, i.e. binding energy, equilibrium density and compressibility. Both NMBT and RMFT can be generalized to take into account the appearance of hyperons. In the following we shall consider some EOS representative of the two approaches, which have been used in the literature.

It should be stressed that different ways of modeling hadronic interactions affect the pulsation properties of a star, which we are going to discuss.

3.1 The axial and polar *w*-modes

As shown in Section 2, the axial perturbations are described by a Schroedinger-like equation with a central potential barrier which depends on the energy and pressure distribution in the unperturbed star, i.e. on the equation of state. The slowly damped modes are not expected to be associated with significant gravitational wave emission, because they are effectively trapped by the potential barrier; in addition they appear if the star has a compactness close to the static Schwarzschild limit, which establishes that constant density star solutions of Einstein's equations exists only for $M/R < 4/9 \simeq 0.44$. Conversely, the w-modes, which are highly damped, exist also for stars with ordinary compactness. They have been shown to exist also for the polar perturbations and in that case they are coupled to negligible fluid

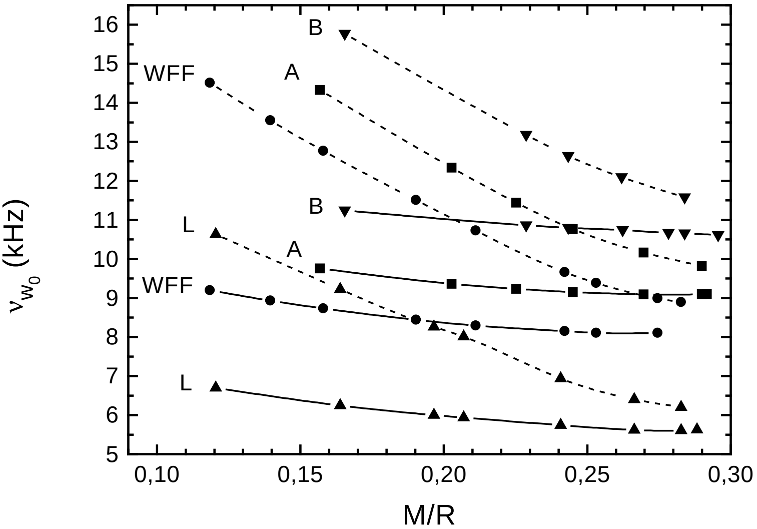


Figure 3. The frequency of the first polar (dashed line) and axial (continuous line) w-modes are plotted as a function of the star compactness for the EOSs A, B, WFF, L.

motion. In Figure 3 we compare the frequencies of the lower axial *w*-modes computed in (Benhar, Berti & Ferrari 1999) with those of the lower polar *w*-modes computed in (Andersson & Kokkotas 1988) for several EOSs. The main features of different EOS are, very briefly, the following. EOS A (Pandharipande 1971a) is pure neutron matter, with dynamics governed by a nonrelativistic Hamiltonian containing a semi-phenomenological interaction potential. It is obtained using NMBT. EOS B (Pandharipande 1971b) is a generalization of EOS A, including protons, electrons and muons in β-equilibrium, as well as heavier baryons (hyperons and nucleon resonances) at sufficiently high densities (NMBT). EOS WFF (Wiringa, Fiks & Fabrocini 1988) is a mixture of neutrons, protons, electrons and muons in β-equilibrium. The Hamiltonian includes two- and three-body interaction potentials. The ground state energy is computed using a more sophisticated and accurate many-body technique (NMBT). In EOS L (Pandharipande & Smith 1975) neutrons interact through exchange of mesons (ω, ρ, σ). The exchange of heavy particles (ω, ρ) is described in terms of nonrelativistic potentials, the effect of σ-meson is described using relativistic field theory and the mean-field approximation.

From Figure 3 we see that for each selected EOS the frequency of the *polar* w-modes is a rather steeply decreasing function of the stellar compactness M/R, whereas for the *axial* modes the dependence of ν_{w_0} on the compactness is weak, and ranges within intervals that are separated for each EOS. This means that if an axial gravitational wave emitted by a star at a given frequency could be detected, we would be able to identify the equation of state prevailing in the star's interior even without knowing its mass and radius. Hence, the detection of axial gravitational waves would allow us to constrain the EOS models, with regard to both the composition of neutron star matter and the description of the hadronic interactions. Until very recently, the common belief was that *w*- modes are unlikely to be excited in astrophysical processes. However, it has been shown that they are excited in the collapse of a neutron star to a black hole, just before the black hole forms (Baiotti *et al.* 2005). Unfortunately the typical frequencies of these modes (of the order of several kHz) are higher than the frequency region where the actual gravitational wave detectors are sensitive.

3.2 Polar quasi normal modes

The polar metric perturbations are physically coupled to the fluid perturbations. As shown in Section 2.1, the frequencies of the polar QNMs can be computed by solving a system of equations involving only the metric perturbations; however, they carry a strong imprint of the internal composition of the star, which is present in Eqs. (15) through the pressure and energy density profiles in the unperturbed star, which appear as coefficients of the differential equations. According to a scheme introduced by Cowling in Newtonian gravity (Cowling 1942), polar modes can be classified on the basis of the restoring force which prevails when the generic fluid element is displaced from the equilibrium position: for g-modes, or gravity modes, the restoring force is due to buoyancy, for p-modes it is due to pressure gradients. The mode frequencies are ordered as follows

$$..\omega_{g_n} < .. < \omega_{g_1} < \omega_f < \omega_{p_1} < .. < \omega_{p_n}..$$

and are separated by the frequency of the fundamental mode (f-mode), which has an intermediate character between g- and p-modes. As discussed in Section 3.1, general relativity predicts also the existence of polar w-modes, that are very weakly coupled to fluid motion and are similar to the axial w-modes (Kokkotas & Schutz 1992). The frequencies of axial and polar w-modes are typically higher than those of the fluid modes g, f and p.

If we are mainly interested in gravitational wave emission, the most interesting mode is the f-mode. For mature neutron stars, its frequency is in the range $1-3$ kHz, which is in the bandwidth of ground based detectors Virgo and LIGO (although not in the region where they are most sensitive); the damping times are of the order of a few tenths of seconds, therefore the excitation of the f-mode would appear in the Fourier transform of a gravitational wave signal as a sharp peak and could, in principle, be extracted from the detector noise by an appropriate data analysis. Moreover, the fundamental mode could be excited in several astrophysical processes, for instance in the aftermath of a gravitational collapse, in a glitch, or due to matter accretion onto the star. For this reason, since the early years of the theory of stellar perturbations, the interest of scientists working in this field has initially been focussed on the determination of the f-mode frequencies. After the work of Lindblom & Detweiler in 1983 and of Cutler & Lindblom in 1987, who respectively computed the $\ell = 2$ and $\ell > 2$ f-mode eigenfrequencies for the EOSs available at that time, more recently this work has been updated, and extended to other modes, by Anderson & Kokkotas (1998) and Benhar, Ferrari & Gualtieri (2004). In particular, in these two papers the f-mode frequency ν_f and the corresponding damping time τ_f have been computed to establish whether ν_f scales with the average density of the star, as it does in Newtonian gravity, and whether there also exists a scaling law for τ_f. The sets of EOSs used in the two works are not identical, because the papers were written six years apart, although some EOSs appear in both (see the two papers for details). The work done by Benhar, Ferrari &

Gualtieri (2004) also includes examples of hybrid stars, namely neutron stars with a core composed of quarks. In Anderson & Kokkotas (1998) ν_f and τ_f have been fitted by a linear function of the average density of the star $(M/R^3)^{1/2}$, and of its compactness M/R, as follows.

$$\nu_f = 0.78 + 1.635\sqrt{\frac{\tilde{M}}{\tilde{R}^3}}, \qquad \frac{1}{\tau_f} = \frac{\tilde{M}^3}{\tilde{R}^4}\left[22.85 - 16.65\left(\frac{\tilde{M}}{\tilde{R}}\right)\right], \tag{20}$$

where $\tilde{M} = M/1.4\ M_\odot$ and $\tilde{R} = R/(10\ \text{km})$. Here and in the following formulae ν_f is expressed in kHz and τ_f in s. The fits for ν_f and τ_f obtained by Benhar, Ferrari & Gualtieri (2004) using the new set of EOSs are

$$\nu_f = a + b\sqrt{\frac{M}{R^3}}, \quad a = 0.79 \pm 0.09\ (\text{in kHz}), \quad b = 33 \pm 2\ (\text{in km}), \tag{21}$$

and

$$\frac{1}{\tau_f} = \frac{cM^3}{R^4}\left[a + b\left(\frac{M}{R}\right)\right], \qquad a = [8.7 \pm 0.2]\cdot 10^{-2}, \qquad b = -0.271 \pm 0.009\,. \tag{22}$$

In Eqs. (21) and (22) mass and radius are in km (i.e. mass is multiplied by G/c^2) and $c = 3 \cdot 10^5$ km/s. The data for the different EOSs used by Benhar, Ferrari & Gualtieri (2004), and the fits given in Eqs. (20)–(22) are shown in Figure 4. ν_f is plotted in the upper panel as a function of the average density; the fit (20) is shown as a black dashed line labelled 'AK fit', whereas the new fit is indicated as a red continuous line labelled 'NS fit'. The NS fit is lower by about 100 Hz than the AK fit, showing that the new EOSs are, on average, less compressible than the old ones. The quantity $(R^4/cM^3)/\tau_f$ given in Eq. (22) is plotted in the lower panel of Figure 4 versus the stellar compactness M/R. In this case the AK fit for τ_f (20), and the NS fit (22) are nearly coincident. For comparison, in both panels of Figure 4 we show the frequency and the damping time of the f-mode of a population of strange stars, namely stars entirely made of up, down and strange quarks, modeled using the MIT Bag-model, spanning the allowed range of parameters, which are the Bag constant, the coupling constant α_S and the quark masses (see Benhar *et al.* 2007 for details). The parameters of the fits for strange stars are

$$\text{for } \nu_f \qquad a = -[0.8 \pm 0.08]\cdot 10^{-2}\,, \quad b = 46 \pm 0.2, \tag{23}$$

and

$$\text{for } \tau_f \qquad a = [4.7 \pm 5\cdot 10^{-3}]\cdot 10^{-2}, \qquad b = -0.12 \pm 3\cdot 10^{-4}\,. \tag{24}$$

In Figure 4 the fits for strange stars are labelled as 'SS fit'. It is interesting to note that the SS fits are quite different from those appropriate for neutron stars (AK- and NS-fits). First of all the errors on the parameters are much smaller, which indicate that the linear behaviour is followed by these stars, both for ν_f and for τ_f, irrespective of the values of the parameters of the model. Moreover, the difference between the fits is much larger for lower values of the average density.

The empirical relations given in Eqs. (20)–(24) could be used to constrain the values of the star mass and radius, where the values of ν_f and τ_f are identified in a detected

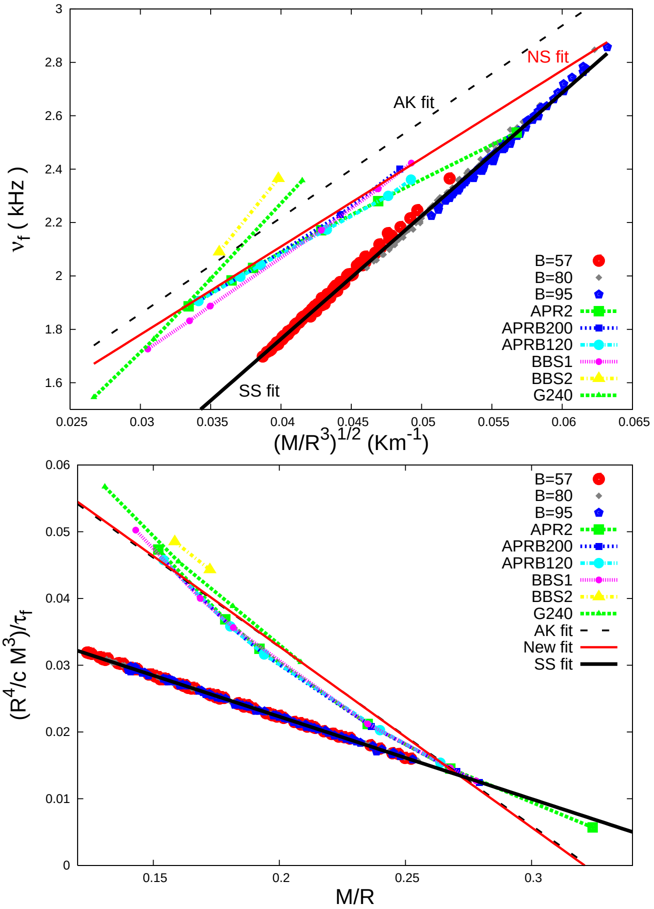


Figure 4. The frequency of the fundamental mode is plotted in the upper panel as a function of the square root of the average density for the different EOSs considered by Benhar, Ferrari & Gualtieri (2004). We also plot the fit given by Anderson & Kokkotas (1998) plotted as AK-fit and our fit (NS-fit). The NS-fit is systematically lower (about 100 Hz) than the AK-fit. The damping time of the fundamental mode is plotted in the lower panel as a function of the compactness M/R. The AK-fit and our fit, plotted respectively as a dashed and continuous line, do not show significant differences.

gravitational signal. The stellar parameters would be further constrained if other modes are excited and detected and, knowing them, we would gain information on the equations of state of matter in the neutron star core, whose uncertainty is due, as explained earlier, to our ignorance of hadronic interactions. Furthermore, if the neutron star mass is known, as it may be if the star is in a binary system, the detection of a signal emitted by the star oscillating in the f-mode may provide some further interesting information (Benhar *et al.* 2007). In Figure 5 we plot ν_f as a function of the stellar mass, for neutron/hybrid stars and for strange stars modeled using the MIT bag model. Note that 1.8 $M_\odot$ is the maximum

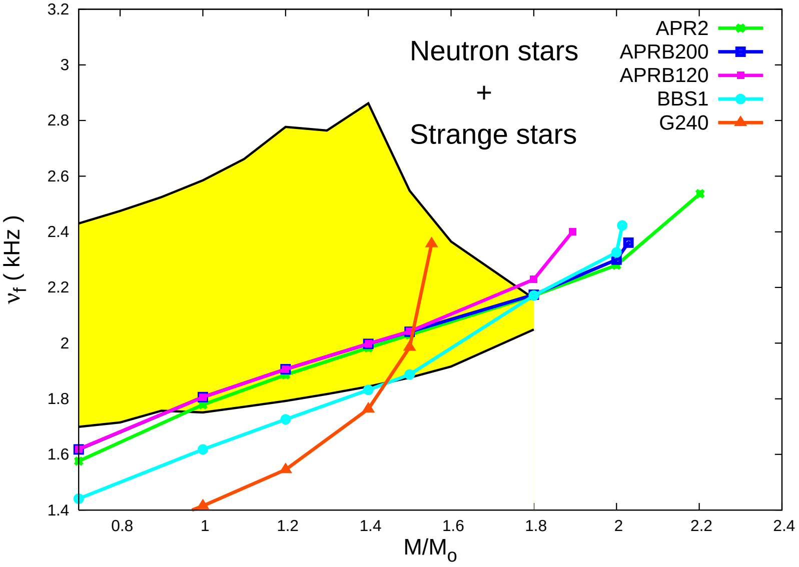


Figure 5. The frequency of the fundamental mode is plotted as a function of the mass of the star, for neutron/hybrid stars (continuous lines) and for strange stars modeled using the MIT bag model, spanning the set of parameters indicated in the range allowed by high energy experiments (dashed region).

mass above which no stable strange star can exist. We see that there is a small range of frequency where neutron/hybrid stars are indistinguishable from strange stars; conversely, there is a large frequency region where only strange stars can emit. Moreover, strange stars cannot emit gravitational waves with $\nu_f \lesssim 1.7$ kHz, for any value of the mass in the range we consider. For instance, if the stellar mass is $M = 1.4\ M_\odot$, a signal with $\nu_f \gtrsim 2$ kHz would belong to a strange star. Figure 5 also shows that, even if we do not know the mass of the star (as it is often the case for isolated pulsars), if $\nu_f \gtrsim 2.2$ kHz, apart from a very narrow region of masses where stars with hyperons would emit (EOS BBS1 and G240), we can reasonably rule out that the signal is emitted by a neutron star. In addition, it is possible to show that, since ν_f is an increasing function of the Bag constant B, if a signal emitted by an oscillating strange star were detected, it would be possible to set constraints on B much more stringent than those provided by the available experimental data (Benhar *et al.* 2007).

In conclusion, the QNM frequencies can be used to gain direct information on the equation of state of matter in a neutron star core.

The crucial question now is: do we have a chance to detect a signal emitted by a star oscillating in a polar quasi-normal mode? Detection chances depend on how much energy is channeled into the pulsating mode, which is unknown, and on whether the mode frequency is in the detector bandwidth. The signal emitted by a star pulsating in a given mode of frequency ν and damping time τ, has the form of a damped sinusoid

$$h(t) = \mathcal{A} e^{-(t-t_0)/\tau} \sin[2\pi\nu(t-t_0)] \quad \text{for } t > t_0, \tag{25}$$

where t_0 is the arrival time of the signal at the detector (and $h(t) = 0$ for $t < t_0$). The wave amplitude $\mathcal{A}$ can be expressed in terms of the energy radiated in the oscillations,

$$\mathcal{A} \approx 7.6 \times 10^{-24} \sqrt{\frac{\Delta E_\odot}{10^{-12}} \frac{1\ \mathrm{s}}{\tau}} \left(\frac{1\ \mathrm{kpc}}{d}\right)\left(\frac{1\ \mathrm{kHz}}{\nu}\right). \tag{26}$$

where $\Delta E_\odot = \Delta E_{GW}/M_\odot c^2$. This quantity is unknown. Therefore to assess the detectability of a signal we can only evaluate how much energy should be emitted in a given mode, in order for the signal to be detected by a given detector with an assigned signal to noise ratio (S/N)

$$\left(\frac{S}{N}\right)^2 = \frac{4Q^2}{1 + 4Q^2} \frac{\mathcal{A}^2 \tau}{2S_n}. \tag{27}$$

In this equation $Q = \pi\nu\tau$ is the quality factor and S_n is the detector spectral noise density. Assuming as a bench-mark for $\Delta E_\odot$ the energy involved in a typical pulsar glitch, in which case a mature neutron star might radiate an energy of the order of $\Delta E_{GW} = 10^{-13} M_\odot c^2$, and assuming $\nu \sim 1500$ Hz, $\tau \sim 0.1$ s, $d = 1$ kpc, we find $\mathcal{A} \approx 5 \times 10^{-24}$. Such a signal is too weak to be seen by actual detectors, therefore we conclude that 3rd generation detectors are needed to detect signals from old neutron stars. More promising are the oscillations of newly-born neutron stars; indeed, since a NS forms as a consequence of a violent, and generally non-symmetric event – the gravitational collapse – a fraction of its large mechanical energy may go into non-radial oscillations and would be radiated in gravitational waves. Thus, during the first few seconds of the NS life more energy could be stored in the pulsation modes than when the star is cold and old. In addition, during this time the star is less dense than at the end of the evolution; consequently, the frequencies of the modes which depend on the stellar compactness (as for instance the f-mode) are lower and therefore span a frequency range where the detectors are more sensitive (Ferrari, Miniutti & Pons 2003). For instance, if we assume that an energy $\Delta E_{GW} = 1.6 \cdot 10^{-9}\ M_\odot c^2$ is stored in the f-mode of a neutron star just formed in the Galaxy, the emitted signal would be detectable with a signal to noise ratio $S/N = 8$ by advanced Virgo/LIGO, and with $S/N = 2.7$ by Virgo+/LIGO, the upgraded configurations now being in operation.

4. Stellar perturbations and magnetar oscillations

Magnetars are neutron stars whose magnetic field is, according to current models, as large as 10^{15} G (Thompson & Duncan 1993, 2001). During the last three decades some very interesting astrophysical events have been observed which are connected to magnetar activity and stellar pulsations. They involve Soft Gamma Repeaters (SGRs), which are thought to be magnetars; these sources occasionally release bursts of huge amount of energy ($L \simeq 10^{44} - 10^{46}$ ergs/s), and these giant flares are thought of being generated from large-scale rearrangements of the inner field, or catastrophic instabilities in the magnetosphere (Thompson & Duncan 2001; Lyutikov 2003). Up to now, three of these events have been detected: SGR 05026-66 in 1979, SGR 1900+14 in 1998 and SGR 1806-20 in 2004. In two of them (SGR 1900+14 and SGR 1806-20), a tail lasting several hundred seconds has been observed, and a detailed study of this part of the spectrum has revealed the presence of quasi-periodic oscillations (QPOs) with frequencies

$$18, \ 26, \ 30, \ 92, \ 150, \ 625 \quad \text{and} \quad 1840 \quad \text{Hz}$$

for SGR 1806-20 (Watts & Strohmayer 2006), and

$$28, \ 53, \ 84 \quad \text{and} \quad 155 \quad \text{Hz}$$

for SGR 1900+14 (Strohmayer & Watts 2006). The discovery of these oscillations stimulated an interesting and lively debate (still ongoing) among groups working on stellar perturbations, about the physical origin of these sequences. Of course in order to study these oscillations the magnetic field and its dynamics have to be included in the picture, while rotation plays a less important role, since observed magnetars are all very slowly rotating. The problem presents extreme complexity both at conceptual and at computational levels; therefore it is usually approached using simplifying assumptions and/or approximations. Some of the studies try to explain the observed modes in terms of torsional oscillations of the crust (Samuelsson & Andersson 2007; Sotani, Kokkotas & Stergioulas 2007), others attribute the observed spectra to global magneto-elastic oscillations (Glampedakis, Samuelsson & Andersson 2006), still others investigate the interaction between the torsional oscillations of the magnetar crust and a continuum of magnetohydrodynamic modes (the Alfven continuum) in the fluid core (Levin 2007; Sotani, Kokkotas & Stergioulas 2008; Colaiuda, Beyer & Kokkotas 2009; Cerdá-Durán, Stergioulas & Font 2009; Colaiuda & Kokkotas 2010; Gabler *et al.* 2011) using different approaches and approximations. In particular, in (Colaiuda & Kokkotas 2010) the torsional oscillations of a magnetar have been studied in a general relativistic framework, perturbing Einstein's equations in the Cowling approximation, i.e. neglecting gravitational field perturbations. By this approach the crust-core coupling due to the strong magnetic field has been shown to be able to explain the origin of the observed frequencies, at least for SGR 1806-20, if a suitable stellar model is considered. With this identification, constraints on the mass and radius of the star, and consequently on the EOS in the core, can be set; estimates of the crust thickness and of the value of the magnetic field at the pole can also be inferred.

Thus, the theory of stellar perturbations has been generalized to magnetized stars, although for now only with considerable restriction, since only torsional oscillations have been considered (i.e. axial perturbations) and only in the Cowling approximation. Nevertheless, it already provides very interesting information on the dynamics of these stars and allows us to confront the predictions with astronomical observations.

5. Concluding remarks

I would like to conclude this review by mentioning the fact that the theory of perturbations of rotating stars has not been developed to the same extent as the theory of non-rotating stars. The main reason is that the mathematical tools appropriate for a successful variable separation has not been found yet. When the perturbations of a non-rotating black hole are studied, separation of variables is achieved by expanding all tensors in tensorial spherical harmonics. In the case of Kerr perturbations, namely of perturbations of an axisymmetric, Petrov type D background, the same result is obtained by expanding the Newman-Penrose quantities in oblate spheroidal harmonics. When perturbing a rotating star, i.e. an axisymmetric solution of Einstein+hydro equations, an expansion in terms of tensorial spherical harmonics leads, as we have seen in the case of slow rotation in Section 2.2, to a coupling between polar and axial perturbations. If rotation is not slow, the number of couplings to be considered increases to such an extent that the problem becomes untreatable, both from a theoretical and from a computational point of view. One may argue that, since the background of a rapidly rotating star is not spherically symmetric, tensorial spherical harmonics are unappropriate, and this is certainly true. However, even if we try to use the oblate spheroidal harmonics and the Newmann Penrose formalism fails: the coupling between the metric and the fluid makes the separation impossible, at least in terms of these harmonics (unlike the Kerr metric, the metric describing a star is not of Petrov type D). For this reason, perturbations of rotating stars have been studied either in the slow rotation regime, or using the Cowling approximation, which neglects spacetime perturbations, or using other simplifying assumptions. For instance, as far as the mode calculation is concerned, the Cowling approximation allows determination with reasonable accuracy the frequency of the higher order p-modes, of the g-modes and of the inertial modes, like the r-modes, thus allowing us to gain information on the onset of related instabilities. Conversely, the determination of the f-mode frequency, which is so important from the point of view of gravitational wave emission, is not very precise, leading to errors as large as ~20%.

However, it should be mentioned that non-linear simulations of rotating stars have produced very interesting results; for instance in a recent paper, Stergioulas and collaborators (Zink *et al.* 2010) have been able to follow the frequency of the non-axisymmetric fundamental mode of a sequence of rotating stars with increasing angular velocity, up to the onset of the CFS instability, making also very optimistic estimates of the amount of gravitational radiation which could be emitted in the process. To describe matter in the neutron star they used a simple model (a polytropic equation of state and uniform rotation); however, their result indicates that numerical relativity is making giant steps in this field. Thus, supercomputers are making accessible very complex problems, which only ten years ago one would not have dreamed of solving; however, perturbation theory still remains a very powerful tool to investigate many physical problems and it should be used in parallel with the numerical work to gain a deeper insight into the physics of stellar oscillations.

References

Andersson N., 1994, CQG, 11, 3003

Andersson N., Araujo M.E., Schutz B.F., 1993a, CQG, 10, 735

Andersson N., Araujo M.E., Schutz B.F., 1993b, CQG, 10, 757

Andersson N., Araujo M.E., Schutz B.F., 1993c, Phys. Rev. D, 49, 2703

Andersson N., Kokkotas K.D., 1998, MNRAS, 299, 1059

Andersson N., Thylwe K., 1994, CQG, 11, 2991;

Baiotti L., Hawke I., Rezzolla L., Schnetter E., 2005, Phys. Rev. Lett., 94, 131101

Benhar O., Berti E., Ferrari V., 1999, MNRAS, 310, 797

Benhar O., Ferrari V., Gualtieri L., 2004, Phys. Rev. D70, 124015

Benhar O., Ferrari V., Gualtieri L., Marassi S., 2007, GRG, 39, 1323

Campolattaro A., Thorne K.S., 1970, ApJ, 159, 847

Cerdá-Durán P., Stergioulas N., Font J. A., 2009, MNRAS, 397, 1607

Chandrasekhar S., 1984, The mathematical theory of black holes, Claredon Press, Oxford

Chandrasekhar S., Detweiler S.L., 1975, Proc. R. Soc. Lond., A344, 441

Chandrasekhar S., Ferrari V., 1990a, Proc. R. Soc. Lond., A428, 441

Chandrasekhar S., Ferrari V., 1990b, Proc. R. Soc. Lond., A432, 247

Chandrasekhar S., Ferrari V., 1991a, Proc. R. Soc. Lond., A435, 645

Chandrasekhar S., Ferrari V., 1991b, Proc. R. Soc. Lond., A433, 423

Chandrasekhar S., Ferrari V., 1991c, Proc. R. Soc. Lond., A434, 449

Chandrasekhar S., Ferrari V., 1992, Proc. R. Soc. Lond., A437, 133

Chandrasekhar S., Ferrari V., Winston R., 1991, Proc. R. Soc. Lond., A434, 635

Colaiuda A., Beyer H., Kokkotas K.D., 2009, MNRAS, 396, 1441

Colaiuda A., Kokkotas K.D., 2010, arXiv:1012.3103v2

Cowling T.G., 1942, MNRAS, 101, 367

Cutler C., Lindblom L., 1987, ApJ, 314, 234

Fano U., 1985, Phys. Rev. A, 32, 617

Ferrari V., Miniutti G., Pons J.A., 2003, MNRAS, 342, 629

Ferrari V., Mashhoon B., 1984a, Phys. Rev. D, 30, 295

Ferrari V., Mashhoon B., 1984b, Phys. Rev. Lett., 52, 1361

Gabler M., Cerdá Durán P., Font J.A., Muller E., Stergioulas N., 2011, MNRAS, 410, L37

Glampedakis K., Samuelsson L., Andersson N., 2006, MNRAS, 371, L74

Ipser J.R., Thorne K.S., 1973, ApJ, 181, 181

Landau L.D., Lifshitz E.M., 1975, The classical theory of fields, New York: Pergamon Press

Levin Y., 2007, MNRAS, 377, 159

Lindblom L., Detweiler S.L., 1983, ApJ Suppl., 53, 73

Lyutikov M., 2003, MNRAS, 346, 540

Kokkotas K.D., 1994, MNRAS, 268, 1015

Kokkotas K.D., Schutz B.F., 1992, MNRAS, 255, 119

Pandharipande V.R., 1971a, Nucl. Phys. A, 174, 641

Pandharipande V.R., 1971b, Nucl. Phys. A, 178, 123

Pandharipande V.R., Smith R.A., 1975, Phys. Lett., 59, 15

Press W.H., 1971, ApJ, 170, L105

Pudliner B.S., Pandharipande V.R., Carlson J., Pieper S.C., Wiringa R.B., 1995, Phys. Rev. C, 56, 1720

Regge T., Wheeler J.A., 1957, Phys. Rev., 108, 1063

Samuelsson L., Andersson N., 2007, MNRAS, 374, 256

Schutz B.F., Will C.M., 1985, ApJ, 291, L33-L36

Sorkin R., 1991, Proc. R. Soc. Lond., A435, 635

Sotani H., Kokkotas K.D., Stergioulas N., 2007, MNRAS, 375, 261

Sotani H., Kokkotas K.D., Stergioulas N., 2008, MNRAS, 385,L5

Strohmayer T.E., Watts A.L., 2006, ApJ 653, 593

Teukolsky S., 1972, Phys. Rev. Lett., 29, 1114

Teukolsky S., 1973, ApJ, 185, 635

Thompson C., Duncan R.C., 1993, ApJ, 408, 194

Thompson C., Duncan R.C., 2001, ApJ, 561, 980

Thorne K.S., 1969a, ApJ, 158, 1

Thorne K.S., 1969b, ApJ, 158, 997

Thorne K.S., Campolattaro A., 1967, ApJ, 149, 591

Thorne K.S., Campolattaro A., 1968, ApJ, 152, 673

Vishveshwara C.V., 1970, Phys. Rev.D, 1, 2870

Walecka J.D., 1974, Ann. Phys., 83, 491

Watts A.L, Strohmayer T.E., 2006, ApJ, 637, L117
Wiringa R.B., Fiks V., Fabrocini A., 1988, Phys. Rev. C, 32, 1057
Wiringa R.B., Stoks V.G.J., Schiavilla R., 1995, Phys. Rev. C, 51, 38
Zerilli J.F., 1970a, Phys. Rev. D, 2, 2141;
Zerilli J.F., 1970b, Phys. Rev. Lett., 24, 737
Zink B., Korobkin O., Schnetter E., Stergioulas N., 2010, Phys. Rev. D81, 084055

Bull. Astr. Soc. India (2011)

The Chandra X-ray observatory

Gordon P. Garmire[*]
Penn State University, PA 16802, USA

Received February 3, 2011; accepted March 12, 2011

Abstract. This paper describes a brief history of the development of the Chandra X-ray observatory, based on the talk which was presented on October 17, 2010 at the Chandrasekhar Centennial Symposium held in the campus of the University of Chicago.

Keywords : X-rays: general – telescopes – space vehicles: instruments

1. Introduction

First, I would like to thank the organizing committee for inviting me to this very interesting conference. Nobel laureate Subrahmanyan Chandrasekhar is certainly deserving of such an accolade. I would like to turn from the theoretical topics that have occupied most of today's lectures, to a more experimental topic, that of the history of the Chandra X-ray Observatory. This Observatory was built to explore the high energy Universe which features such exotic objects as neutron stars and black holes. Chandra was very interested in black holes as is illustrated by his description that "The black holes of nature are the most perfect macroscopic objects there are in the Universe: the only elements in their construction are our concepts of space and time." One of the closest examples of a massive black hole is the one at the centre of our Galaxy in SgrA* shown in Figure 1 from a 500 ks Chandra image obtained by Muno *et al.* (2003). This remarkable object is emitting X-rays at a rate of only 2×10^{33} ergs/s which is nearly nine orders of magnitude below that of an active galactic nucleus (AGN). One interesting feature of this object is the flaring activity that was detected by Chandra (Baganoff *et al.* 2001). These flares occur aperiodically with a frequency of approximately once per day and represent increases of up to a factor of one hundred in luminosity for up to an hour in duration.

A brief outline of the paper is as follows which describes different phases of its history in the following Sections: a brief history of the Observatory; the early years; the big test; construction at last; testing; launch. The last Section describes a few results.

[*] e-mail: garmire@astro.psu.edu

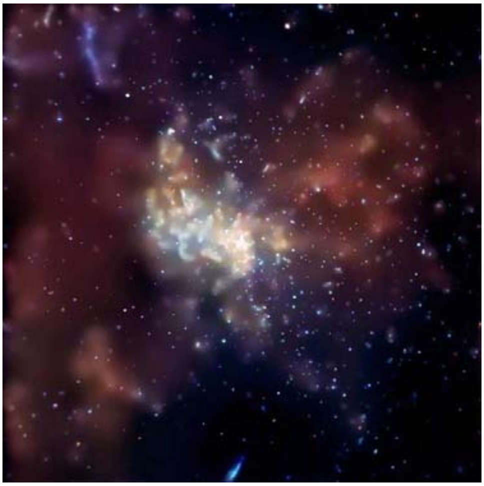


Figure 1. The 16 arcmin square region of the Galactic Centre with SgrA* at the centre of the image was obtained from a 500 ks exposure using the Chandra X-ray Observatory (Chandra archives at http://chandra.harvard.edu/photo/2003/0203long).

2. A brief history of the Observatory

The idea of building an X-ray telescope was advanced by Riccardo Giacconi and Bruno Rossi in a seminal paper in 1960 (Giacconi & Rossi 1960). The idea used grazing incidence X-rays on highly polished metal surfaces shaped into paraboloidal and hyperboloidal shapes to form a sharp image. A schematic diagram of this concept is shown in Figure 2.

Following the discovery of an extra-solar X-ray source in 1962 (Giacconi *et al.* 1962), Giacconi and his collaborators at American Science and Engineering proposed a program of X-ray astronomy to NASA which included a large focusing X-ray telescope of 4 feet (1.2 m) with a focal length of 30 feet ($\sim$10m) in 1963 and an angular resolution of about one arcmin. NASA accepted the Sounding Rocket portion of the proposal but deferred on the large telescope. In 1968 NASA initiated the High Energy Astrophysics Program of four large observatories to cover the field from X-rays to high energy cosmic rays which included a 1.2 m diameter X-ray telescope. The program was cancelled in 1973 by NASA and reconstituted as a much smaller program that included a 0.6 m X-ray telescope which became the Einstein Observatory in 1978. In 1976 Dr. Riccardo Giacconi and Dr. Harvey Tananbaum, then at Harvard, submitted a letter proposal to NASA to begin the study of a 1.2 m diameter X-ray observatory. NASA accepted the idea and organized a study group to define the Observatory. In order not to call attention to the fact that this was another large observatory program being initiated by NASA like the Hubble Telescope, the group decided to call the observatory the Advanced X-ray Astrophysics Facility to make it sound less expensive. In 1985 four focal plane instruments and two grating designs were selected for study. These included the High Resolution Camera, the Advanced CCD Imaging Spectrometer to the AXAF CCD Imaging Spectrometer (later changed to the Advanced CCD Imaging Spectrometer after the name Chandra was chosen for the Observatory), the Bragg

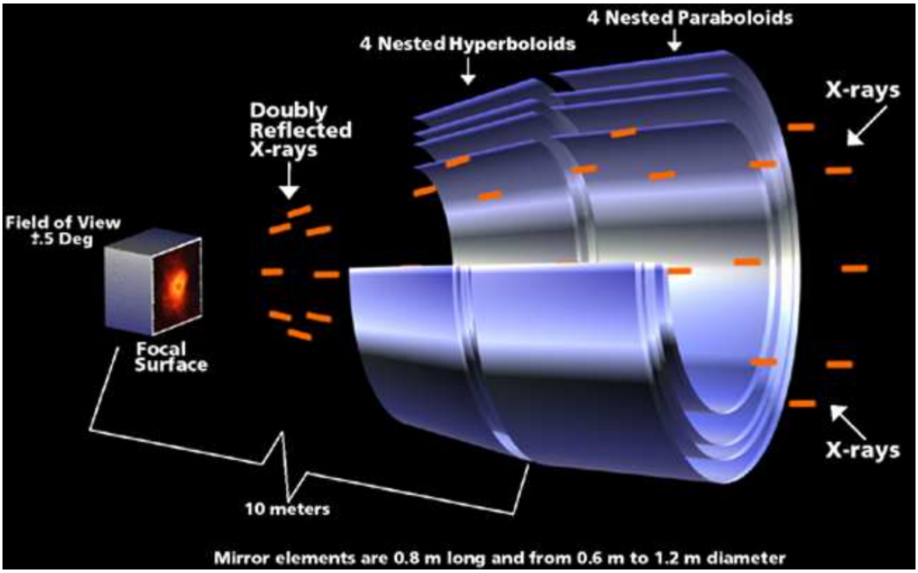


Figure 2. Schematic representation of the Giacconi-Rossi concept as applied to the Chandra X-ray Observatory mirrors. In Chandra the focal surface contains the High Resolution Camera and the Advanced CCD Imaging Spectrometer on a translation table (http://chandra.harvard.edu/graphics/resources/illustrations/cxcmirrors-72.jpg).

Crystal Grating Spectrometer and the X-ray Calorimeter together with the Low Energy and Medium/High Energy objective transmission gratings. The Observatory was to be in low earth orbit and be serviced by the Space Shuttle astronauts. The instruments were designed to be changed for new improved instruments as new technology became available.

3. The early years

Even though instruments had been selected and a spacecraft contractor was selected shortly thereafter, there was no guarantee that the program would receive a ‘New Start’ by Congress. The AXAF program received the highest ratings in the 1980 and 1990 Decadal Surveys by the astronomical community, but Congress was reluctant to fund a program with such a high technical risk, that of producing 0.5 arcsec X-ray mirrors which were better than any mirrors made thus far by an order of magnitude. Dr. Charles Pellerin, director of the NASA Astrophysics Division, proposed a clever way to sell the program. By combining the Hubble Space Telescope, the Compton Gamma-Ray Observatory, AXAF and the Space Infrared Telescope into a Great Observatories Program (a name suggested by George Field) he produced a package that was more saleable to Congress. To put the Congressional staffers at ease, a bargain was made to make the largest mirror and test it to prove that the technology was up to the challenge. If the mirror failed the test, then the program would not go forward. In 1991 the largest of the AXAF mirror pairs was completed and ready for testing after a tremendous effort by the Telescope Scientist, Dr Leon van Speybroeck, and

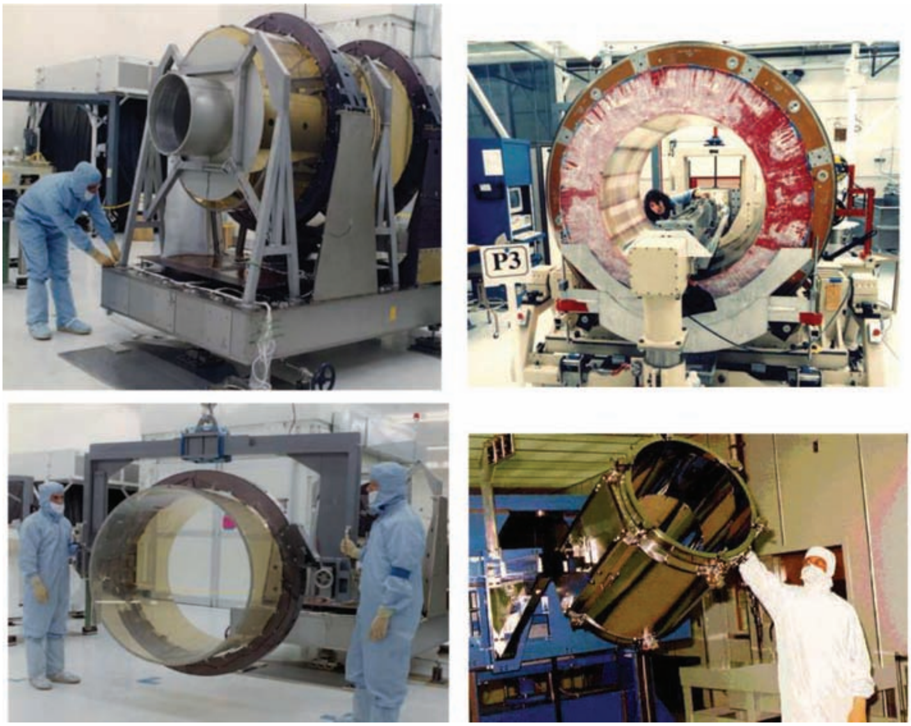


Figure 3. Upper left: The mirror assembly being readied for testing at the MSFC testing facility (http://chandra.harvard.edu/graphics/resources/illustrations/veta7_72.jpg); upper right: the polishing process at HDOS (http://chandra.harvard.edu/graphics/resources/illustrations/presentPolish1-72.jpg); lower left: the completed P1 mirror before coating it with Iridium (http://chandra.harvard.edu/graphics/resources/illustrations/2c_mounted_primary.jpg); lower right: one of the AXAF mirrors after the coating process (http://chandra.harvard.edu/graphics/resources/illustrations/barrel.jpg).

the mirror fabricators at Hughes Danbury Optical Systems (HDOS). Dr. Martin Weisskopf, the AXAF project scientist, and Danny Johnson, the Project Engineer, were in charge of the design and construction of the testing facility at Marshal Space Flight Center (MSFC), which they successfully completed in time for the X-ray testing of the mirror pair. The testing revealed that the mirrors sagged in the 1g field of the test facility and special fixturing had to be created to remove the effects of gravity on the mirrors. Once this was done, the mirrors passed the test with flying colours. Figure 3 shows the mirrors ready for testing at the MSFC facility.

Although the mirror test had been successful, the program was still not out of the woods. The Super Conducting Super Collider had experienced very significant over runs in its budget and was cancelled in 1992. The Space Station was also experiencing very large over runs in costs. Attempts within NASA to cancel the Space Station failed in Congress and NASA was told to proceed using what money it had with no budgetary increases. This placed severe constraints on the science budget. The Office of Management and Budget

(OMB) had projected that the AXAF program would have a total run out cost in excess of 6 billion dollars, making it a target for cancellation as well. Charlie Pellerin at NASA HQ called a meeting to discuss ways to reduce the cost of AXAF and to control the run out costs of the mission. The Hubble Space Telescope was alerting Congress to the very substantial costs of servicing a mission in low Earth orbit. There was some reluctance among the AXAF scientists to descope AXAF, but a presentation by Leonard Fisk, the Associate Administrator of NASA, convinced the group that descoping was the only way to keep the program alive. The decision after much debate was to reduce the number of mirror pairs from six to four and to carry only the two focal plane instruments which could be used for enhanced spectroscopy by employing the objective gratings behind the mirrors to produce a dispersed spectrum. In consultation with TRW, the prime contractor, and the instrument teams, the observatory was redesigned to go to a high Earth orbit that could not be serviced by the Space Transportation System, thereby guaranteeing that there would be no mission servicing costs. With this change of design, Congress agreed to fund the construction of the observatory. In order to salvage the higher resolution spectroscopy portion of the mission, a second mission called AXAF-S was designed to carry the high-resolution calorimeter and Bragg crystal spectrometer. Unfortunately, after a preliminary design of this mission, it was cancelled.

4. Construction at last

In 1993 the program went into high gear. The mirror facility at the Hughes Danbury Optical Systems plant began the process to grind, figure and polish the remaining three mirror pairs. The polishing process is illustrated in Figure 3, where the axis of the mirror is nearly horizontal and narrow shaping and polishing tools are used to polish and figure the surface between metrology measurements in the metrology facility, specifically designed for the AXAF mirrors.

The creation of the AXAF mirrors, which are an order of magnitude more precise than any such mirrors ever made, are the result of the leadership and ability of the Telescope Scientist, Dr. Leon van Speybroeck of the Center for Astrophysics at Harvard and the dedicated and skillful engineers and opticians at HDOS. After the mirrors were figured and polished and tested at HDOS they traveled to the Optical Coatings Laboratory Inc (OCLI) in Santa Rosa, CA, where they were coated with Iridium to provide the highest reflectivity at X-ray wavelengths. One of the mirrors is shown in Figure 3 at the OCLI facility. Following the coating of the mirror surface the mirrors were shipped to Eastman Kodak where they were assembled into a holding fixture that provided the accurate alignment of the mirror pairs to complete the High Resolution Mirror Assembly (HRMA). This process is shown in Figure 4.

This process was crucial to the formation of a high quality image. Each mirror pair must focus on the same point to provide a sharp image. Unfortunately, the inner mirror pair slipped slightly in the gluing process so that a ghost image was formed about one half arc second away from the focus of the other mirror pairs. Since this was the smallest mirror pair with the least area at the lower energies, this image is not usually apparent. It can be detected if a source is piled up in the primary image, thereby reducing the intensity of the center of the image and revealing the fainter image from the inner pair of mirrors which is

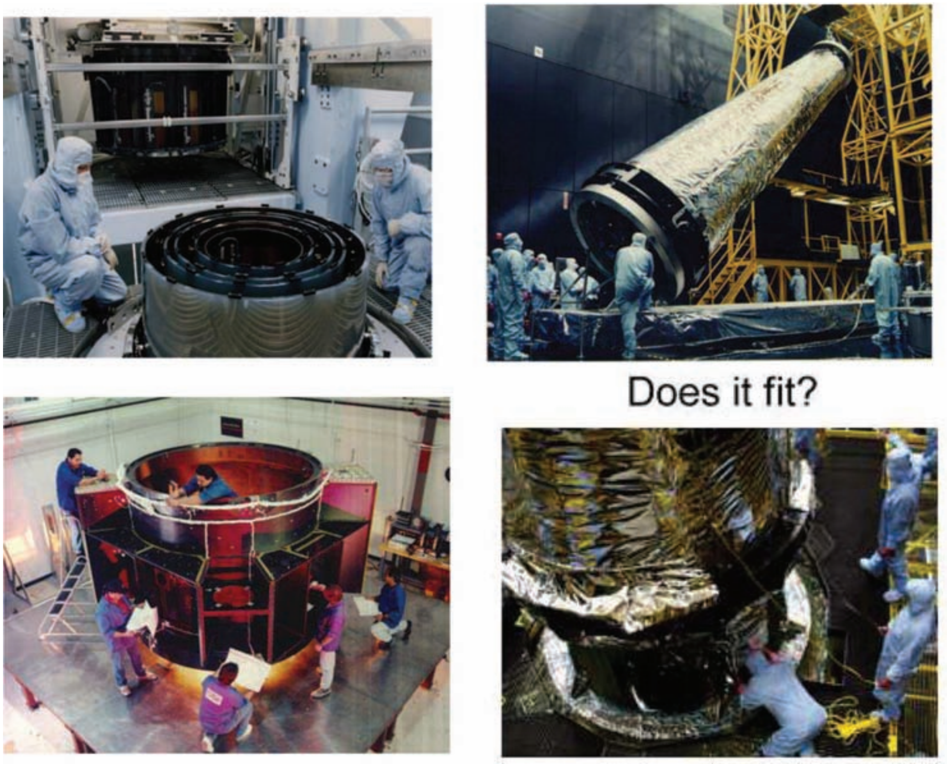


Figure 4. Upper left: The fabrication of the High Resolution Mirror Assembly (HRMA) at Eastman Kodak (http://chandra.harvard.edu/graphics/resources/illustrations/hrma_11.jpg); upper right: the carbon fiber epoxy telescope tube covered with a protective coating (http://chandra.harvard.edu/graphics/resources/illustrations/craftOptBench45-72.jpg); lower left: the epoxy structure of the spacecraft at TRW's Space and Electronics Group. This material was adopted to reduce the weight of the AXAF (http://chandra.harvard.edu/graphics/resources/illustrations/craftBusRed1-72.jpg); lower right: The telescope being inserted into the spacecraft (courtesy of Robert Burke and Blake Bullock of Northrop Grumman Aerospace Systems).

not piled up. The optical bench or telescope tube was fabricated at Kodak as well as the HRMA. This large carbon fiber epoxy structure is shown in Figure 4. The telescope tube was not baked to reduce the volatiles that were trapped in the matrix. These volatiles slowly escaped after the telescope was in orbit and may have been part of the reason that the cold filter of the Advanced CCD Camera was slowly coated with an unknown layer of material. The spacecraft was designed, assembled and tested at TRW Space and Electronics Group (now part of Northrop Grumman Aerospace Systems). Figure 4 shows the spacecraft under construction at the Redondo Beach, CA facility of TRW. The telescope tube and HRMA were brought together in the spring of 1998. Figure 4 shows the mating of the telescope to the Observatory.

5. The instruments

The scientific instruments for AXAF consist of two focal plane cameras, the High Resolution Camera (HRC) and the Advanced CCD Imaging Spectrometer (ACIS) as well as two objective transmission grating assemblies located just behind the mirrors. The objective transmission gratings were the responsibility of the two Principal Investigators, Dr. Claude Canizares of MIT for the high and medium energy transmission gratings, and Dr. Albert Brinkman of the Space Research Organization of the Netherlands for the low energy transmission grating. The HRC team was led by Dr. Steven Murray of the Center for Astrophysics at Harvard, and the ACIS team was led by Dr. Gordon Garmire of the Pennsylvania State University.

The objective transmission grating were challenging to construct in that they were extremely thin and had to be self supporting in order to transmit the lowest energy X-rays as well as survive the high acoustic levels encountered during a Space Shuttle launch. I don't have time to go into the details of the fabrication process other than to say that it uses high resolution lithography. An example of a completed grating assembly is shown in Figure 5.

The High Resolution Camera employed micro-channel plates similar to the Rosat and Einstein Observatory high resolution imagers. A large micro-channel plate formed the imaging array and three smaller plates formed the spectroscopic array imager. The micro-channel plates used 10 micron pores which formed the limiting spatial resolution of the camera. With the 10 m focal length of the HRMA this provided an angular resolution of about 0.2 arcsec on the sky, over sampling the point spread function of the HRMA which is about 0.5 arcsec. The HRC does not provide a very accurate determination of the X-ray energy, but when combined with the transmission gratings, particularly the low energy grating, it can achieve an energy resolution E/ΔE of about 1000. Another advantage of the HRC is that it provides 16 microsecond time resolution for the X-ray events for sources that are positioned on the central micro-channel plate of the spectroscopy array and are below the telemetry saturation level. Figure 5 (upper right) shows a picture of the HRC and Figure 5 (lower left) shows the completed HRC with its electronics ready for mounting on the translation table that moves the camera in and out of the focal plane of the HRMA.

The ACIS instrument employs CCDs as the basic detecting element. The CCDs for ACIS, which were fabricated by MIT's Lincoln Laboratory, are of two types: front illuminated and back illuminated. Front illuminated CCDs expose the portion of the silicon chip that the CCD is constructed on that is covered by the readout gates and insulators. This limits the lowest energy that the CCD can detect to about 0.4 keV with a small narrow (in energy) window at 0.25 keV. The depth of the depleted silicon under the gate structure is about 50 microns, which results in a upper energy cut-off at around 8 keV. The energy resolution of the CCD varies from about 100 eV at 1 keV to about 160 eV at 6 keV. The back illuminated CCD has been thinned to about 40 microns thick and the readout gate structure is away from the incoming X-rays such that the incident X-rays fall on the thinned silicon layer that is only covered by about 0.03 microns of a special backside treatment. This increases the quantum detection efficiency of the back illuminated CCD to about 60% at 0.25 keV. The CCDs must be covered by an optically opaque film to prevent them from responding to visible light which is focused by the HRMA onto the focal plane.

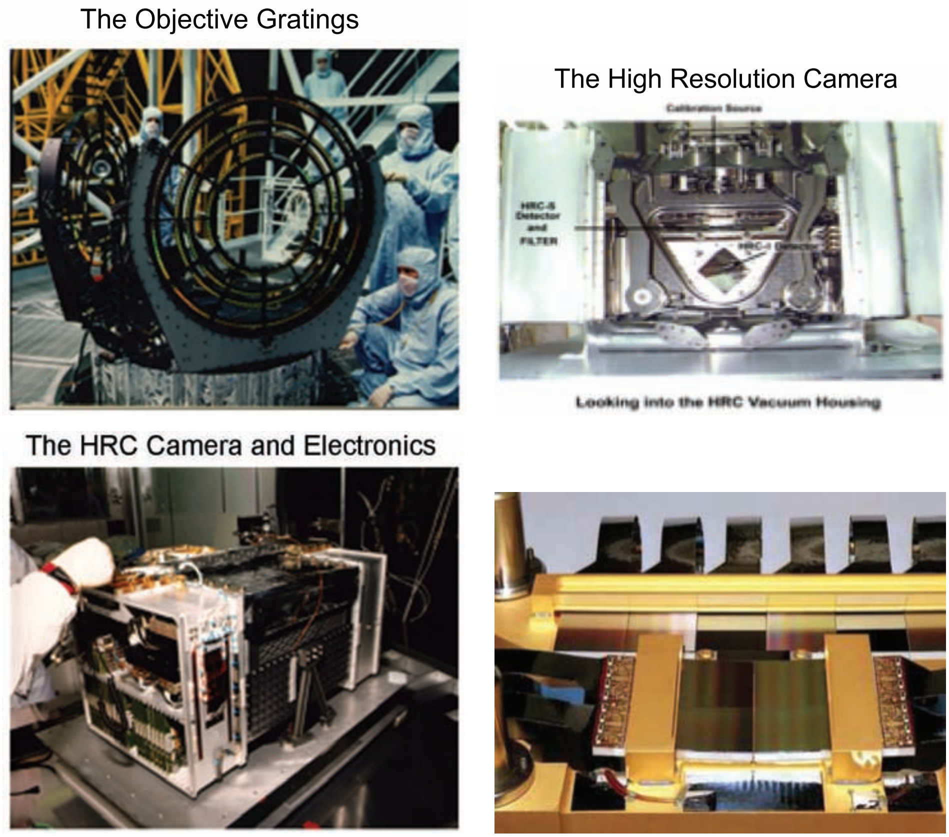

The Objective Gratings
The High Resolution Camera
Calibration Source
HRC-S Detector and FILTER
HRC-I Detector
Looking into the HRC Vacuum Housing
The HRC Camera and Electronics

Figure 5. Upper left: The objective transmission grating assembly mounted behind the HRMA. Either the high and medium transmission gratings or the low energy transmission gratings can be rotated into place behind the HRMA to provide a dispersed spectrum of the object under study (http://chandra.harvard.edu/graphics/resources/illustrations/gratingsLow1-72.jpg); upper right: The High Resolution Camera as observed from the HRMA (http://chandra.harvard.edu/graphics/resources/illustrations/HRClabel-72.jpg); lower left: The HRC with its electronics assembly ready for mounting to the translation table (http://chandra.harvard.edu/graphics/resources/illustrations/HRCbox-72.jpg); lower right: The ACIS array of 10 CCDs. The four CCDs in the square array are the imaging array front illuminated CCDs and the six CCDs in a linear array form the spectroscopic array to image the spectrum dispersed by the transmission gratings. The two mirror surface CCDs are the back illuminated CCDs. The gold colored bars across a portion of the CCDs are radiation shields to prevent X-rays from impinging upon the frame store portion of the CCDs. The optical blocking filters are not shown in this view (Courtesy of MITs Lincoln Laboratory, Bernie Kosicki).

The polyimide plastic film, provided by the Luxel Corporation, that is 2000 Angstroms thick and coated by 1600 Angstroms of aluminum covers the four front illuminated CCDs comprising the imaging array. The spectroscopic array of 6 CCDs, two of which are back illuminated, is covered by the polyimide film of the same thickness and a 1300 Angstrom

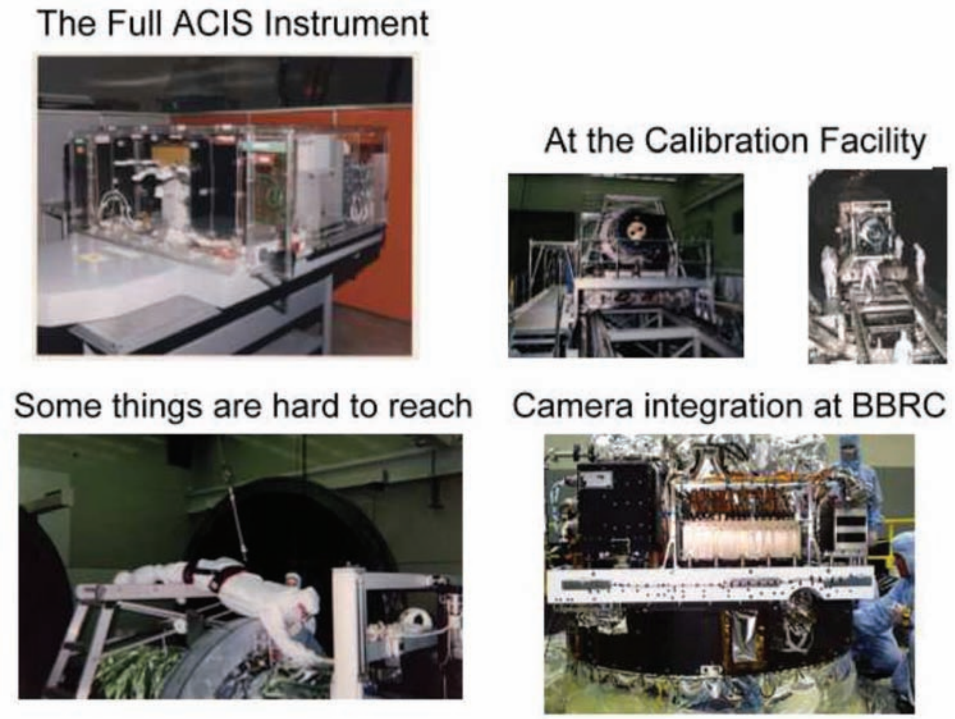

The Full ACIS Instrument
At the Calibration Facility
Some things are hard to reach
Camera integration at BBRC

Figure 6. Upper left: The complete ACIS camera and electronics assembly. The white straps are thermal conductors which will be connected to radiators on the spacecraft to cool the CCDs and camera housing (Courtesy of Ball Brothers Aerospace Corporation); upper right: The HRMA being prepared for insertion into the vacuum chamber at the MSFC test facility (photo courtesy of Robert Burke, Northrop Grumman Aerospace Systems); lower left: Making some adjustments to the HRMA mount at the MSFC test facility (Courtesy of Robert Burke, Northrop Grumman Aerospace Systems); lower right: The instruments being assembled and tested at BBRC (http://chandra.harvard.edu/graphics/resources/illustrations/modul1-72.jpg).

aluminum film. Over the course of the mission, a slow buildup of some form of contaminant has coated the filter, decreasing the low energy efficiency of the ACIS instrument. By using the onboard calibration source, it has been possible to measure this buildup and correct the quantum efficiency accordingly. A picture of the ACIS CCD array is shown in Figure 5 (lower right). The completed ACIS instrument with its electronics is shown in Figure 6.

The CCDs are read out every 3.24 seconds in their normal mode of operation. Special modes, such as using a reduced number of CCDs or using only a portion of a CCD, can reduce the sample time to 0.2 seconds. Continuously clocking of a CCD can reduce the sample time to 3 milliseconds but one loses a spatial dimension of the image. The pixel size is 24 microns resulting in an image with a sampling every 0.492 arc second, about the same as the point spread of the HRMA.

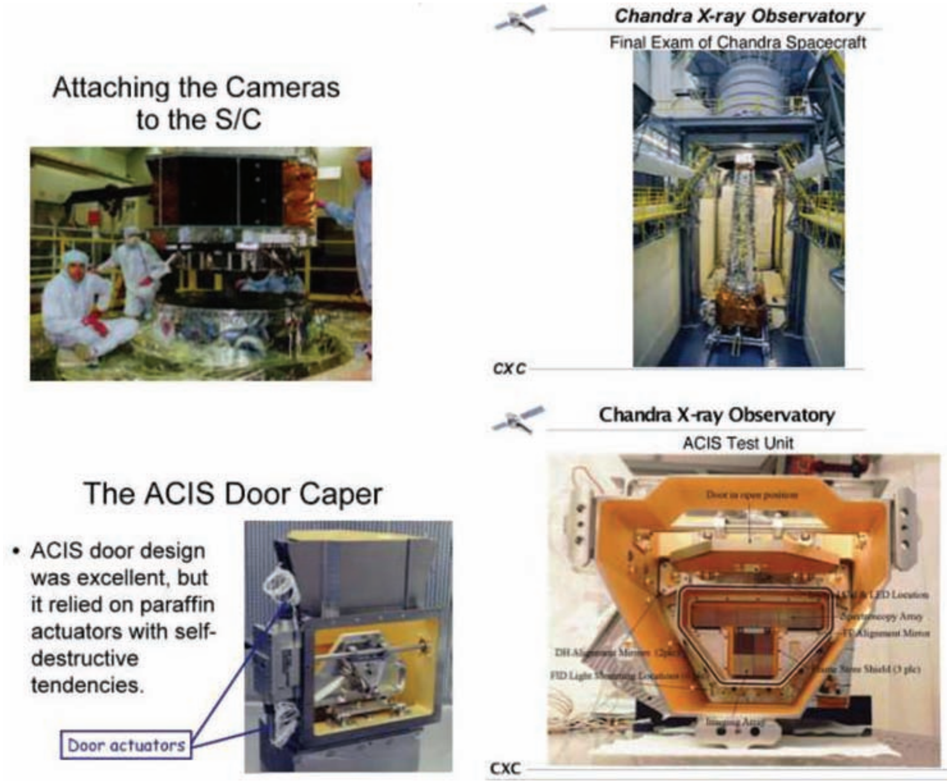
Several problems developed during the construction of the ACIS instrument. During the assembly of the flight unit focal plane array, it was discovered that the flex prints that carry the voltages and clocking signals to the CCD and the data stream from the CCDs were failing. The printed-through holes in the circuit boards attached to the CCDs were cracking, creating open circuits upon thermal cycling. This was a major problem, since the flight unit used these boards and they would have to be debonded from the CCDs and replaced. A company was found called Speedy Circuits that could quickly supply new boards that could survive the thermal cycling. The company said you can have two of the following three choices: quick delivery; reliable units; or low cost. Obviously at this point we opted for the first two choices.

6. Testing

The delays caused by the flex print problem made it impossible to provide the finished camera in time for the calibration of the instruments at the Marshall Space Flight Center in Huntsville, Alabama which was scheduled in March through May of 1997. ACIS provided a “two chip” camera for the calibration to at least see how the CCDs would perform in the focal plane of the HRMA. The full ACIS arrived in Huntsville after the HRMA was taken to TRW for integration into the spacecraft. This calibration in the facility did provide some useful information about the CCDs in a calibrated X-ray beam. The setup at the calibration facility is shown in Figure 6. The ACIS instrument travelled to Ball Brothers Research Center in Boulder, CO next for integration onto the translation stage that carries the two cameras into the focal plane of the HRMA. The translation stage is shown in Figure 6 (lower right).

The next step was to integrate the translation table and instruments with the telescope and spacecraft at TRW (Figure 7). The next major problem encountered, besides the difficulty of producing high quality CCDs, especially the back illuminated versions, was encountered during the vacuum test of the full Observatory at TRW (Figure 7, upper right). In order to verify that there were no light leaks that might degrade the CCD operation on orbit, the protective vacuum sealed door covering the CCDs and filter had to be opened in the vacuum chamber and lights shown onto the spacecraft to simulate solar, lunar and Earth shine illumination. When the command was given to open the door, the mechanism failed and the door did not open. There was no way to open the door with this kind of failure. The only remedy for this problem was to remove ACIS from the Observatory and return the unit to Lockheed Martin Aerospace Corp., where the door was designed and fabricated, for testing and redesign as needed. After extensive testing, no failure mechanism was found that could cause the door to stick shut. Some redesign of the opening mechanism permitted an evaluation of the door opening process so that it might be possible to try opening the door without breaking the opening mechanism and thereby seek solutions to the sticking should it occur on orbit. Thankfully, the door opened without a problem on orbit, but we still do not know what caused the failure at the TRW test. The ACIS door is shown in Figure 7.

In making a complex observatory there are literally thousands of people involved. I cannot give credit to all of them in the space here, but I do want to mention all of the people involved in the ACIS experiment. They are given in Figure 8 (left panel).



Attaching the Cameras to the S/C

Chandra X-ray Observatory

Final Exam of Chandra Spacecraft

The ACIS Door Caper

- ACIS door design was excellent, but it relied on paraffin actuators with self-destructive tendencies.

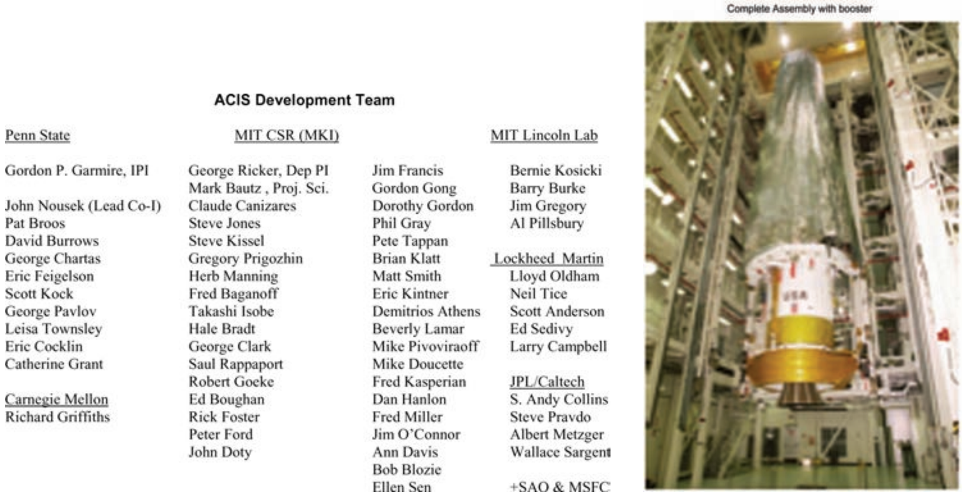
Chandra X-ray Observatory

ACIS Test Unit

Figure 7. Upper left: Mating the translation table assembly to the spacecraft at TRW (http://chandra.harvard.edu/graphics/ resources/illustrations/obs_assemb7_72.jpg); upper right: The Observatory ready to be lifted into the vacuum chamber for the thermal vacuum testing (http://chandra.harvard.edu/graphics/resources/illustrations/chandaFinalExam-72.jpg); lower left: The ACIS door and door opening mechanism. The horizontal shaft rotates to pull the door into this position, which is the open position. The opposite rotation closes the door against the o-ring seal shown in the lower right panel (Courtesy of Mark Bautz, MIT); lower right: The ACIS camera showing the door, CCDs and the o-ring seal. The camera is under vacuum at launch to protect the thin optical blocking filters from the acoustic load generated by the launch vehicle (http://chandra.harvard.edu/graphics/resources/illustrations/ACISlabel-72.jpg).

7. Launch!

The Chandra X-ray Observatory finally arrived at the Kennedy Space Center in the spring of 1999. Figure 8 (right panel) shows the full Observatory with the rocket booster attached that will send it into a highly elliptical orbit. The onboard rocket, which is part of the spacecraft, will then increase the orbital altitude and circularize the orbit. Eileen Collins was the Mission Commander, the first woman to assume this role. After two unsuccessful launch attempts the Space Shuttle Columbia roars into space with the Chandra X-ray Observatory



ACIS Development Team

Penn State

Gordon P. Garmire, IPI

John Nousek (Lead Co-I)
Pat Broos
David Burrows
George Chartas
Eric Feigelson
Scott Kock
George Pavlov
Leisa Townsley
Eric Cocklin
Catherine Grant

Carnegie Mellon
Richard Griffiths

MIT CSR (MKI)

George Ricker, Dep PI
Mark Bautz , Proj. Sci.
Claude Canizares
Steve Jones
Steve Kissel
Gregory Prigozhin
Herb Manning
Fred Baganoff
Takashi Isobe
Hale Bradt
George Clark
Saul Rappaport
Robert Goeke
Ed Boughan
Rick Foster
Peter Ford
John Doty

Jim Francis
Gordon Gong
Dorothy Gordon
Phil Gray
Pete Tappan
Brian Klatt
Matt Smith
Eric Kintner
Demitrios Athens
Beverly Lamar
Mike Pivoviraoff
Mike Doucette
Fred Kasperian
Dan Hanlon
Fred Miller
Jim O'Connor
Ann Davis
Bob Blozie
Ellen Sen

MIT Lincoln Lab

Bernie Kosicki
Barry Burke
Jim Gregory
Al Pillsbury

Lockheed Martin
Lloyd Oldham
Neil Tice
Scott Anderson
Ed Sedivy
Larry Campbell

JPL/Caltech
S. Andy Collins
Steve Pravdo
Albert Metzger
Wallace Sargent

+SAO & MSFC

Figure 8. Left: The ACIS development team and the most of the original co-investigators; right: The complete Chandra X-ray Observatory with the booster rocket attached in the hanger at KSC (http://chandra.harvard.edu/graphics/resources/illustrations/99pp0704-72.jpg).

on board. The launch turned out to be a nail biter. Quoted below are the mission notes from NASA.

"During the countdown for launch on the third attempt, a communications problem occurred that resulted in the loss of the forward link to Columbia. The problem was corrected at the Merritt Island Launch Area (MILA) ground facility and communications was restored. As a result of this problem, the time of the planned launch was slipped seven minutes to 12:31 a.m. EDT on July 23.

About 5 seconds after liftoff, flight controllers noted a voltage drop on one of the shuttle's electrical buses. Because of this voltage drop, one of two redundant main engine controllers on two of the three engines shut down. The redundant controllers on those two engines — centre and right main engines — functioned normally, allowing them to fully support Columbia's climb to orbit.

The orbit attained, however, was 7 miles short of that originally projected due to premature main engine cutoff an instant before the scheduled cutoff. This problem was eventually traced to a hydrogen leak in the No. 3 main engine nozzle. The leak was caused when a liquid oxygen post pin came out of the main injector during main engine ignition, striking the hotwall of the nozzle and rupturing three liquid hydrogen coolant tubes.

The orbiter eventually attained its proper altitude and successfully deployed the Chandra X-ray Observatory into its desired orbit."[1]

After the Space Shuttle achieved orbit, the bay doors were opened and the Chandra X-ray Observatory with its attached booster rocket was ejected from the payload bay. Figure 9 (left panel) shows the last views of the Observatory as it drifts off in preparation for the

[1] http://www.nasa.gov/mission_pages/shuttle/shuttlemission/archives/sts-93.html

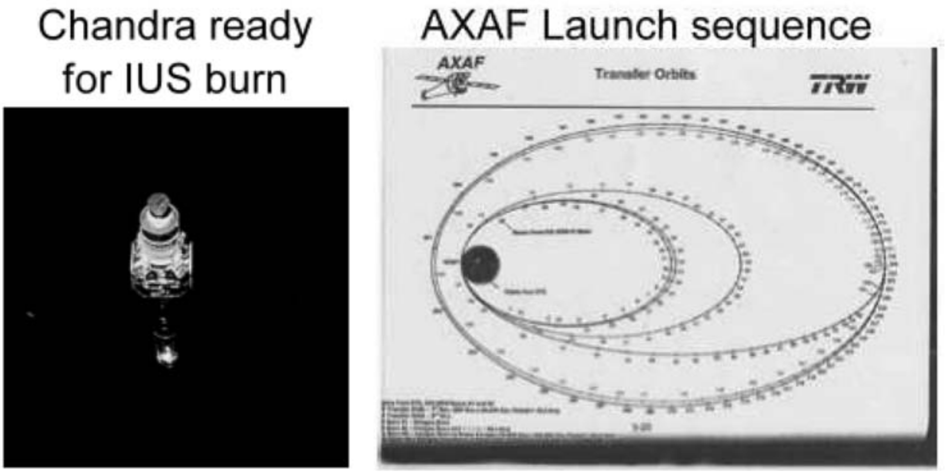


Figure 9. Left: The view of Chandra and its attached booster as it drifts away from STS-93. (http://chandra.harvard.edu/graphics/resources/illustrations/deploy/sts93-deploy1-72.jpg); right: The orbital insertion sequence following the STS-93 launch and ejection of the Observatory from the Space Shuttle (TRW document).

Figure 10. Left: The ACIS team feeling much relief now that the ACIS door was open including yours truly (Courtesy of Mark Bautz, MIT); right: The Project Scientist, Martin Weisskopf, pointing and the ACIS PI, Gordon Garmire, watching the data from Leon X-1 being acquired as the first image of an X-ray source viewed by the Chandra X-ray Observatory (http://chandra.harvard.edu/graphics/resources/illustrations/occ/group/group7-721.jpg).

insertion into a high elliptical orbit. It took almost another week before the Observatory reached its final orbit of 10,000 by 138,000 km, requiring five different burn sequences using the onboard rocket. Each burn was a source of worry. The orbital insertion sequence is shown in Figure 9 (right panel). Once the final orbit was achieved, the activation of the spacecraft and instruments followed. The moment that the ACIS Team was waiting for occurred on August 12th, when the ACIS door was finally opened without a hitch. Figure 10 shows the relief of the some of the team members.

The next step was to verify that the telescope was working and focusing X-rays. After suitable guide stars were located, Chandra began to accumulate data on the ACIS camera back illuminated CCD. After a few tense moments an image began to appear, somewhat off-axis but in reasonably good focus. The telescope scientist, Leon van Speybroeck, breathed a huge sigh of relief at this point and said in his low-key manner, “At least we don’t have a pile of glass at the bottom of the telescope tube!” The project scientist and the ACIS PI are shown in Figure 10 watching the data come in on the source we called Leon X-1 in honour of the Telescope Scientist.

8. Results

Next I would like to show just a few images from the past ten years of observations. The Observatory and instruments have performed essentially flawlessly for this time span. There was a brief period at the very beginning of the mission when ACIS was exposed to protons from the trapped radiation belts that were scattered by the telescope onto the CCDs. This caused radiation damage to the front illuminated CCDs, reducing their ability to transfer charge, but by placing ACIS out of the telescope focal plane during the radiation passages, further damage could be avoided. The back illuminated CCDs are protected by 40 microns of silicon before the protons could reach the transfer portion of the CCD. This was enough shielding to prevent damage to these devices. Data analysis techniques have been developed by the ACIS team to partially mitigate this problem in the front illuminated CCDs (Townsley *et al.* 2000).

The closest massive black hole is the one at the Galactic Centre associated with the radio source Sgr A*. This object is found to emit X-rays, but at a very low level of about 2×10^{33} erg/s, which is some nine orders of magnitude below typical AGN activity (see Figure 1). This may be the result of a supernova remnant that has engulfed the black hole and its environs, thereby making accretion difficult (see Maeda *et al.* 2002). The X-ray source has been observed to flare on a daily basis, increasing in intensity by nearly two orders of magnitude for a period of order an hour, then falling rapidly back to its quiescent level (Baganoff *et al.* 2001). The cause of the flares is not known. I’m sure Chandra would have been interested in this phenomenon.

Another object that exhibits relativistic plasma phenomenon is the Crab Nebula (Figure 11). In a time laps image of the Crab Nebula, some of the wisps are seen to move at velocities as high as 0.5c (Hester 2008). The pulsar is clearly the centre of the wisp activity.

Another supernova remnant containing a pulsar is G292.0 +1.8. This is one of the few oxygen rich remnants in a nearby galaxy and shows clumps of gas rich in Mg, Si and S (Park *et al.* 2007). A strong shock front can be seen along some of the outer perimeter of the remnant. The pulsar is in the blue nebulosity to the southwest portion of the nebula in Figure 11 (upper right).

Another supernova remnant, RCW 103, is shown in Figure 11 (lower left). This remnant was the first SNR with a pulsar located at the very centre of the nearly circular nebula (Tuohy & Garmire 1981). The pulsar has been found to be the slowest rotating neutron star with a period of 6.67 hr (De Luca *et al.* 2006). It is likely to be in a binary system with the same period (Pizzolato *et al.* 2008) or a magnetar with a fall-back disk (Li 2007). This pulsar experienced a large outburst in 2000, increasing in luminosity by a factor of 100,

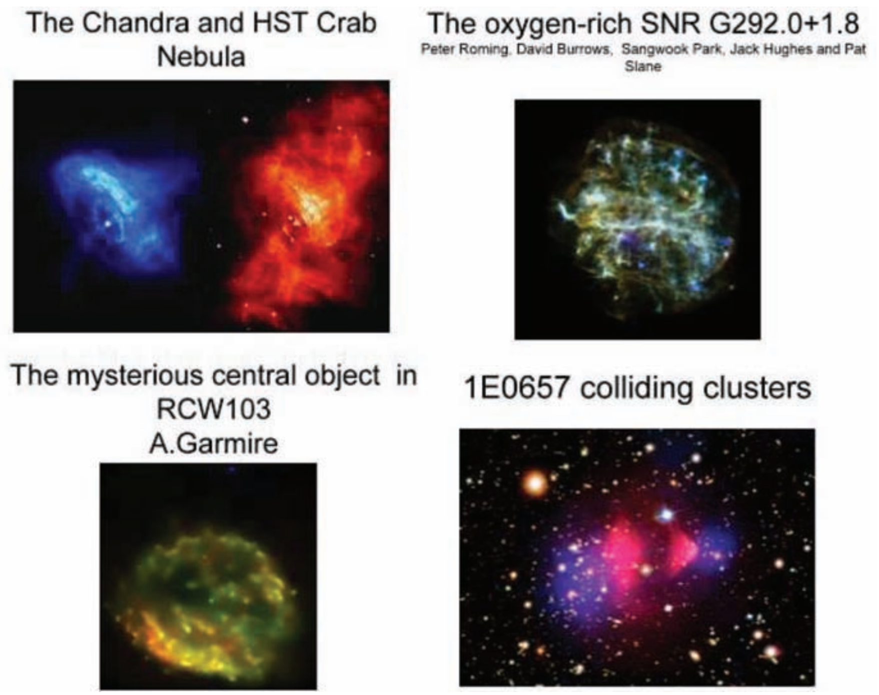

The Chandra and HST Crab Nebula
The oxygen-rich SNR G292.0+1.8
Peter Roming, David Burrows, Sangwook Park, Jack Hughes and Pat Slane
The mysterious central object in RCW103
A.Garmire
1E0657 colliding clusters

Figure 11. Upper left: The Crab Nebula as viewed by Chandra (left) and Hubble (right); (Courtesy of David Burrows, Penn State); upper right: the oxygen-rich SNR G292.0+1.8. The pulsar is located just south of the central bar in the blue nebulosity (courtesy of Peter Roming, Penn State); lower left: The supernova remnant RCW 103 with a central pulsar of extremely slow rotation period of 6.67 hr (courtesy of Audrey Garmire, Penn State); lower right: the colliding clusters of galaxies, 1E0657 (http://chandra.harvard.edu/photo/1e0657/1e0657.jpg).

then decaying very slowly over the next seven years (Garmire *et al.* 2000). No optical or IR candidate has conclusively been found for this object (De Luca *et al.* 2008).

The last object I wish to show is the colliding clusters of galaxies 1E0657 (Figure 11, lower right panel). This remarkable collision reveals that the dark matter (which does not interact with baryons and is traced by gravitational lensing) follows the galaxies through the collision process (the blue clouds), while the hot plasma shows strong interaction (pink clouds). This collision has been used as strong evidence for the presence of dark matter in the clusters as opposed to modified gravity to explain the velocity dispersion of galaxies and the confinement of the hot plasma found in clusters of galaxies (Clowe *et al.* 2006).

The naming of the Chandra X-ray Observatory was the result of a contest conducted by NASA and open to the world. There were more than 6000 entries from 61 countries. The winners were a high school student from Idaho, Tyrel Johnson and a high school teacher from California, Jatila van der Veen. These two submitted the winning essays that selected Chandra in honour of Subrahmanyan Chandrasekhar as the name of this ‘Great Observatory’.

Acknowledgments

I would like to thank the many, many scientists, engineers, technicians and managers who contributed to the wonderful success of this Great Observatory, the Chandra X-ray Observatory. In particular, I would like to thank Dr. Mark Bautz of the Kavli Institue for Astrophysics at MIT and Dr. Robert Burke of the Northrop Grumman Aerospace Systems group for several images which they provided. For much more detail concerning the history and development of the Observatory, the reader is referred to the books *Revealing the Universe: the making of the Chandra X-ray Observatory* by W. and K. Tucker, and *Secrets of the Hoary Deep* by R. Giacconi.

References

Baganoff F.K., *et al.*, 2001, Nature, 413, 45

Clowe D., Bradac M., Gonzalez A.H., Markevitch M., Randall S.W., Jones C., Zaritsky D., 2006, ApJ, 648, L109

De Luca A., Caraveo P.A., Mereghetti S., Tiengo A., Bignami G.F., 2006, Science, 313, 814

De Luca A., Mignani R.P., Zaggia S., Beccari G., Mereghetti S., Caraveo P.A., Bignami G.F., 2008, ApJ, 682, 1185

Garmire G.P., Pavlov G.G., Garmire A.B., Zavlin V.E., 2000, IAUC, 7350, 2

Giacconi G., Rossi B., 1960, JGR, 65, 773

Giacconi R., Gursky H., Paolini F.R., Rossi B.B., 1962, PhRvL, 9, 439

Hester J.J., 2008, ARA&A, 46, 127

Li X.-D., 2007, ApJ, 666, L81

Maeda Y., *et al.*, 2002, ApJ, 570, 671

Muno M.P., *et al.*, 2003, ApJ, 589, 225

Park, S., Hughes J.P., Slane P.O., Burrows D.N., Gaensler B.M., Ghavamian P., 2007, ApJ, 670, L121

Pizzolato F., Colpi M., De Luca A., Mereghetti S., Tiengo A., 2008, ApJ, 681, 530

Townsley L.K., Broos P.S., Garmire G.P., Nousek J.A., 2000, ApJ, 534, L139

Tuohy I., Garmire G., 1980, ApJ, 239, L107

Some memories of Chandra[*]

Robert M. Wald[†]

Department of Physics, University of Chicago, Chicago, USA

Abstract. Five noted scientists, all close colleagues and friends of Subrahmanyan Chandrasekhar, share thoughts and memories of the man whose centennial we celebrate.

Chandra was particularly intolerant of scientists motivated primarily by the hope of receiving recognition from others rather than by a deep, inner conviction that their work was of importance and interest. — Robert Wald

I first met Subrahmanyan Chandrasekhar in December 1972, but did not get to know him well until early 1976, more than a year after I arrived at the University of Chicago as a postdoc in the relativity group. For nearly 20 years after that, until his death in 1995, we interacted on an almost daily basis. My memories of those conversations and interactions have faded considerably over the past 15 years — I simply do not have Chandra's remarkable ability to recall all details of events that occurred long ago. However, the overall impression that Chandra left on me and many other scientists is something that will never fade away.

To many who met him but did not get to know him well, Chandra must have seemed an exceptionally austere and formidable figure — an impression with a great deal of validity.

[*] Reproduced with permission from *Physics Today*. © 2010 American Institute of Physics.

[†] Robert Wald is the Charles H. Swift Distinguished Service Professor.

Of all the scientists I have met, Chandra had the highest standards for both intellectual rigor and personal integrity. He applied those standards most uncompromisingly to himself, but he also did not tolerate failings by others in such matters. He was particularly intolerant of scientists motivated primarily by the hope of receiving recognition from others rather than by a deep, inner conviction that their work was of importance and interest, whatever anyone else might think. He was equally intolerant of scientists who rested on their laurels or were otherwise lazy or sloppy, rather than applying their full intellectual efforts toward their work. It was not unusual for Chandra to ask questions of a seminar speaker that were aimed at discerning the speaker's convictions or at probing how carefully the speaker had thought through the relevant issues. Often those were uncomfortable moments for the speaker.

To get to know Chandra well, a barrier first had to be crossed, a barrier undoubtedly enhanced by the man's impeccable dress — a suit and tie on all occasions — and by his impeccable speech and manners. It is unfortunate that this barrier had the effect of isolating him from a portion of the scientific community. I believe all that was needed to cross the barrier was some expression to him of the depth of one's passion for research or other intellectual endeavors. With the barrier crossed, the very sensitive, caring, and above all loyal nature of Chandra's personality would become readily apparent. The combination of those very human qualities with Chandra's almost superhuman discipline, self-sacrifice, and dedication to science had a profound and lasting effect on all who knew him.

In his scientific career of more than 65 years, Chandra's enthusiasm for the pursuit of science never declined, nor did his fortitude in carrying out major projects. I do not recall a single instance in which he appeared to be motivated by personal gain, nor a single occasion when he made an excuse for not doing something he felt should be done. If he thought a visit to a collaborator or other scientist would help advance his research, he would make the visit without seeking reimbursement for his travel expenses. Similarly, he never requested summer salary from his NSF grant. It appears that the free pursuit of his own scientific research was so important to Chandra that he did not want it tainted or encumbered with issues involving personal gain or accountability.

Chandra will be remembered for the next hundred years and beyond primarily for his truly major contributions to a remarkably broad range of areas in physics and astronomy. He ensured that his scientific legacy will pass on to future generations in unadulterated form by writing a definitive monograph on each of the topics on which he worked. It is highly appropriate that Chandra be remembered primarily for his scientific work. But it also is important that he be remembered for his personal qualities.

To convey a more complete picture of what Chandra was like as a person, I present four reminiscences from scientists who knew him well. John Friedman, professor of physics at the University of Wisconsin–Milwaukee, was one of Chandra's last students and closely collaborated with him in the early 1970s. Abhay Ashtekar, Eberly Professor of Physics at the Pennsylvania State University, was a student in the Chicago relativity group in the early 1970s, a postdoc in the group in the late 1970s, and a close friend of Chandra's thereafter. Valeria Ferrari, Professor of Physics at the University of Rome I ("La Sapienza"), was Chandra's closest collaborator during the last 10 years of his life. Roger Penrose, Emeritus Rouse Ball Professor of Mathematics at Oxford University, was someone whose research Chandra particularly admired and whose scientific advice Chandra sought when he encountered particularly challenging problems. The excerpts below were written about a year after Chandra's death and are taken from *S. Chandrasekhar*: *The Man Behind the Legend*, edited by Kameshwar C. Wali (Imperial College Press, 1997). They are reprinted here with the permission of the publisher and the authors.

Chandra's meticulous script was as elegant as ever, lengthy error-free art, ink on bond. He smiled with mischievous pleasure that I had also been working by candle. — John Friedman

John Friedman

Despite the fact that he was still sole editor of *The Astrophysical Journal*, Chandra spent as much time on research as did his most dedicated students. Beginning his work by 5 am, he finished each 13-hour workday late in the evening. As part of his moral instruction to us, Chandra did not hesitate to point out that by the time his colleagues arrived in the morning, he had already put in half as many hours as they would work in a day. He described a visit to Caltech mainly by noting that the physicists had spent several evenings during the week at cocktail parties. How, he asked, could they get anything done if this was the way they lived? If a few supremely talented physicists could afford such lapses, Chandra placed himself (and, of course, us) among that vast majority for whom success in science was a matter of character. . . .

In my last year of graduate work, Chandra and [his wife] Lalitha were scheduled to spend six months at Oxford, and Chandra asked me to come with him to finish up my thesis work, a collaboration with him on the stability of rapidly rotating "configurations," none of which had, at that time, been observed. [My wife] Paula, Mack (our six-month old son), and I traveled to Oxford in time for the great blackout of '72, one of the miners' strikes.

In the darkness of that winter, when Chandra went home to his apartment with Lalitha and I to the row house we rented from the Rev. Gauntlett of Maid Marion Way in Nottingham, we worked by candlelight . . . . It was dim and as damp as England's winters have always been. I might have been feeling a little down myself, tired from our son's cries and straining to check equations in the dark. But when I came in to work, Chandra's meticulous script was as elegant as ever, lengthy error-free art, ink on bond. He smiled with mischievous pleasure that I had also been working by candle. Amid 13th century stone walls, built to sequester from the town a secular clergy that once comprised Oxford, he was obviously proud that we each had again spent a day and an evening showing our devotion. It was, he said, as if we were medieval scribes.

The beautiful hand in which his equations were written mirrored Chandra's understanding of the equations themselves. For most physicists on the mathematical side, equations are viewed abstractly in a way that highlights the properties their expressions share as operators on a Hilbert space, while astrophysicists usually take from mathematics only what is needed for the problem at hand. Chandra, however, fell in neither camp. For his time, Chandra was, to my knowledge, unique in the way he treated the equations of relativistic astrophysics seriously as objects in themselves, their structure clear in the manner he displayed them, their meaning to be found in this structure. That mathematics was the language of nature he never doubted, and he served nature all his life.

Chandra was also unique in the way he combined a deep understanding of classical mathematics, of astrophysics and of the history of science, particularly the history of classical physics and astronomy. [Andrzej] Trautman and Roger Penrose were then the physicists to whom Chandra seemed closest in temperament and perspective, while his interests were closest to those of the astrophysical relativists, Kip Thorne and James Bardeen. The understanding that grew from Chandra's history distinguished the problems he worked on, and the unmatched artistry with which he handled his language of equations distinguished their solutions. He was as devoted to science as anyone I have ever met.

Chandra got up spontaneously and told some wonderful ghost stories — one told to him by Dirac! They were short, dry and crisp and we all gasped when the punch line came and then laughed. — Abhay Ashtekar

Abhay Ashtekar

I first met Chandra when I arrived at the University of Chicago as a green graduate student in '71. He had just turned sixty. I had done my undergraduate work in India and to me — as to most other Indian students in science — Chandra's stature was god-like. We had heard of the innumerable discoveries he had made whose meaning and scope we understood only in the vaguest terms. But there was a feeling of awe and admiration and a conviction that for a single person to accomplish all this, he had to be superhuman. And so, I was very surprised when I first met him. Yes, he did have that pristine air about him, and yes, everything he did — the way he dressed, the way he sat in seminars, even the hard-backed chairs he chose to sit on — everything had an aura about it that set him apart. One immediately sensed a refined, dignified and austere personality, just of the type one would expect of a legendary figure like him. Yet. when it came to science, there was unexpected openness. He treated us, students in the newly formed relativity group at Chicago, as if we were his colleagues, his equals. He would come to all seminars, including the ones given by students. He would ask us technical questions with genuine interest. When discussions began, he seemed to become genuinely young, almost one of us. I still remember the smile that would light up his face in the middle of a talk when he heard a beautiful result. Sometimes, when he had cracked a hard problem, something that he found truly satisfying, he would tell us about it. The joy he experienced was so manifest and so contagious! . . .

Chandra was a master storyteller; I have yet to encounter his equal. He had such a fantastic memory for dates and details that, in the anecdotes he recounted, everything became alive. And his anecdotes ranged from incidents that took place in the lofty halls of the Trinity College in Cambridge to his small cabin in the ship he took across the North Sea when he went to Russia. He would recount the events as if they had happened yesterday. We would later shake our heads in astonishment. For, here was Chandra telling about a storm he encountered during the North Sea passage in 1934, or his interesting meetings with the then President of the University of Chicago in 1946, with such clarity and in such detail that we could not have matched in describing events that took place in our own lives just a year before!

I still vividly recall the first time that I heard him tell a story. The students and postdocs in the relativity group had organized a potluck dinner. Chandra and his wife Lalitha came with a delicious vegetarian casserole. When it came to coffee time, there was some unease about how the event was going to end. Do we just say good-bye and leave? Students had planned the menu well but hadn’t thought of anything specific as an after-dinner activity. So, there was some unease. Chandra got up spontaneously and told us some wonderful ghost stories — one told to him by Dirac! They were short, dry and crisp and we all gasped when the punch line came and then laughed. Then other people got up to tell other stories and the evening ended in a relaxed and friendly mood.

Valeria Ferrari

My collaboration with S. Chandrasekhar started in October of 1983. We had met in Rome after the X International Conference on General Relativity, held in Padova in the summer of 1983, and he had invited me to work with him on some relations existing between the mathematical theory of black holes and exact solutions of Einstein’s equations possessing two space-like Killing vectors.

I arrived in Chicago a few days after he had been awarded the Nobel Prize. I was afraid that the commitments associated with such an important event would prevent Chandra from working with me. But my fears were unwarranted, because he was more interested in the work we were doing than in giving interviews to the press. Our first paper was completed in two weeks.

For me, this first interaction with Chandra was surprising in many respects. Knowing the breadth and wide range of his scientific accomplishments and having listened to his

Chandra turned out to be entirely different from my preconception. In our work, for example, he never used his authority to impose his view on a subject; we always discussed and confronted our ideas as if we were on the same footing. — Valeria Ferrari

lectures at conferences, I had nurtured the idea that he was very strict and rigorous, a man totally and exclusively dedicated to science, and so overwhelming that it would be difficult for me even to talk to him. But Chandra turned out to be entirely different from my preconceptions. In our work, for example, he never used his authority to impose his view on a subject; we always discussed and confronted our ideas as if we were on the same footing. At the same time he was an extraordinary teacher, and shared with me his knowledge and the secrets of his technical ability.

I had to change my views also about Chandra's personality. In spite of his strict appearance, he was a very warm person, to whom friendship was of great importance. Although I came to know him only during the last twelve years of his life, from many episodes that he narrated to me I think that this had always been the case. For example, in remembering [Arthur] Eddington, with whom he had had the famous scientific dispute that strongly affected his life and his career, he never expressed feelings of resentment or disrespect. I was surprised to learn that while Eddington attacked Chandra's work in international conferences (he characterized the theory of the limiting mass for the white dwarfs [as] "a stellar buffoonery"), in private they remained on good terms, joining for tea or for a bicycle ride. Chandra was convinced that Eddington's opposition to his theory was motivated by honest scientific disagreement, and his enormous respect, admiration and affection for him were unharmed by these events. At that time Chandra was in his mid-twenties. Chandra told me that when he used to see Eddington walking the streets of Cambridge with an umbrella under his arm, he thought that this was the picture of a man who had dedicated his life to the pursuit of science and finally had reached a sense of harmony and contentment. Thinking of his own future, he would think that he would also experience a similar sense of harmony, peace and contentment in his old age. "But," he would add, "it hasn't turned out that way." He had a feeling of disappointment because the hope for contentment and a peaceful outlook on life as a result of single-minded pursuit of science had remained unfulfilled. I used to wonder, how could a man like Chandra have this feeling of discontentment about his life? Chandra did not exactly know the reason himself. However, I used to feel a sense of relief in seeing that the excitation for a new result, or the occurrence of a problem difficult to solve, was always able to divert his mind from these sad thoughts.

It is almost as though he had made a tactical retreat, circling around and exploring the details of the surrounding terrain — stellar dynamics, radioactive transfer, and the stability of various types of astrophysical structures — before he felt ready for an assault on the profound issue that his early work had uncovered. — Roger Penrose

Roger Penrose

This world has seen some scientists of extraordinary ability — some who are quick and often arrogant, others cautious and possessing genuine humility. Among that small proportion who are of real and rare distinction are the very few who are truly great. It has been my considerable good fortune to have made the acquaintance of some four or five of those that fall into this final category, but only one of them could I claim to have known at all well — Subrahmanyan Chandrasekhar . . . .

My acquaintance with Chandra dates back to1962, when I first encountered him at the Warsaw International Conference on General Relativity and Gravitation. That occasion had a particular significance for Chandra with regard to general relativity, as it marked his entry into the world of general relativists. In fact, he attended that meeting as a "student," as his way to acquaint himself best with the current activity in that subject.

Why did Chandra have such determination, at the age of 51, to break entirely into a new field, demanding the learning of many new concepts and techniques, where much of the vast expertise that he had built up over many decades would have little direct relevance? It would be natural to suppose, and as I would strongly suspect myself, that it was his desire finally to address the profound conundrum that his early work had thrown up, dating back to his calculations in 1930 on the boat from India to England — that white dwarf stars of more than about one and one-half solar masses cannot sustain themselves against gravitational collapse. It seems clear that even at that time, Chandra was basically aware of the awesome implications of this conclusion, namely that the collapse of the star must eventually take it out of the realm of known physics and into an area shrouded in puzzlement and mystery. But he was by nature an extremely cautious individual, as is made manifest in the modest way he stated his conclusion:

> The life-history of a star of small mass must be essentially different from the life-history of a star with large mass. For a star of small mass the natural white-dwarf stage is an initial step towards complete extinction. A star of large mass cannot pass into the white-dwarf stage and one is left speculating on other possibilities.

He was not the sort who would attempt, without due preparation, to make "authoritative" assessments of the likely fate of the material of a body indulging in gravitational collapse. There are, indeed, still many possible loopholes in the arguments which lead to the final conclusion that has now become an accepted implication of present-day theory — that, at least in some cases, the fate of a body in gravitational collapse must be to encounter a space-time singularity, representing, for the constituents of that body, an end to time!

The issue had been at the root of his difficulties with Eddington, when Eddington had so unfairly attacked his work at a meeting of the Royal Astronomical Society in 1935. Eddington, also, was aware of the implications of Chandra's findings, but regarded this as a *reductio ad absurdum* and preferred to move along his own highly speculative route towards a fundamental theory, thereby rejecting the sound reasoning within the accepted tenets of procedure that had characterized what Chandra had achieved. Chandra appears to have been deeply hurt by Eddington's reaction — the reaction of a man whom Chandra had previously so admired and looked up to. In response, Chandra turned his back on Cambridge and on the immediate problems thrown up by the structure of white dwarfs, apparently devoting his attention entirely to other problems. Yet the question of the ultimate fate of a gravitationally collapsing body must have continued to nag at his physical understandings for many intervening years — even while he was engaging in thorough studies of matters pertaining to quite other astrophysical questions. It is almost as though he had made a tactical retreat, circling around and exploring the details of the surrounding terrain — stellar dynamics, radiative transfer, and the stability of various types of astrophysical structures — before he felt ready for an assault on the profound issue that his early work had uncovered.

His assault was carefully prepared, and required many years of study of the intricacies of Einstein's general relativity. Not only did he familiarize himself with the standard mathematical techniques and conceptual notions that had been developed for that subject over the years, but he engaged the assistance of certain relativists, such as Andrzej Trautman (and even myself), who had specialist knowledge of some of the less familiar modern mathematical procedures, to give a series of lectures in Chicago to him, his coworkers, and students.

Chandra's first contributions in which he was able to bring general relativity to bear on astrophysical questions showed that there were additional instabilities, beyond those of Newtonian theory, making their mark earlier than had been expected, and leading even more surely to the ultimate situation of a black-hole fate for a collapsing star. He then moved to the study of black holes themselves, and became fascinated by the beauty of these structures — particularly the Kerr geometry that pertains to a stationary rotating black hole, the ultimate configuration of gravitational collapse. He eventually referred to black holes, in the prologue to his epic book on the subject, *The Mathematical Theory of Black Holes,* as "the most perfect macroscopic objects that there are in the universe." ...

His fascination with black holes gained as much from aesthetics as from a desire to push forward the frontiers of scientific knowledge. In his later years Chandra became quite explicit as to the importance of aesthetic qualities in science and in his own work in particular.

This brings out what must surely be one of Chandra's very special qualities: his profound appreciation of the beauty of mathematical formulae. This appreciation extended into pure mathematics as well as applied, and he had an especial admiration for the work of Srinivasa Ramanujan. (He often expressed to me his delight in the fact that the only known photograph of Ramanujan was one that he had himself retrieved. Ramanujan had served as an important inspiration for Chandra in his early aspirations to become a scientist.) Chandra's wonderful way with mathematical formulae must have been a quality that benefited him also through his earlier work — and provided a thread of continuity throughout his scientific researches in various disparate fields of endeavor. However, this quality is particularly apparent in his work in relativity theory. No doubt he was struck by the fact that the closer his researches took him to fundamental issues in physics — in the analysis of the very nature of space-time — the greater was the mathematical elegance that he encountered in the equations . . . .

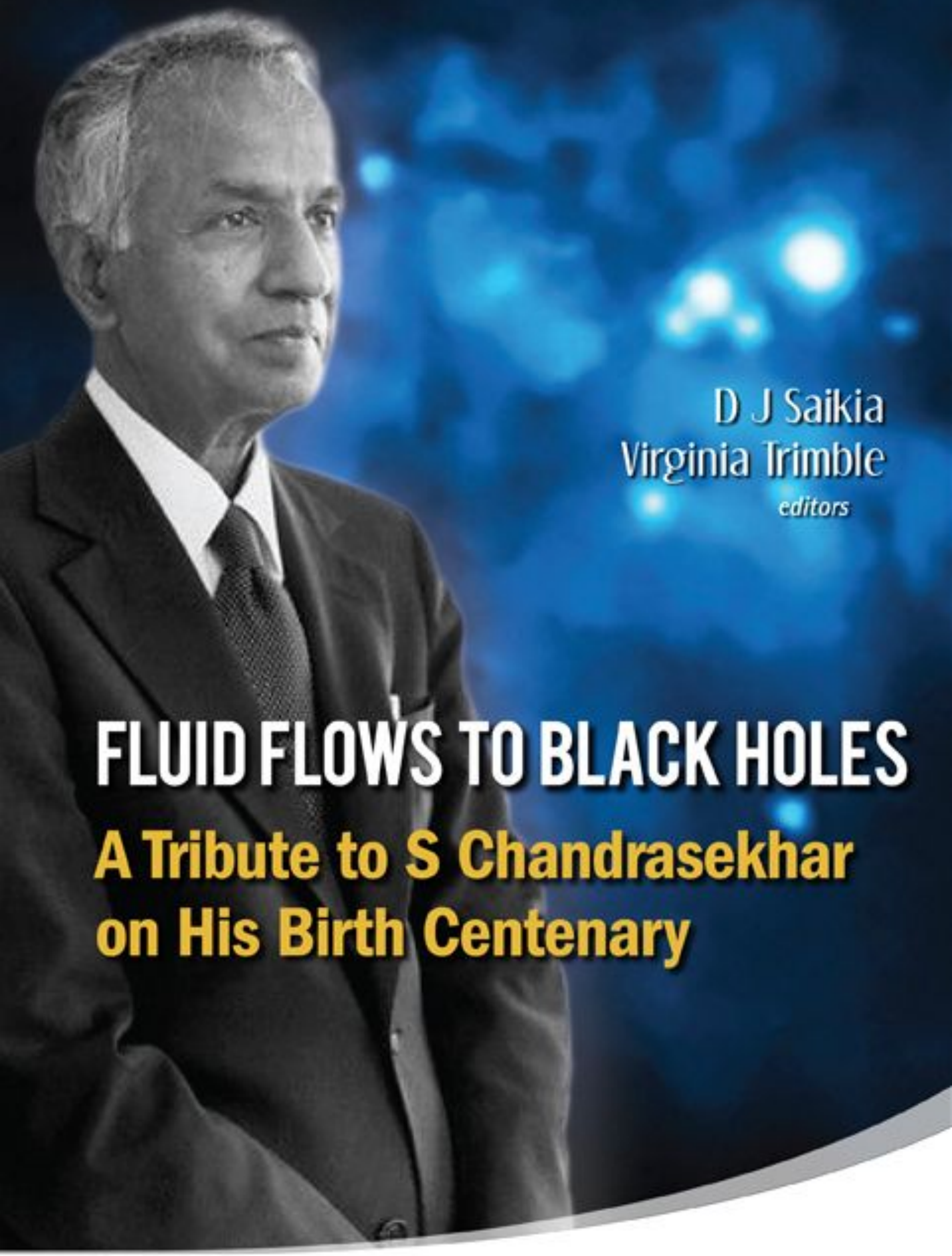
What are the qualities that stand out in my memories of Chandra? That he was a great and prolific scientist, there is no doubt, and a deeply individual original thinker. He was enormously systematic and well organized, and he worked incredibly hard. He was a rigorous and somewhat autocratic taskmaster, but he had a genuine appreciation of quality in others. He was a loyal friend, reliable, and totally honest. He was deeply sensitive, but proud. He was a difficult man to criticize, and on occasion his pride might get the better of him — but he would be scrupulously generous with his critics if he could be found to be in error. He was polite and enormously dignified: a greatly cultured individual with a feeling for what is valuable in humanity wherever it might be found. He respected life in all forms (he was a strict vegetarian) and had deep appreciation of the works of Nature. He particularly valued the arts and took great pleasure in them, perceiving profound links between artistic and scientific values. [See S. Chandrasekhar, Beauty and the quest for beauty in science, *Physics today*, 1979 July, pp. 25–30.]

How did he view the status of his own scientific contributions in relation to his initial aspirations? One recalls Chandra's distinctive way of working — reminiscent of the great mathematician David Hilbert — whereby (in essence) Chandra would devote different decades of his life to different topics, culminating each with a definitive book, and leaving each topic behind when he embarked on the next. What does one conclude from this? It might seem that these decades must have represented, to him, completed work that would be neatly wrapped up in the final book. Perhaps so; yet I detected a restlessness in him indicative of a dissatisfaction with what he had ever been able to achieve.

I suspect that his work in relativity theory was what brought him closest to the ultimate goals that he was striving for. He must have derived great satisfaction from his study of black holes, but there were always profound questions left open — and the more that were resolved, the more new ones would appear. Moreover, in his black-hole work, it was the vicinity of the horizon that was being studied, and this lay far outside the central region where the matter of the collapsing star would meet its fate. To gain insights into the nature of this region one must study the space-time singularities — where space and time themselves reach their final termination. Chandra's work on colliding plane waves must surely have been directed towards gaining an understanding of these singularities, for they provide specific models where one can examine the generation of singularities explicitly.

It is inevitable that the results of this work must remain inconclusive, despite the power and insights that Chandra and his associates were able to provide. If the problem of the ultimate fate of a collapsing star — or a collapsing universe — remains unresolved, it is no discredit to him. He opened our eyes to this profound and deeply important problem and he made great strides towards resolving it. Quite apart from all his other achievements, that in itself might be thought to be enough for any man.

Index



D J Saikia
Virginia Trimble
editors

FLUID FLOWS TO BLACK HOLES

A Tribute to S Chandrasekhar on His Birth Centenary

World Scientific