Beyond Pluto

Exploring the outer limits of the solar system

John Davies

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
10 Stamford Road, Oakleigh, VIC 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

© Cambridge University Press 2001

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2001

Printed in the United Kingdom at the University Press, Cambridge

Typeface Nimrod MT 9/14 pt *System* QuarkXPress™ [SE]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data
Davies, John Keith.
Beyond Pluto : exploring the outer limits of the solar system / John Keith Davies.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 80019 6
1. Kuiper Belt. I. Title.
QB695.D38 2001
523.2–dc21 00-049364

ISBN 0 521 80019 6 hardback

Contents

The edge of the solar system

Like the planets Pluto and Neptune, the existence of what is frequently called the Kuiper Belt was predicted theoretically long before it was actually observed. Probably the first fairly detailed speculation about a cometary ring beyond Neptune was put forward by Kenneth Essex Edgeworth in 1943. Edgeworth was an interesting character who had progressed from soldier and engineer to retired gentleman and amateur theoretical astronomer. He was born on 26th February 1880 in County Westmeath, Ireland, into a classic well to do literary and scientific family of that era. As a young man he joined the Royal Military Academy at Woolwich, England, and attained a commission in the Royal Engineers. He spent his next few years stationed around the world building bridges, barrack blocks and the like. With the outbreak of the First World War he served with the British Army in France as a communications specialist and was decorated with both the Distinguished Service Order and the Military Cross. He remained in the army until 1926 and then took up a position with the Sudanese department of Posts and Telegraphs in Khartoum. Edgeworth remained in the Sudan for five years before retiring to Ireland to live out the remainder of his life.

Although retired, Edgeworth was by no means inactive. During the 1930s he studied economic theory and published several books on this topic. Although never affiliated with a university or other astronomical institution, he also pursued an interest in astronomy which he had acquired in his youth (he had joined the Royal Astronomical Society in 1903). After he retired he wrote a number of articles, mostly theoretical in nature, dealing with the process of star formation and developing ideas about the origin of the solar system. In 1943 he joined the British Astronomical Association, whose journal published his first

paper on the evolution of the solar system that summer. It was a short note which Edgeworth himself described as containing 'Not so much a theory, but the outline of a theory with many gaps remaining to be filled'. In his paper he described how a cloud of interstellar gas and dust might collapse to form a disc. He speculated that within such a disc numerous local condensations of higher density might then develop and collapse upon themselves. Noting that the real solar system does not comprise a huge number of small objects, but rather a few large planets and moons, Edgeworth suggested that these condensations then coalesced to form the nine known planets and their satellites. Crucially, Edgeworth recognised that there was no obvious reason why the disc of planet-forming material should have been sharply bounded at the orbit of the outermost planet. He suspected that the disc probably extended far beyond this distance and reasoned that, so far from the Sun, the density of material in the disc would be

Figure 1.1 A caricature of Kenneth Edgeworth as a young man. Comparison with photographs of him in later life suggests that it is a good likeness. (Royal Signals Museum Archive.)

very low. So, although individual condensations of reasonable size might form beyond Neptune, there would be little likelihood of them encountering each other frequently enough to form large planets. He suggested instead that these condensations would simply collapse upon themselves to form a large number of small bodies. Echoing then current theories of comets as concentrated swarms of meteoroids he described these distant condensations as astronomical heaps of gravel. He added that perhaps from time to time one of these condensations 'Wanders from its own sphere and appears as an occasional visitor to the inner solar system'. Here was the genesis of the idea of a trans-Neptunian comet belt.

Edgeworth developed his ideas further, writing a longer paper along similar lines a few years later. This second paper was submitted to the *Monthly Notices of the Royal Astronomical Society* in June 1949. Although by today's standards it contained numerous poorly justified assumptions, it was accepted almost immediately and appeared in an issue of the journal dated late 1949. In this paper Edgeworth expanded on his model for the formation of the planets and once again mentioned the likely existence of a vast reservoir of potential comets beyond the orbit of Neptune.

About the same time as Edgeworth's musings, the Dutch-born astronomer Gerard Kuiper was also considering the existence of tiny worlds beyond Pluto. Kuiper was working at the Yerkes Observatory in Chicago and, in 1951, he wrote what became a classic book chapter summarising the state of knowledge about the solar system. Kuiper noted that the distribution of material in the outer solar system seemed to come to an unnaturally sharp edge in the region of the planet Neptune and that there was no obvious reason why this should be so. Perhaps taking a lead from newly published theories about the composition of comets, Kuiper suggested that during the formation of the planets many thousands of kilometre-sized 'snowballs' might have been formed in a disc beyond the planet Neptune. Like Edgeworth, Kuiper reasoned that at such great distances from the Sun, where the relatively tiny snowballs would occupy a huge volume of space, it was unlikely that these snowballs could come together to form large planets. He suggested that instead their orbits were disturbed by the gravitational influence of the planet Pluto[†] and they were either ejected into deep space or sent in towards the Sun to

[†] Pluto was then thought to be much more massive than we now know it to be.

appear as comets. However, in a world in which observational astronomy was still dominated by the photographic plate, the detection of such tiny objects remained impracticable.

Of course, speculation about missing planets is not a new phenomenon. Ever since William Herschel's discovery of Uranus in 1781, astronomers have been fascinated by the possibility that there might be other unknown worlds. On the 1st of January 1801 the Italian astronomer Giuseppe Piazzi made a chance discovery of what was at first thought to be a new planet. The object, which was soon shown to be orbiting between Mars and Jupiter, was named Ceres after the Roman goddess of the harvest. It was soon found that Ceres, even though it was quite close, did not show a detectable disc when viewed through a telescope. This suggested that it was smaller than any of the other known planets. Three similar objects, Pallas, Juno and Vesta, were found in 1802, 1804 and 1807 respectively. All appearing as slow-moving points of light, this group of new objects was referred to as asteroids (star-like) by William Herschel. All went quiet for a while until the mid 1840s when new asteroids began to be found in quite large numbers. By the end of 1851 fifteen of them had been found and we now know that Ceres is just the largest of many small rocky objects in what became known as the asteroid belt.

However, by the middle of the nineteenth century attention had once again swung to the outer solar system. Irregularities in the motion of Uranus hinted that it was being tugged by the gravitational pull of another, more distant world still waiting to be discovered. In a now classic story of astronomical detective work, the mathematicians Urbain Le Verrier of France and John Couch Adams of England independently calculated the position of the unseen planet, making its discovery a relatively simple matter once someone could be persuaded to look in the appropriate direction. In the event, it was Le Verrier whose prediction was first tested. While Adams' calculations lay almost ignored by the English Astronomer Royal, the director of the Berlin Observatory J. G. Galle and his assistants searched the region suggested by Le Verrier. They found the predicted planet on 23rd September 1846. However, the discovery of Neptune was not the end of the issue as far as distant planets were concerned. After a few decades it seemed that Neptune alone could not explain all the problems with the orbit of Uranus. This hinted that there might be yet another planet lurking in the darkness of the outer solar system. Two Americans set out to see if this was the case.

William Pickering was one of these planet hunters, suggesting in 1908 that a planet with twice the mass of the Earth should lie in the direction of the constellation Cancer. His prediction was ignored. Eleven years later he revised his calculations and pointed to a position in nearby Gemini. This time astronomers at the observatory on Mt Wilson, California, responded, using their 24 cm telescope to search around Pickering's predicted coordinates. They failed to find anything. Meanwhile, American millionaire Percival Lowell was also turning his attention to the outer solar system. Lowell, who had earlier convinced himself that intelligent life existed on the planet Mars, firmly believed that deviations from the predicted positions of Uranus meant that there must be another unseen planet remaining to be discovered. He called this distant object 'Planet X' and, like Pickering, he tried to calculate where in the sky it might be found. However, Lowell had an advantage over his rival, for he had the means to pursue his search without relying on the whims of others. Lowell owned a private observatory which he had founded in 1894. It was built on Mars Hill, just outside the town of Flagstaff, Arizona. Unlike modern observatories, which are usually located on barren mountaintops, Lowell placed his telescopes in a delightful setting. The Lowell observatory was surrounded by pine trees and had a fine view back across the town.

Lowell's Planet X was predicted to be quite large, but very distant, and so was unlikely to show an obvious disc in the eyepiece of a small telescope. The best way to find it would be to detect its daily motion relative to the fixed background of stars and galaxies. In the previous century such searches had been made by laboriously sketching the view through a telescope and then comparing this with sketches of the same region made a few days earlier. However, by Lowell's time, astronomical photography had come on the scene and offered a much faster and more reliable way to survey the sky. Lowell's first search was made between 1905 and 1907 using pairs of photographic plates which he scanned by eye, placing one above the other and examining them with a magnifying glass. He soon realised that this method was not going to work.

Lowell's next step was to order a device known as a blink comparator to assist in the examination of the photographs. The comparator provided a magnified view of a portion of the photographs but, more importantly, it allowed the searcher to switch rapidly between two different images of the same patch of sky. Once the photographs were

aligned correctly, star images remained stationary as the view flashed from one plate to the other. However, should there be a moving object in the field of view, its image would jump backwards and forwards as the images were interchanged. Naturally enough, the process was known as 'blinking' the plates.

A search of the constellation Libra was made in 1911, but was abandoned after a year when nothing was found. Undeterred by this failure, Lowell began another search in 1914. Between then and 1916 over 1000 photographic plates were taken, but once again nothing was found.† Lowell died suddenly from a stroke on the 16th of November 1916, his planet-finding ambition unfulfilled. He was buried in a small mausoleum, shaped to resemble the planet Saturn, in the grounds of his observatory on Mars Hill. For a time the search for Planet X was halted as Lowell's widow tried to break the provisions of his will. Mrs Lowell wanted to remove funds from the operation of the observatory and preserve the site as a museum in her late husband's memory. The resulting litigation siphoned off funds from the observatory for a decade.

Eventually, under the directorship of Vesto Slipher, the Lowell Observatory returned to the problem of the missing planet. Slipher recruited a young amateur astronomer named Clyde Tombaugh, a farmboy from Kansas, as an observing assistant. Tombaugh arrived in Flagstaff during January 1929 and was set the task of taking photographs which could be searched for Lowell's Planet X. It took a while to get the new 31 cm telescope, built especially for the search, into full operation, but by April all was ready. Tombaugh took a number of photographic plates covering the region around the constellation of Gemini, the latest predicted location of Planet X. The plates were 33.5 cm by 40 cm in size and covered nearly 150 square degrees of sky. Each contained many thousands of star images. Vesto Slipher and his brother blinked the plates over the next couple of weeks, but they failed to find anything. In the meantime, Tombaugh continued to photograph the sky and soon a large backlog of unexamined plates had built up. Slipher then asked Tombaugh to blink the plates as well as taking them, explaining that the more senior observatory staff were too busy to devote much time to the onerous and time-consuming blinking process.

† In fact the missing planet did appear on two of these images, but it was much fainter than expected and its presence was missed.

Figure 1.2 Clyde Tombaugh entering the dome of the Lowell Observatory's 33 cm telescope. He is carrying a holder for one of the photographic plates. After exposing the plate he had to search it millimetre by millimetre for Planet X. Few astronomers now go to their telescopes so formally dressed, as can be seen by comparison with figures 4.1 and 5.7. (Lowell Observatory archives.)

Tombaugh regarded the prospects of his new assignment as 'grim', but he dutifully continued with his programme. Night after night he made a systematic photographic survey of the sky. He concentrated on regions close to the ecliptic, an imaginary line across the sky which marks the path traced out by the Sun across the constellations of the zodiac during the course of a year. The ecliptic is not the precise plane of the solar system, which is better defined by taking account of all the planets and not just of the Earth. When this is done the result is known as the invariable plane. However, when projected onto the sky, the ecliptic and the invariable plane are not much different and it is common, if careless, to use the two terms interchangeably. Since the orbit of Planet X would presumably be close to the invariable plane, the ecliptic was the obvious region around which to search.

Tombaugh's method was to take three photographs of each region of sky at intervals of two or three days. Each photograph was exposed for several hours. During each exposure Tombaugh painstakingly guided the telescope to make sure that the images of the stars were sharp, with all their light concentrated onto as small an area of the photographic emulsion as possible. Only then would his plate reveal the very faintest objects and give him the best chance of success. At dawn he developed the plates, careful lest a tiny mistake ruin them and waste his hours of work in the telescope dome. Later he examined the plates for anything which might have moved between the two exposures.

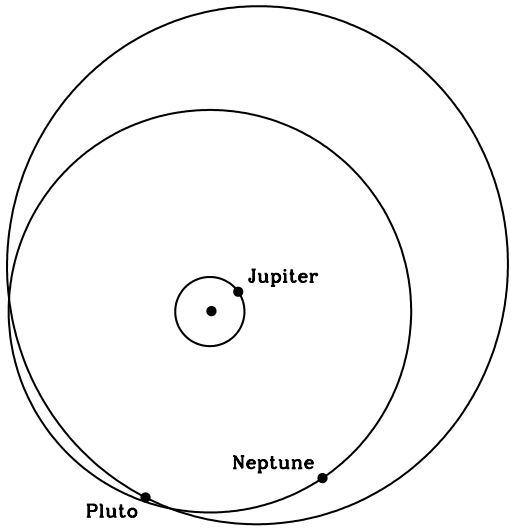

Figure 1.3 The orbits of Jupiter, Neptune and Pluto. Pluto's eccentric orbit crosses that of Neptune, although the significance of this was not realised at the time of its discovery. (Chad Trujillo.)

Although the technique sounds simple in principle, Tombaugh's task was a huge one. The long nights of observing were tiring and the blinking of the frames was tedious in the extreme. Many false detections appeared, caused by things such as variable stars, chance alignments of main belt asteroids and photographic defects which mimicked moving objects. To eliminate these false detections, Tombaugh used his third plate to check if any of the candidate objects were visible again. Usually, of course, they were not. Tombaugh's patience was finally rewarded on the 18th of February 1930 when he was examining a pair of plates he had taken a few weeks earlier. Blinking them, he found a moving object that was clearly not a star, a nearby asteroid or a flaw in the photographic emulsion. What was more, the object's slow motion across the sky suggested that it must be well beyond Neptune. After a few more weeks of observations had been made to define the object's motion more accurately, the discovery was announced on 13th March. The date was chosen since it would have been Lowell's 75th birthday if only he had lived to see the day. After a certain amount of debate, to which we shall return later, the new object was named Pluto, after the god of the underworld.

Clyde Tombaugh continued his search for another 13 years. He estimates that in this time he covered about 70% of the heavens and blinked plates covering some 90000 square degrees of sky.† All in all he spent some 7000 hours scanning every square millimetre of about 75 square metres of plate surface. Although he marked 3969 asteroids, 1807 variable stars and discovered a comet, he never found another object as distant as Pluto. This was a little odd since it gradually became clear that the new planet was rather smaller than predicted. The first clue that Pluto was small came from its faintness, which suggested it could not be any larger than the Earth. Worse still, even the largest telescopes of the day could not resolve Pluto and show it as a disc. Under even the highest magnifications, the planet remained a tiny point of light, devoid of any features. This was worrying since if Pluto was very small it could not affect the orbit of Uranus to any significant extent. None the less, the intensity of Tombaugh's efforts seemed to rule out any chance that any other massive planet could exist near Neptune's orbit.

It was not until much later that theoretical work, notably by

† The total area of the sky is less than this, but some regions were examined more than once.

American E. Myles Standish in 1993, explained the apparent deviations in the motion of Uranus. Standish based his calculations on improved estimates of the masses of the giant planets which had been determined during the flybys of the Voyager spacecraft. Using these he showed that any remaining errors in the measurements of Uranus' position were tiny and could be explained by systematic observational uncertainties. There was no need to invoke the gravitational influence of a missing planet, massive or otherwise. Lowell's hypothesis of a Planet X had been completely wrong. The discovery of Pluto was a consequence of the thoroughness of Tombaugh's systematic search and the fact that Pluto was fairly close to Lowell's predicted position was just a coincidence.

It was well into the 1970s before the true nature of Pluto was revealed. The planet's orbit was quite well defined within a year of its discovery, but Pluto's faintness made determining details of its physical make-up almost impossible for decades. In the mid 1950s it was established that Pluto has a rotation period of 6.39 days and in 1976 methane frost was detected on its surface. Since methane frost is quite reflective, this implied that Pluto was even smaller than at first thought. Pluto soon shrank again. In 1977 James Christy was examining images of Pluto when he noticed that the planet seemed to be elongated some of the time and not others. He soon realised that this was due to the presence of a large satellite going around the planet every 6.39 days, the same as Pluto's rotation period. As its discoverer, Christy had to name the new moon and he chose Charon, the name of the ferryman who transported souls to the underworld. Strictly speaking Charon should be pronounced Kharon, but it is often enunciated as Sharon since Christy's wife, Sharlene, is known to her friends as Shar. Once the details of Charon's orbit had been established, it was possible to determine the combined mass of Pluto and Charon. This turned out to be no more than 0.0024 times the mass of the Earth. Pluto was a small and icy world. Although the true size of Pluto was unclear in the 1940s, it may have been the realisation that there was no massive Planet X that made Edgeworth and Kuiper speculate about the edge of the solar system. Certainly the existence of small icy worlds at the fringe of the planetary region seemed a natural conclusion from theories of how the solar system formed.

It had once been suggested that the solar system was produced when a close encounter between our Sun and another star pulled out a filament of material which condensed into planets. However, it was

soon shown that this could not be the case. The realisation that the distances between the stars were very large made such an encounter unlikely, but more importantly, it can be shown mathematically that material pulled out from the Sun could not form planets. Ejected material would either fall back onto the Sun or disperse into space. So astronomers rejected this near-encounter model. Instead, they embraced an idea put forward by the philosopher Immanuel Kant in 1755 and subsequently developed by a French scientist, Pierre Simon, Marquis de Laplace. In 1796 Laplace suggested that the Sun formed in a slowly rotating cloud of gas and that, as the cloud contracted, it threw off rings of material which formed the planets. Although many of the details have been improved, the general outline of this nebular hypothesis survives today.

Modern theories of the formation of our solar system begin from the assumption that stars like the Sun form in the clouds of gas and dust which exist throughout interstellar space. These clouds often contain as much as a million times more mass than the Sun and each spreads over a huge volume of space. From time to time, instabilities develop within these clouds and bursts of star formation are triggered. About five billion years ago, an instability in just such a cloud triggered one such collapse. At the centre of this collapsing region, itself buried deep within the larger interstellar cloud, a dense clump of material began to form. As this protostellar core contracted, it increased in mass and so generated a more powerful gravitational field. This in turn attracted in more material, increasing the mass of the core still further in a rapidly accelerating process. As material fell in towards the centre it was slowed down by friction and gave up its kinetic energy as heat, gently warming the central regions of the core. For a while, the heat could leak out in the form of infrared and sub-millimetre radiation and so the collapsing core remained quite cool. However, as the cloud got more and more dense, a point was reached when its central regions became opaque to most forms of radiation. When this happened, heat could no longer escape easily and the temperature at the centre began to rise rapidly. After a while conditions reached the point at which nuclear reactions could begin and the core began to convert hydrogen to helium. The energy released by these nuclear reactions generated sufficient pressure to halt any further collapse and the star we call the Sun was born.

Of course the details of the star formation process are far more complicated than can be described in a single paragraph. In particular,

a mathematical analysis of the fate of a spherical collapsing cloud immediately throws out a simple, but vitally important question. If the Sun formed from the collapse of a huge cloud of gas, why does it rotate so slowly, taking about 11 days to turn on its axis? This fact alone hints at the existence of planets as a consequence of the physical law that angular momentum, or spin energy, must be conserved.

The conservation of angular momentum can be observed when an ice dancer skating with arms outstretched enters a tight turn and begins to spin on the spot. If, as her spin begins to slow down, the dancer brings her arms in close to her body, her rate of rotation suddenly speeds up. A similar effect can be experienced, without getting cold feet, by sitting on a well oiled office chair and spinning around on it with your arms held out. If you pull in your arms you can feel the spin rate increase, push them out again and the spinning slows down. Try again with a heavy book in each a hand and you will find it works even better. This simple observation is revealing two important things about physics. Firstly, angular momentum depends on both the rate at which something is spinning and upon the distance of its constituent masses from the axis of rotation. Secondly, the total amount of angular momentum in a spinning system is conserved. So, as demonstrated by our ice dancer, as mass is brought in towards the axis of rotation of a spinning system, the spin rate must increase to keep the total amount of angular momentum, or spin energy, the same. The more mass there is on the outside of the spinning region, and the further the mass is from the spin axis, the more angular momentum the system has.

The problem faced by the forming Sun was that as the protostellar cloud collapsed, it had to lose considerable amounts of angular momentum. This is necessary because unless the original cloud was completely at rest when the infall began, then as material fell inwards, it would have transferred its angular momentum to the central regions. This would have increased the rate of rotation of the protosun quite dramatically. Unless this angular momentum could be removed, the spin rate would continue to increase as the collapse proceeded. By the time the core had shrunk to stellar dimensions, the rotation would be far too rapid to allow a star to form. So, somehow during its collapse, the core must have transferred angular momentum to material further out in the cloud. This occurred as magnetic fields and gas drag gradually forced the outer reaches of the cloud to spin around with the core. As this continued, the outer regions of

what had been a spherical cloud fell in towards the equator and the nebula became a huge, slowly spinning disc surrounding a small stellar embryo. The forming Sun continued to grow as material in the disc fell inwards onto it.

The conditions across the protoplanetary disc depended on the balance between the energy generated during the collapse, the light emitted by the still-forming Sun and the rate at which energy was transported through the disc. In the central regions it was too hot for icy material to survive. Here, in what became the inner solar system, most of the ices were evaporated and blown outwards, leaving behind more robust dusty material. Within the spinning disc, tiny grains began to bump into each other. The grains, remnants of the original interstellar cloud, were probably smaller than a micron across to start with, but the collisions were gentle enough that they began to stick together. At first they formed fluffy structures which were mostly empty space, but as they grew still further, they began to compact. Soon they reached the point were they were more like small pebbles, jostling each other as they orbited the Sun. Inexorably these lumps of debris grew still further. Then, once a few objects had reached a size of about ten kilometres in diameter, a dramatic change of pace occurred.

These larger lumps, or planetesimals, were now massive enough that their gravitational fields began to attract other passing material onto themselves. Once this started it dramatically accelerated the growth process. Before long a few planetesimals began to dominate all of the space around them, clearing away the remainder of the orbiting material by dragging it down onto their surfaces. Within 100000 years or so many rocky bodies about the size of the Earth's moon had formed. After this brief period of runaway growth, the pace slowed again. By now each planetary embryo had swept up all the material within reach and the distances between the larger objects were too great for them to encounter each other. It took another 100 million years for the planet-building process to be completed. Gradually, subtle gravitational interactions between the planetesimals stirred up their orbits enough for occasional dramatic collisions to occur. One by one the surviving embryos were swept up into the four terrestrial, or Earth-like, planets, which we see today.

Further out, about half a billion kilometres from the Sun, temperatures were low enough that ices could survive. So, as well as dust, the outer regions of the disc contained considerable amounts of water ice and frozen gases such as methane, ammonia and carbon monoxide.

Here, the growing planetary cores swept up this extra material to form giant planets dominated by the gases hydrogen and helium with a seasoning of various ices. Jupiter, the largest of these giants, was so large that, even while it was still forming, its gravitational field had a dominating effect on its neighbourhood. Jupiter's gravity stirred up the region between itself and the still forming planet Mars and prevented a single object dominating this region. Instead of forming a fully fledged planet, the growth stopped, leaving a population of smaller, rocky asteroids. Beyond Jupiter, the other giant planets Saturn, Uranus and Neptune grew as they too swept up the icy planetesimals from the space around them.

At great distances from the Sun the protoplanetary disc became much more diffuse. Here, although there was sufficient material to reach the stage of forming small planetesimals, there was not enough time, or enough material, for them to combine into a major planet. Instead they formed a diffuse zone of small icy objects in almost permanent exile at the fringes of the solar system. This is the frozen boundary of the planetary region; beyond it lies only the huge, more-or-less spherical cloud of planetesimals ejected into deep space by gravitational interactions with the forming planets and the rest of the stars in our galaxy.

After Edgeworth's and Kuiper's articles, thinking about a possible disc of planetesimals beyond Neptune lapsed until the early 1960s. A brief revival of interest began in 1962 when naturalised American physicist Alistair Cameron† wrote a major review about the formation of the solar system. Cameron's review appeared in the first issue of *Icarus*, a new scientific journal devoted exclusively to the study of the solar system. Using the same arguments as Kuiper and Edgeworth, namely that material in the outer regions of the protoplanetary disc would be too diffuse to form a planet, Cameron wrote that 'It is difficult to escape the conclusion that there must be a tremendous mass of small material on the outskirts of the solar system'. Soon after Cameron wrote his review, another astronomer turned his attention to the possible existence of a comet belt beyond Neptune.

American Fred Whipple, who had done much to explain the composition of comets a decade earlier, began by accepting the likely existence of what he called a comet belt. From his knowledge of comets, he reasoned that if a trans-Neptunian belt of icy planetesimals

† Cameron was originally a Canadian.

existed, then even objects as large as 100 km in diameter would be very faint. This would make the discovery of individual objects highly unlikely with then existing astronomical technology. Turning the question around, he then asked himself if the comet belt would be detectable if the comets within it were very small. Would the combined light of large numbers of small comets produce a faint, but detectable glow across the sky? His conclusion was that any glow from the comet belt would be too faint to see against the background of the night sky. In particular it would be masked by the diffuse glow of the zodiacal light, a band of light along the ecliptic plane produced by sunlight scattering off interplanetary dust in the inner solar system. Having decided that it was impossible to detect a hypothetical comet belt directly, he set out to attack the problem dynamically. Harking back to Adams, Le Verrier and Lowell, Whipple tried to find out if a comet belt could have any measurable gravitational effects on the rest of the solar system.

Whipple first considered the gravitational effect of the belt on the motion of the planets Uranus and Neptune. He concluded that a comet belt having a mass of 10–20 times that of the Earth might exist beyond Neptune, but that the evidence for such a belt was not conclusive. He merely noted that a hypothetical comet belt provided a better explanation of the apparent irregularities in the motion of Neptune than assigning a mass to Pluto that was much larger than seemed justified by other observations of the tiny planet. He even went as far as to say Pluto could not affect the other two planets significantly even if it were made of solid gold. In 1967 Whipple, together with S.E. Hamid and a young astronomer called Brian Marsden, tried to estimate the mass of the comet belt another way. They looked for its effect on the orbits of seven comets which all travelled beyond Uranus. They then calculated the gravitational effect that a hypothetical comet belt containing as much material as the Earth would have on the orbits of each of these comets. Since they found that the real comets had suffered no such effects, they concluded that any unseen comet belt could not have a mass of much more than one Earth mass. Thus Edgeworth's and Kuiper's ideas remained largely in limbo for a number of years. It was only when a number of advances in our understanding of comets began to come together that it was gradually realised that there was a problem that could best be solved by postulating the existence of an ecliptic comet belt.

The existence of comets, as ghostly apparitions that appear without warning, move slowly across the sky and then fade away, has been known throughout history. However, only in the latter half of the twentieth century was a reasonable physical model of a comet developed. Although Edmund Halley noticed the similarity between the orbits of a number of comets, realised they were the same object and predicted the return of what has become known as Halley's Comet, neither he nor his contemporaries really understood what a comet actually was. By the early 1900s the favoured model was of a loose aggregation of dust and rocks, little more than a loosely bound cloud of material, carrying with it gas molecules trapped both on the surfaces of the grains and deep within pores of the larger pieces. When the comet was warmed by the Sun, these gases were apparently released to form a tail. There were serious flaws with this model, the most significant being that such a system could not supply enough gas to explain the rate at which gas was known to be produced as a comet approached the Sun. There was really only one thing that was known for certain about comets: dynamically speaking, they were of two distinct types. Comets of one kind appeared unpredictably from random directions on the sky and made a single trip around the Sun before disappearing for thousands of years. Those of the other kind, which were generally much fainter, reappeared regularly and their returns could be predicted quite accurately.

Comets of the first kind, called long-period comets, follow very elongated (parabolic) orbits which range from the inner solar system at one extreme into deep space at the other. As with most solar system objects, it is convenient to describe these orbits in terms of astronomical units (AU). An astronomical unit is defined as the average distance of the Earth from the Sun and amounts to about 150 000 000 km. Using these units, Jupiter is 5 AU from the Sun and Neptune's orbit is at about 39 AU. The long-period comets which can be observed from Earth have perihelia, or closest approaches to the Sun, of less than a few astronomical units and aphelia, or furthest distances from the Sun, of many thousands of AU. In the late 1940s the Dutch astronomer Jan Oort examined the statistics of the few hundred long-period comets then known and suggested that they came from a huge, more-or-less spherical shell around the Sun which extended to about half way to the nearest stars. Oort believed that the comets were ancient planetesimals that had been gravitationally ejected from the region of what is now the asteroid belt during the planet-building process about

four and a half billion years ago. He suggested that they then remain in the distant cloud until random gravitational forces from other nearby stars change their orbits slightly and cause them to fall inwards towards the warm heart of the solar system. Gerard Kuiper, an ex-student of Oort's, soon pointed out that the comets, being icy, were probably formed in the region from 35 to 50 AU rather than in the asteroid belt. Kuiper believed they were ejected by Pluto, not Jupiter. However, the broad outline of Oort's theory for the origin of comets was generally accepted and the hypothetical shell of distant comets became known as the Oort Cloud.

About the same time as Oort was explaining the dynamics of the long-period comets, Fred Whipple brought forward his icy conglomerate or 'dirty snowball' model of a comet. He suggested that the essence of a comet was a single solid body a few kilometres across called the nucleus. Each comet nucleus comprises frozen ices such as water, carbon monoxide, ammonia and methane together with a small amount of dust. As the nucleus approaches the Sun, solar heating warms it and causes the frozen gases in its outer layers to sublime. This creates a physically large, but very tenuous cloud around the nucleus. This cloud is called the coma. The coma is not entirely gas, since as the gases leave the nucleus they carry with them tiny dust particles. The pressure of sunlight, and the solar wind of material

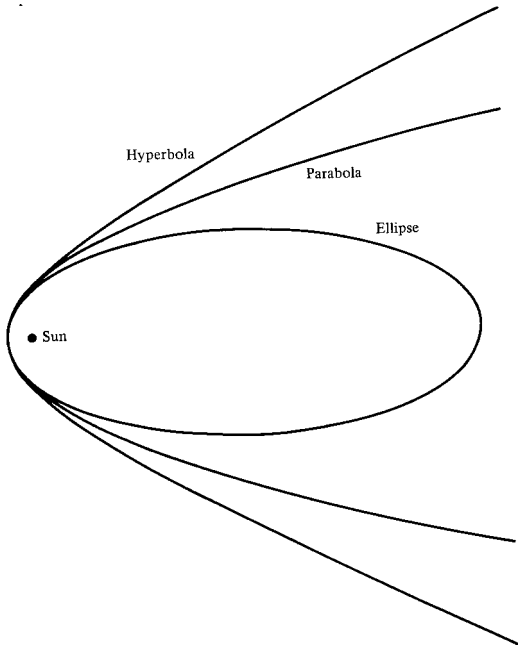

Figure 1.4 Comet orbits. Long-period comets approach the Sun along parabolas. Short-period comets orbit the Sun in ellipses usually, but not always, quite close to the plane of the planets. A comet on a hyperbolic orbit would be approaching from outside the solar system. The fact that no such hyperbolic comets are seen is evidence that comets are part of the Sun's family. The elliptical orbit has an eccentricity of 0.9.

constantly flowing out from the Sun, act on the coma and blow material away to form the comet's tail. Most comets actually have two tails, a long straight bluish one comprising gases that have been ionised and are moving directly away from the Sun and a curved, yellowish one comprising individual dust grains being blown away from the nucleus into independent solar orbits. In most comets, depending on the ratio of gas to dust in the nucleus, one of these tails is much more prominent than the other. Sunlight reflected from the coma and the dust tail makes the comet visible from the Earth. Once the comet has passed around the Sun and begins to recede back into deep space, the nucleus cools and the sublimation of the ices slows down and finally stops. Once this happens the comet rapidly becomes too faint to detect. Unseen, the frozen nucleus returns to the Oort Cloud from where, thousands of years hence, it may return to visit the Sun again.

The second class of comets are those of short period. These are confined to the inner solar system and most of them travel around the

Figure 1.5 Comet Hale–Bopp seen from Mauna Kea, Hawaii. Comet Hale–Bopp is a long-period comet and displayed a long, bright dust tail. Short-period comets are almost never visible to the unaided eye. (John Davies.)

Sun in elliptical orbits with periods of about a dozen years. In general, they have orbits of low inclination, that is to say they are close to the plane of the solar system. The short-period comets are generally much fainter than comets coming from the Oort Cloud. This is because short-period comets approach the Sun very frequently and on each trip more and more of the volatile ices which form the coma and tail are removed. So even when heated by the Sun at perihelion, short-period comets are pale shadows of their fresh, bright cousins making their rare appearances in the inner solar system. The faintness of the short-period comets indicates that they are gradually running out of volatile material and that they cannot survive for long in their present orbits. The short-period comets are fated to fade away completely, and to do so quite quickly in astronomical terms. By estimating how much material is removed on every trip around the Sun, astronomers have shown that short-period comets cannot survive in their present locations for even a small fraction of the age of the solar system. However, the very fact that numerous short-period comets do exist means that new ones must be arriving regularly to top up the present supply and replace them as they vanish. Many of these short-period comets have aphelia in the region of Jupiter's orbit; these are called Jupiter family comets. This link with the giant planet is a clue to their origin. Comets from the Oort Cloud which happen to approach Jupiter too closely have a chance of being captured into the inner solar system by Jupiter's gravity, making them doomed to make frequent approaches to the Sun until they disappear forever.

Although the Dutch astronomer Van Woerkom had noticed in 1948 that there seemed to be about twenty times more short-period comets than he would have expected, the idea that short-period comets were really just ordinary comets from the Oort Cloud which had the misfortune to be captured by Jupiter was accepted for a number of years. However, as more and more comets were discovered, it became clear that something was wrong. The observed population of short-period comets was too large to be explained by the effects of Jupiter's gravity on comets from the Oort Cloud. The developing problem was twofold. Firstly, the capture of an individual comet by Jupiter is a very unlikely event. Only if a comet flies quite close to Jupiter can enough gravitational energy be exchanged to slow the comet down and trap it in the inner solar system. With comets arriving from all directions, including from above and below the plane of the planets, the chances of crossing Jupiter's orbit just when the planet happens to be there,

and to do so close enough to the plane of the solar system to be captured, are very small. The probability of capture in this way is so low it has been compared with the likelihood of hitting a bird with a single bullet fired into the sky at random. Also, most of the short-period comets are in low-inclination orbits and most go around the Sun in the same sense as the rest of the planets. Since comets from the Oort Cloud approach the Sun at all angles to the ecliptic plane, including in orbits that go around the Sun in the opposite direction to the planets, it was puzzling that orbits of the short-period comets were not more randomly distributed.

In 1972, physicist–astronomer Edgar Everhart tried to resolve this problem. He suggested that the short-period comets were derived not from the capture of just any Oort Cloud comet, but rather from a subset of such comets which had specific orbital characteristics that made them likely to be captured. In particular, Everhart suggested that the short-period comets were Oort Cloud comets which had entered the zone of 4–6 AU from the Sun close to the plane of the solar system. Everhart's model could explain why the short-period comets population looked the way it did, but it was not long before another problem showed up. The following year Paul Joss from Princeton University looked at Everhart's model and put in some estimates for the capture rate, the likely lifetime of a typical short-period comet, and so on. In just two pages of text he showed that there were hundreds of times too many short-period comets to be explained by Everhart's capture model. Of course, not everyone agreed with him, but it did look as if something was missing from the equation. While the dynamicists pondered this problem, a new piece of the puzzle was about to be turned up by a strictly observational astronomer.