The Cosmos: A Historical Perspective

Craig G. Fraser

Greenwood Press

THE COSMOS

Recent Titles in
Greenwood Guides to Great Ideas in Science
Brian Baigrie, Series Editor

Electricity and Magnetism: A Historical Perspective
Brian Baigrie

Evolution: A Historical Perspective
Bryson Brown

The Chemical Element: A Historical Perspective
Andrew Ede

The Gene: A Historical Perspective
Ted Everson

Planetary Motions: A Historical Perspective
Norriss S. Hetherington

Heat and Thermodynamics: A Historical Perspective
Christopher J. T. Lewis

Earth Cycles: A Historical Perspective
David Oldroyd

Quantum Mechanics: A Historical Perspective
Kent A. Peacock

Forces in Physics: A Historical Perspective
Steven Shore

THE COSMOS

A Historical Perspective

Craig G. Fraser

Greenwood Guides to Great Ideas in Science
Brian Baigrie, Series Editor

Greenwood Press
Westport, Connecticut • London

Library of Congress Cataloging-in-Publication Data

Fraser, Craig G.
The cosmos : a historical perspective / Craig Fraser.
p. cm. — (Greenwood guides to great ideas in science, ISSN: 1559–5374)
Includes bibliographical references and index.
ISBN 0-313-33218-5 (alk. paper)
1. Cosmology—History. 2. Astronomy,
Ancient. I. Title. II. Series.
QB981.F68 2006
523.109—dc22 2006006367

British Library Cataloguing in Publication Data is available.

Library of Congress Catalog Card Number: 2006006367
ISBN: 0-313-33218-5
ISSN: 1559-5374

First published in 2006

Greenwood Press, 88 Post Road West, Westport, CT 06881
An imprint of Greenwood Publishing Group, Inc.
www.greenwood.com

Printed in the United States of America

The paper used in this book complies with the Permanent Paper Standard issued by the National Information Standards Organization (Z39.48–1984).

10 9 8 7 6 5 4 3 2 1

CONTENTS

SERIES FOREWORD

The volumes in this series are devoted to concepts that are fundamental to different branches of the natural sciences—the gene, the quantum, geological cycles, planetary motion, evolution, the cosmos, and forces in nature, to name just a few. Although these volumes focus on the historical development of scientific ideas, the underlying hope of this series is that the reader will gain a deeper understanding of the process and spirit of scientific practice. In particular, in an age in which students and the public have been caught up in debates about controversial scientific ideas, it is hoped that readers of these volumes will better appreciate the provisional character of scientific truths by discovering the manner in which these truths were established.

The history of science as a distinctive field of inquiry can be traced to the early seventeenth century when scientists began to compose histories of their own fields. As early as 1601, the astronomer and mathematician, Johannes Kepler, composed a rich account of the use of hypotheses in astronomy. During the ensuing three centuries, these histories were increasingly integrated into elementary textbooks, the chief purpose of which was to pinpoint the dates of discoveries as a way of stamping out all too frequent propriety disputes, and to highlight the errors of predecessors and contemporaries. Indeed, historical introductions in scientific textbooks continued to be common well into the twentieth century. Scientists also increasingly wrote histories of their disciplines—separate from those that appeared in textbooks—to explain to a broad popular audience the basic concepts of their science.

The history of science remained under the auspices of scientists until the establishment of the field as a distinct professional activity in middle of the twentieth century. As academic historians assumed control of history of science writing, they expended enormous energies in the attempt to forge a distinct and autonomous discipline. The result of this struggle to position the history of science as an intellectual endeavor that was valuable in its own right, and not merely in consequence of its ties to science, was that historical studies of the natural sciences were no longer composed with an eye toward educating a wide audience that included non-scientists, but instead were composed with the aim of being consumed by other professional historians of science. And as historical breadth was sacrificed for technical detail, the literature became increasingly daunting in its technical detail. While this scholarly work increased our understanding of the nature of science, the technical demands imposed on the reader had the unfortunate consequence of leaving behind the general reader.

As Series Editor, my ambition for these volumes is that they will combine the best of these two types of writing about the history of science. In step with the general introductions that we associate with historical writing by scientists, the purpose of these volumes is educational—they have been authored with the aim of making these concepts accessible to students—high school, college, and university—and to the general public. However, the scholars who have written these volumes are not only able to impart genuine enthusiasm for the science discussed in the volumes of this series, they can use the research and analytic skills that are the staples of any professional historian and philosopher of science to trace the development of these fundamental concepts. My hope is that a reader of these volumes will share some of the excitement of these scholars—for both science, and its history.

Brian Baigrie
University of Toronto
Series Editor

PREFACE

I was drawn to the history of modern cosmology because of a childhood fascination with astronomy and because the development of theories of the universe is one of the most exciting stories in all of science. I began to ponder the mysteries of the universe as a teenager, when for several years I made observations of variable stars with a 2.4-inch refractor and submitted them to the American Association of Variable Star Observers. The present book grew out of courses in the history of astronomy and cosmology that I have taught over the past 10 years. In some of these courses the subject was presented from the point of view of the history of mathematics, and there was a greater emphasis on mathematical details than is the case here.

I am grateful to Brian Baigrie for inviting me to contribute a book on the history of cosmology to the Greenwood series. I have benefited from discussions with Alexander Jones, John Steele, Nathan Sidoli, Elizabeth Burns, and Matthew Edwards as well as with the students in my courses. The help of two student research assistants, Steven Teasdale and Shayan Hamidi, is greatly appreciated. Jeff Kent composed the illustrations and James Ingram of the University of Toronto library photographed images from the library's collection. Finally, I would like to thank my wife and daughter for their patience during the hours I sat in front of the computer writing.

LIST OF ILLUSTRATIONS

1

INTRODUCTION

The history of Western cosmology may be divided into five periods. The beginnings of the subject were marked by the primarily mythological understanding of the world characteristic of Middle Eastern and Mediterranean societies and lasted until the emergence of Greek scientific thought in the period from 500 to 300 B.C. The second period began with the conceptions of the early Greeks and ended with the publication of Nicholas Copernicus's *On the Revolutions of the Heavenly Spheres* in 1543. Throughout these two millennia, thinkers placed the Earth at the center of the cosmos and described the motions of the planetary system in terms of a combination of spheres rotating in various ways about the Earth. The third period commenced with Copernicus and may be said to have finished around 1750; it was characterized by a view of the cosmos centered on the Sun and was codified in Isaac Newton's great work *Principia Mathematica* of 1687. After 1750, there was an increasing tendency to view cosmology as a subject that primarily concerned the many stars, star clusters, and nebulae that are distributed throughout the sky. During the period the universe beyond the solar system was opened up to observation and physical study, although its large-scale structure was a subject of continued speculation and debate. The final period had a very clear beginning in 1925 with the determination of the extragalactic character of the spiral nebulae and the realization that the universe is populated by a myriad of galaxies, of which the Milky Way galaxy is just one member. This finding was followed only four years later by the formulation of the red shift law and the discovery of universal expansion. By the end of the twentieth century, the standard big bang model had become the accepted theory of the universe. The universe originated in a creation event around 13 billion years ago, in conditions of exceedingly high density and temperature, and has been expanding ever since, becoming less dense and cooler with time.

The subject of this book is how conceptions of the universe's origin and large-scale structure developed and changed throughout history. What is of interest is how assumptions about cosmology influenced astronomical work and how astronomical work in turn generated cosmological beliefs and constructions. The development of prescientific cosmologies, such as the creation story of the book of Genesis or the mythology set out in the Babylonian epic *Gilgamesh,* are representative of a literary and religious outlook of considerable interest, but they fall outside the scope of an inquiry devoted to the history of scientific cosmology.

An important theme in the study of ancient cosmology concerns the question of whether the Greek geometrical modeling of planetary motions corresponded to a belief in physically real mechanisms in the heavens to produce these motions. Opinions on this question are divergent. Some prominent commentators hold that the ostensible statements by the Greeks about the nature of the heavens should be viewed as having only a nominal significance and that at a deeper level their devices to account for the motions were meant simply as mathematical constructions useful in prediction. Another body of opinion sees a unity of conception in the Ptolemaic mathematical and cosmological constructions of the planetary system and views attempts to separate them as unhistorical. Any interpretation of Greek astronomy must grapple with this question, and it will be a subject of some concern in our account.

Until the end of the seventeenth century the Western cosmos referred, for all intents and purposes, to the arrangement of the Earth, Moon, Sun, and the five planets. Before Copernicus the stars were supposed to lie on a sphere only slightly beyond the orbit of Saturn. The fundamental shift in scientific understanding that occurred with Copernicus involved a transition from a closed, Earth-centered universe to a vastly larger Sun-centered universe. This change in understanding would require a fundamental revision of both astronomy and physics, a great historical event known as the Scientific Revolution.

The Scientific Revolution is one of the most closely studied subjects in the history of science. A recurring topic of reflection concerns Copernicus's role in the early modern upheaval in cosmology. Although he is traditionally depicted as a hero of science, much historical writing over the past half century has been critical of the Polish canon's work as an astronomer. From a certain point of view his writing of a systematic treatise on the heliocentric system seems to have been an improbable achievement by a figure who was more medieval than modern. The problem of Copernicus's place in the history of science is connected in an interesting way with the larger question of how we are to understand the technical, philosophical, and cultural facets of the Scientific Revolution.

The modern revolution leading up to the consolidation of the big bang theory resulted from a series of discoveries made possible by advances in instrumentation. Large telescopes placed on mountaintop observatories analyzed the faint light from distant nebulae. These observations were aided by a new theory of gravity—Einstein's general theory of relativity—as well as by new physical knowledge of energy processes in the centers of stars and in the very early universe. We have today, for the first time in history, a clear picture of the structure of the whole universe, its origin and evolution in time. We are privileged to live in one of the most exciting times in the history of civilization, when ultimate truths about the cosmos are continually being revealed by ever more sophisticated instruments of observation.

A central question in the history of modern cosmology concerns the relationship between the exciting discoveries in nebular astronomy and the almost exactly contemporaneous emergence of Einstein's general theory of relativity. Although we look back and recognize the historical and even epochal significance of the stunning discoveries in observational cosmology, at the time it was Einstein's theory that seemed novel and revolutionary. Throughout the decade leading up to Edwin Hubble's 1929 breakthrough, speculation about the red shifts was often tied in with theorizing in relativistic cosmology. It is true that world models based on solutions of the relativistic field equations reflected assumptions that were later found to be true of the universe as a whole. Nevertheless, that the investigation of relativistic solutions occurred at the same time as the exciting advances in nebular astronomy was, in the final analysis, an interesting historical coincidence.

Theory and observation have had a somewhat ambivalent relationship in modern cosmology. The postulation of critical theoretical entities or phenomena did not result in a concerted program of observation to detect them. Examples are provided by the microwave background radiation and gravitational lensing, both of which were discovered by accident in the course of projects devoted to other purposes. The background radiation was predicted to exist in the 1940s by theorists working on the conditions that must hold in the early universe but was not found until 1965. Gravitational lensing is now seen as an important phenomenon explained by the general theory of relativity, but examples were only first identified in 1979. Another less clear-cut example is provided by gravitational waves. Attempts to find such waves were carried out only by a few isolated researchers, and the first confirmation of their existence was a serendipitous discovery in the 1980s involving the radio observation of binary pulsars. Quasars, dark matter, and universal acceleration are all observational phenomena for which there was little or no theoretical precedent. An interesting exception to this general pattern was the postulation in the 1960s of black holes, initially a purely theoretical concept that has proved to be durable and of increasing importance in the study of quasars and galaxies.

The causes and historical courses of the two revolutions—the Copernican revolution and the big bang revolution—were very different. The events of the sixteenth and seventeenth centuries involved a reconceptualization of a body of observations that had been around for close to 2000 years. The construction of the new world picture occurred primarily through the elaboration of theory and was aided only later—if very importantly—by advances in instrumentation. One has only to look at Copernicus and Kepler, neither particularly notable observationalists, to appreciate what was achieved by thought and reflection alone. By contrast, modern cosmology grew directly out of the fantastic discoveries being made with the great American telescopes. The subsequent detection of the cosmic background radiation and universal acceleration was also made possible by the immense resources of scientific technology. Technology has come to play an ever more prominent role in modern cosmology. What the two revolutions do have in common was a profound change in the way in which the universe as a whole was understood and a consequent fundamental rethinking of our place within the cosmos.

2

BABYLONIAN AND CHINESE ASTRONOMY AND COSMOLOGY

Two civilizations of antiquity other than the Greeks who cultivated astronomy in a quantitative and systematic way were the Seleucid Babylonians in the period from 320 B.C. to 600 A.D. and the Chinese, during the Han Dynasty, from 200 B.C. to 200 A.D. Both civilizations, at an earlier stage, had possessed well-developed, imaginative cosmologies, but these contained very few references to findings of astronomy, and observational work in astronomy was uninfluenced by cosmological assumptions rooted in the prevailing mythology. Babylonian astronomy seems to have been carried out without any assumptions about the nature of the universe in which the objects of astronomy were located. The Chinese contributions to astronomy consisted of a fairly complete record of celestial phenomena that was relatively crude in comparison to Babylonian data. However, in Han China, empirical astronomy became combined with an explicit, if fairly elementary, spatial cosmology.

BABYLONIAN ASTRONOMY

Although the Babylonians did not contribute to cosmology, their astronomy is of great interest because the astronomical data they accumulated would later be of the utmost importance in the development of Greek geometric astronomy and cosmology. The emergence of Babylonian astronomy was preceded many centuries earlier by the appearance of a very advanced mathematics, documented in cuneiform clay tablets dating back to 1700 B.C. and earlier. This mathematics was based on a base-60 positional numeration system and contained solutions to quadratic equations and algorithms to compute the square roots of numbers. Although there was some interest in geometry, the Babylonians emphasized the arithmetic and algebraic parts of mathematics. There was, during this older period, no comparable development of astronomy in even its most rudimentary empirical form. It was only much later, beginning around 600 B.C., that a sophisticated numerical astronomy was cultivated. The Seleucid Babylonians compiled very accurate tables giving the positions of the Sun, Moon, and planets as a function of time. They did so using the ancient mathematical tools; thus the base-60 system of notation was used to measure time and angles and has survived up to this day in timekeeping and navigation.

In considering Babylonian mathematics and astronomy we are in the unusual position of having a substantial collection of original artifacts—the clay tablets on which the tables and procedures were recorded—but very little or no information about the individual astronomers and no explanation of the methods and outlook that guided their work. We know that at a fairly early stage the Babylonians divided the ecliptic into twelve parts, each part being 30 degrees wide. These parts would become associated with constellations in a way that is familiar to everyone today. The zodiacal divisions, or signs, provided a convenient way of identifying the location of a celestial body, which would be given in terms of the sign and the number of degrees along the ecliptic within the sign.

Some indication of the character of Babylonian astronomy may be obtained from a tablet from 133 B.C. giving the position of the Sun each month when it is in conjunction with the Moon. (The following account is based on Neugebauer (1969, chap. 5).) The speed of the Sun's motion along the ecliptic varies, with motion being faster in the winter and slower in the summer. The total variation in speed is not large, being only about 3 percent of the average speed. Babylonian astronomers not only detected the variable solar speed but compiled tables accurately, giving it as a function of time. The table in question contains three columns. The month is listed in the first column, the number N of degrees traveled by the Sun in a one-month period following conjunction with the Moon is in the second column, and the position P of the Sun at conjunction is in the third column. (The structure of the table is indicated in table 2.1, which describes three successive rows. It should be noted that specific numbers rather than variables appear in the original table.) In order to find the position of the Sun for the next month, one adds P to N, and this gives the next entry in the third column. The second column gives the solar velocity along the ecliptic since it lists the degrees traveled by the Sun in successive, constant, one-month time periods. It turns out that the function giving the solar velocity is what is called a linear zigzag function, in which the dependent variable increases in a linear fashion, stops, and then decreases in a linear fashion.

Table 2.1: Babylonian Solar Table, 133 B.C.

Month	Degrees traveled by the Sun in one-month period following conjunction with Moon	Position of Sun at conjunction with Moon
T	N	P
$T+1$	N'	$P+N$
$T+2$	N''	$(P+N)+N'$

In a comment on this and other tables Neugebauer (1969, 110) observes that "at no point of this theory are the traces of a specific geometrical model visible." Babylonian astronomy, even more so than Babylonian mathematics, avoided any use of geometrical figures or constructions. The positivist dream of a science without hypotheses was realized by the Babylonians in their computations of planetary positions. From the existing evidence it appears that only functional numerical patterns inferred from the data were used to compile predictive tables. That the Babylonians were able to attain such high levels of observational accuracy with no underlying geometrical cosmology is one of the great marvels of ancient exact science.

CHINESE ASTRONOMY AND COSMOLOGY

In ancient China, astronomy was a state-sponsored activity, and astronomers were members of the imperial bureaucracy. The demands of the emperor included the construction of accurate calendars and the keeping of a complete record of celestial events. Because detailed histories were produced for each dynasty, we have an unusually complete record of the activities of Chinese astronomers. As in Babylonian astronomy, the Chinese relied on arithmetical-algebraic procedures to study the motions of the Sun, Moon, and planets. As in Greek astronomy, the Chinese developed an explicit cosmology and used geometry—albeit of a very elementary sort—to determine some of the numerical constants of the model. Chinese cosmology in the sense of a spatial physical conception of the celestial world was more primitive than its Greek counterpart and never played much of a role in the primary subject of calendrical astronomy.

Both astronomy and cosmology reached a certain level of maturity during the Han Dynasty. We have the treatise *Zhou bi suan jing,* which dates from the first century B.C., one of the earliest surviving Chinese scientific works and one that has been closely studied by modern scholars. There is also a treatise on cosmology, the *Ling xian,* written around 100 A.D. by the great Han astronomer, Zhang Heng.

During the Han age an older cosmology, the *Gai tian,* gave way to the *Hun tian,* the latter remaining the dominant cosmology for the following centuries. The *Gai tian,* or "Doctrine of the Heaven as a chariot-cover" (Cullen 1996, 35), is described in the *Zhou bi suan jing.* It posited a flat, stationary Earth beneath the heavens, the latter rotating rather like a large umbrella about a point on the surface of the Earth. The rising and the setting of the Sun was explained as an optical phenomenon that resulted as the Sun merged in the distance with the horizon. The *Hun tian,* which is set out in the *Ling xian,* replaced the chariot cover by a sphere. The heavens revolved as a sphere on an axis inclined to the flat base of the Earth in much the same way that the skies revolve about the auditorium floor of a planetarium. The celestial sphere became the fundamental astronomical concept, and coordinates on this sphere were used to locate the positions of celestial objects.

In addition to the *Gai tan* and *Hun tian,* there was a third cosmology in early China, the *Xuan Ye.* According to this view, the planets moved through empty space without the assistance of mechanical spheres. Despite its somewhat modern-looking outlook, the *Xuan Ye* represented a general and indefinite conception of the cosmos and never played a role in the study of astronomy.

The Chinese understanding of cosmology encompassed much more than the simple geometrical modeling of the motion of the planets. In all of the stages of its development Chinese thinking was informed by a belief in the organic unity of the universe, in the existence of various correlations or resonances between the earthly and celestial worlds. The numerical schemes employed to describe the solar-lunar calendar and the motions of the planets were influenced by a priori assumptions derived from a kind of numerological astrology. For example, ancient Chinese astronomers believed that the number five was special and that all things could be described in terms of five phases. The number five was the number of the wandering stars that accompanied the Sun and the Moon in their cyclical journeys through the heavens. Other numerical relations of cosmological significance were taken from the *I Ching,* the Book of Changes, a work that emphasized the balancing of opposites, of the yin and the yang, a dynamical process that was believed to be pervasive in the universe.

Astrology was much more integrated into Chinese astronomy than was the case in Babylonian and Greek astronomy. In the latter its role was largely that of an external agent that provided formal motivation to develop accurate planetary and eclipse tables. The actual construction of these tables followed empirical and theoretical principles that used observation and (in the case of the Greeks) geometric modeling, influenced, to be sure, by abstract philosophical beliefs about the mathematical nature of reality. By contrast, in Han China, numerical harmonies rooted in astrology influenced the selection of the cycles of numbers at the foundation of calendrical astronomy.

CALENDRICAL SYSTEMS

The calendar today is based on the Sun and its annual circuit around the sky against the background of the stars. The fundamental unit is the tropical year, the time from the summer solstice to the next summer solstice. Because one year is not exactly equal to an integral number of days, it is necessary to vary slightly the number of days in a year so that the calendar remains synchronized with the seasons over time. Thus every four years we add one extra day to the year. By contrast, the Chinese calendar was based on both the Sun and the Moon. A lunar-solar calendar was also used by the Babylonians and is fairly natural in any society that makes the month a fundamental unit to mark the passage of time. In ancient China the month was taken to begin on the day of conjunction of the Moon and the Sun. The last day of the month coincided with the last day that the crescent moon was observed in the eastern sky before sunrise. The lunar month as defined in this way actually varies slightly in value because the last visibility of the crescent moon is affected by such factors as the inclination of the ecliptic to the horizon and the latitude of the Moon with respect to the ecliptic. In addition, the length of the year measured as an integral number of days varies, holding to an average value that is known today to equal 365.2422 days.

In the Chinese lunar-solar calendar it was necessary to adjust the length of the month so that the calendar kept in step with the seasons. The fundamental problem was to find a cycle in which an integral multiple of years was equal to an integral multiple of months and to find other cycles that yielded both of these quantities in terms of an integral number of days. Another basic concept of the Chinese calendar was the original time, or epoch, from which all dates were computed. In any given dynastical system of calendrical astronomy the selection of the epoch was determined by political factors, astronomical considerations, and various numerological beliefs.

A fundamental cycle of the Chinese lunar-solar calendar was the equality 19 years = 235 months. The number 19 was regarded as significant because it was equal to 10 + 9, the yin and yang numbers from the *I Ching*. Another fundamental cycle was the equality 76 years = 940 months = 27,759 days. This is the smallest cycle that gives rise to a whole number of days, months, and years. We have $940 - (76 \times 12) = 28$, and so it follows that during a 76-year period, there will need to be 28 calendar years in which a 13th month is added to the standard 12-month year.

Up to the second century B.C., the traditional value used for the average length of the year was 365 1/4 days. In 100 B.C. the Han emperor Wu decided to institute a new calendar, or *li*. The most significant innovation was to change slightly the value adopted for the average length of the month. In the old system a month was equal, on average, to 29 499/940 days, while in the new system it was equal to 29 43/81 days. This change implied a change in the average length of the year, from 365 1/4 days to the value 365 385/1539. This change required various adjustments in the cycles making up the calendrical system.

The change in *li* under Emperor Wu illustrates how considerations of a numerological sort influenced thinking about the calendar. The basis of the change was a new value for the fractional part of the average month, namely 43/81. The number 81 in the denominator of this fraction was the square of nine, and nine was a special yang number in the mode of thinking that saw the cosmos as balanced between the cosmic forces of yin and yang. Eighty-one was also the volume capacity in conventional units of the standard pitch pipe, an instrument whose notes were believed to resonate with the cycles of the cosmos. The organic unity of the universe was expressed in resonances between the cycles of the calendar and the cycles found elsewhere in the affairs of man and nature.

STARS AND CELESTIAL COORDINATES

Chinese astronomy as it developed up to the early Han period was oriented toward the circumpolar stars. The heavens were divided into 28 parts, or lodges, each lodge consisting of a slice of the sky beginning at the north celestial pole and extending to the "Red Road," what in modern terms would be called the celestial equator. In terms of the concept of the celestial sphere, as it later became known, each lodge is analogous to an orange slice, in which the orange, or celestial sphere, is cut into 28 slices. The width of the mansions varied considerably, from 1.5 degrees to 30 degrees. The primary concept in describing the motion of a planet was the time at which it reached the meridian. Thus time rather than angular measure was the fundamental conceptual parameter of interest in astronomy. The coordinates of a celestial object were specified by the lodge in which it was located, its distance from the edge of the lodge, and the distance from the celestial pole. Distances were expressed in terms of a unit called a *du,* there being 365 1/4 *du*s in a whole circle. The *du* was the distance traveled by the Sun on the celestial sphere in one day and was therefore a temporal unit of measurement.

With the advent of the *Hun tian* and the explicit appearance of the concept of the celestial sphere, angular measure on this sphere supplanted time as the basic element of interest in astronomy. Such angular measurements were made with an armillary sphere, a physical model of the celestial sphere. Conceptually, the Chinese focus on the north celestial pole and the celestial equator distinguished it from Babylonian astronomy, where the fundamental object of reference was the ecliptic. The Chinese convention is followed in modern astronomy, where an object is located on the celestial sphere by its right ascension and declination.

An interest in the angular separation of objects on the celestial sphere did not extend to the development of the mathematical subject of trigonometry, which appears to have been an exclusively Western invention. The Chinese were able to use a comparison of triangles and some basic geometrical facts to calculate the height of the Sun above the flat Earth. Although they possessed a form of the Pythagorean theorem, they never produced a systematic body of results in geometry. The concept of a deductive proof was not part of Chinese mathematics.

As early as the fourth century B.C., the Chinese had compiled detailed star catalogs. Astronomers were also very interested in transitory and unpredictable celestial phenomena such as novae, comets, meteors, and sunspots and maintained a complete record of such events over many centuries. In the West, such phenomena received no attention at all, a fact that is sometimes attributed to the Greek interest only in what was regular and law-like in nature. In imperial China, all that happened in the heavens was viewed as important for the affairs of state, and so it was necessary for state specialists on *tianwen* ("celestial patterns") to keep a detailed record of all celestial phenomena.

CONCLUSION

Chinese astronomy of the Han period never reached a level of development comparable to its Babylonian and Hellenistic Greek counterparts. Han astronomers failed to detect the variable motion of the sun, used the assumption of a flat Earth to calculate the variation in the length of the Sun's shadow at different locations, and were unsuccessful in predicting solar eclipses. The later, more advanced development of Chinese astronomy, which reached its highest stage during the Yung Dynasty of the thirteenth century, was made possible by the rejection of the older numerology and was also aided to some extent by the absorption of ideas from Indian astronomy. Although it is currently unfashionable to compare Chinese scientific accomplishments to contemporary Western work, it is the case that a study of the Chinese case serves to highlight and bring into focus the remarkable innovations of Greek cosmology and predictive astronomy.

3

GREEK ASTRONOMY AND COSMOLOGY

INTRODUCTION

Cosmology in early civilizations was bound up with mythology and creation stories, with attempts to ground one's experience on Earth in an imaginative and religious interpretation of the world. Understood as an attempt to explain rationally the physical constitution of the universe, cosmology was first developed by Greek thinkers beginning about four centuries before the birth of Christ. The *Timaeus*, composed by the philosopher Plato around 380 B.C., combined the imaginative thinking characteristic of traditional cosmology with the theoretical outlook of contemporary Greek philosophy and mathematics. Although it was qualitative and speculative, the *Timaeus* set the general groundwork for subsequent scientific cosmology. The Earth is motionless at the center of the universe; the stars, Sun, Moon, and planets move in circles about the Earth. Plato idealized the objects of geometry, and the most perfect geometrical objects of all are the circle and sphere. The principle that all celestial motions are compounded of circular motions was adopted in various forms by all later ancient Greek thinkers and is commonly referred to as the Platonic axiom. It would dominate astronomy for the next two millennia.

The Earth-centered cosmology was set within a physically coherent picture of the world by Plato's younger contemporary, Aristotle. In his books *Physics* and *On the Heavens* Aristotle distinguished between the sublunary realm, the world below the Moon, and the celestial realm, the world of the Moon, Sun, planets, and stars. The four terrestrial elements were fire, water, earth, and air, and their natural motions were straight-line motions toward or away from the center of the Earth. Celestial bodies were made up of a fifth perfect element, quintessence, or ether, which naturally moved in a circle about the Earth. This dichotomy between the celestial and terrestrial realms was perhaps the most significant conceptual feature of ancient Greek cosmology.

GREEK COSMOLOGY: THE FIRST STAGE

The fourth and third centuries B.C. marked the emergence and flowering of Greek mathematical science. The three greatest figures of Greek mathematics—Euclid, Apollonius, and Archimedes—worked during the period from 320 to 200 B.C. Their efforts were preceded in the fourth century by the seminal contributions of Thaetetus and Eudoxus of Rhodes. Eudoxus studied at Plato's Academy and went on to establish a school of mathematics in Cnidus in Asia Minor. Eudoxus created the proportion theory at the foundation of Greek mathematics, and he was also the one who founded geometric cosmology.

The basis of the Greek geometrical view of the universe is what is known as the two-sphere model, a conception suggested by Plato in the *Timaeus* and developed more formally by Eudoxus. The Earth is a very small sphere at the center of the universe, surrounded at an immense distance by a celestial sphere, on which lie the fixed stars. The celestial sphere rotates once every 24 hours, taking with it the fixed stars, the planets, the Moon, and the Sun on their daily circuits through the sky.

The sphericity of the Earth was a fact that was supported by several pieces of evidence. The mast of a ship sailing off in the distance is the last part of the ship to disappear from view, just as we would expect if it moved on a curved arc on the spherical Earth. An eclipse of the Moon occurs when the Sun, Earth, and Moon are aligned, and the zone of darkness as it passes across the Moon possesses a circular shape, apparently the result of the Moon passing into the shadow of the spherical Earth. It is possible to travel within the Mediterranean region a considerable distance from south to north. As one does so, changes are observed in the altitude of the Sun at noon and in the total length of day at different times of the year, observational facts that seem explicable only by assuming that the Earth is a sphere.

The celestial sphere was both a conceptual object that facilitated the measurement of the position of objects in the sky and a material body to which the stars were attached and that rotated daily. Today, in surveying and navigation the celestial sphere endures as a mathematical idealization useful in organizing line-of-sight observations. The Greek conception of it as a material body seems to have derived primarily from the fact of its daily rotation: the sphere moved as one would expect a rigid body to move, with the relative distances of the different parts remaining unchanged during the motion. Thus it was the diurnal motion of the heavens which led to the reification of what otherwise would have been a purely mathematical concept.

The two-sphere model of the universe was well established in Greek astronomical thinking by the beginning of the fourth century B.C. and is believed to have been the inspiration for a system of planetary models created by Eudoxus. Although Eudoxus was also responsible for fundamental contributions to mathematics, none of his original writings have survived. The basic idea of his planetary system was adopted by Aristotle, who wrote about it in his *Metaphysics,* and there is also an account of the system by the Aristotelian commentator Simplicius in the fifth century A.D. Modern interest in the Eudoxan spheres stems from the writings of the nineteenth-century Italian astronomer Giovanni Schiaparelli (1835–1910), who reconstructed the Eudoxan explanation of planetary motion.

The Eudoxan conception is known as the system of homocentric or concentric spheres. In a slightly simplified form it works as follows. Each celestial body is assumed to be moved by a set of spheres concentric with the Earth and all at the same distance from the Earth. In the case of the Sun, there are two motions to be modeled: the daily motion of the Sun westward in the sky and the much-slower annual motion of the Sun eastward on the ecliptic, that is, on the great circle it traces annually on the celestial sphere. The two motions are understood to result from the action of two rotating spheres, to which the Sun is affixed in some manner. One of the spheres produces the daily motion of the Sun westward in the sky; this motion coincides with the daily rotation of the celestial sphere. A second sphere produces the slower motion eastward of the Sun along the ecliptic, with a period of rotation equal to the sidereal period of the Sun, that is, the time it takes for the Sun to complete a 360-degree circuit of the ecliptic with respect to the fixed stars. The axes of rotation of the two spheres are inclined to each other at an angle of approximately 23 degrees. (Yet a third sphere was added to model the motion of the Sun, although its purpose is not clear.)

Similarly, two spheres produce the motion of the Moon. The first coincides with the daily rotation of the celestial sphere and produces the Moon's daily westward circuit of the sky, and the second carries the Moon eastward along the ecliptic, completing one rotation in 27 1/2 days, the Moon's sidereal period. A third sphere was added, apparently, to account for some variations in the Moon's motion with respect to the ecliptic.

The main difference between the planets on the one hand and the Moon and the Sun on the other is that the planets exhibit periodic retrograde motions in their passage eastward along the ecliptic. For the sake of simplicity we consider the case of the superior planets. Figure 3.1 depicts the path of Saturn along the ecliptic over a two-year period. The backward motion occurs around the time when Saturn is in opposition, that is to say, when it is 180 degrees opposite the Sun in the sky. Today, we are aware that retrogradation is an optical effect that results as the faster-moving Earth in its orbit about the Sun passes the slower-moving Saturn in its orbit. As Saturn is sighted against the distant stellar background, it appears to move backward for a while, with the midpoint of the retrogradation occurring at opposition.

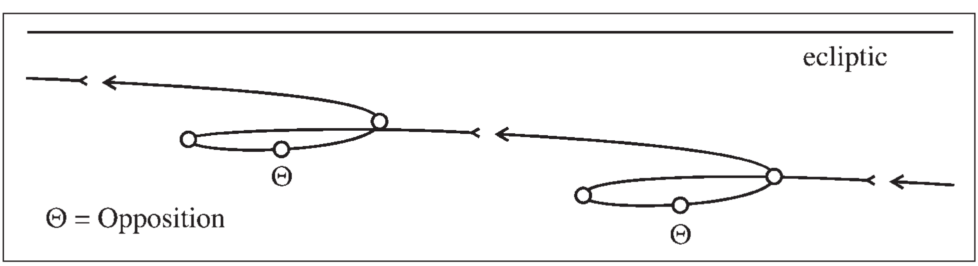

Figure 3.1: Retrograde motion of a superior planet.

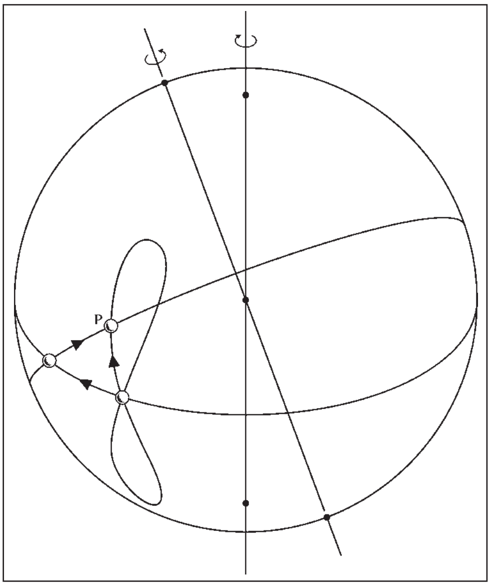

Figure 3.2: Eudoxus's model for retrograde motion.

Eudoxus was able to come up with a geocentric model that at least qualitatively produced the retrograde motion of a planet. Consider the case of Saturn. He first introduced two spheres to produce the daily rotation and the eastward circuit of Saturn around the ecliptic; the first sphere has a rotational period of 24 hours, and the second sphere has a period of 29 1/2 years, the sidereal period of Saturn. A third and fourth sphere were introduced to account for the retrograde motion. These spheres rotate with equal and opposite angular velocities about axes that are tilted with respect to each other (see Figure 3.2). Consider a point that is initially on the intersection of the equators of the two spheres. It will be carried by the two motions in a figure eight–shaped curve, whose axis of symmetry lies perpendicular to one of the equators. If the two spheres are positioned so that this axis of symmetry lies along the ecliptic path of the planet, the effect of the combined action of the third and fourth spheres will be to superimpose the figure-eight-shaped motion on the steady eastward motion of the planet, producing the retrogradations of the planet that are periodically observed. The motion of the planet is thereby successfully modeled using a set of four spheres.

The Eudoxan system was adopted by Callipus (370–300 B.C.) and by Aristotle, both of whom further developed the conception and added some embellishments of their own. Aristotle supposed that the motion of the spheres for a given planet occurs as the result of the transference of motion from the outermost sphere inward. In order to make this work mechanically, it was believed necessary to introduce additional "counterturning" spheres to counteract the westward rotational motion of the outer sphere. This modification introduced some complications into the original Eudoxan conception. In the Eudoxan system, there were a total of 25 spheres, while in the final Aristotelian scheme, 64 spheres were required to make the system work.

The Eudoxus-Aristotelian system was the first geometrical attempt to model the motions of the planets, and it continued to be upheld by some writers well into the Middle Ages. Nevertheless, it possessed some serious defects that led to its abandonment by virtually all later astronomers of note. First, in this system the distance of each planet from the Earth always remains the same, a fact that seems to contradict the periodic and substantial variations in the observed brightness of the planets. Whatever the cause of a planet's brightness, it is difficult to account for its changes other than to suppose that they result from changes in distance of the planet from the Earth. Second, the mechanism to produce retrograde motion succeeds only in a crude qualitative way in accounting for this phenomenon; in the case of certain of the planets it does not succeed at all. The basic problem is that there is only one degree of freedom, given by the angle of inclination between the axes of rotation of the two spheres that produce the retrogradation. This angle determines the width in latitude of the retrograde loops, and the latter is fixed once this width is given. There is then no flexibility in the model to produce the correct period relations for the planet's phases, its stationary points, opposition, and so on.

Despite its limitations, the Eudoxan system of homocentric spheres was a significant step forward in theorizing about the cosmos. It extended the geometric method of modeling evident in the two-sphere model to the motions of the Sun, Moon, and planets. There was now a complete geometric system of the heavens, consonant with the fundamental geometric outlook of Greek mathematical thought, which described a cosmos with a stationary Earth at the center of the universe. A crucial new intellectual element had entered into astronomy: each motion was revealed to result from a definite cause. For example, retrograde motion was (at least in principle) a consequence of the combined motion of two planetary spheres rotating in a specified way. The notion of causality was completely absent in Babylonian mathematical astronomy. Its emergence in Greek astronomy was connected to an interest in spatial geometrical modeling and was reflective, at a very general level, of the concern for the notion of cause in Greek philosophy and for deductive proof in Greek mathematics.

GREEK COSMOLOGY: THE SECOND STAGE

Greek rational cosmology emerged in an intellectual milieu dominated by a geometrical conception of mathematics, and it was natural that geometrical modeling was integral to the Greek astronomical outlook. The astronomy of Eudoxus and Aristotle was theoretical and qualitative, based on highly idealized models of how the planets move. During the second century B.C., the Greeks came into contact with a large body of observational data compiled by the Babylonians. Although the details of how this contact occurred are not known, it is believed that the acquisition of some of the Babylonian data and the associated numerical methods transformed Greek astronomy, eventually leading to the mature geocentric theory of the universe that would endure unchallenged as the dominant cosmology until the sixteenth century.

The legacy of ancient mathematical astronomy is contained in Ptolemy of Alexandria's masterpiece, *The Mathematical Syntaxis,* a work that is customarily known by its Arabic title, the *Almagest* (or "greatest"). Written around 150 A.D., the *Almagest* is, along with Euclid's *Elements,* one of the two or three most significant works of Greek exact science. Perhaps the greatest astronomical book ever written, it is a comprehensive exposition of the methods and theory needed to produce empirically reliable tables of motion for the Sun, Moon, and planets. In addition to the *Almagest,* Ptolemy composed a work on geography and some comparatively minor treatises on cosmology and astrology as well as some specialized studies of mathematical subjects.

Among his predecessors, Ptolemy credited the work of Hipparchus of Nicaea, an astronomer who lived around 140 B.C. and who appears to have been the most important figure in astronomy before Ptolemy. It is believed that the *Almagest* solar theory and some of the lunar theory were due to Hipparchus, as were many of the observations cited by Ptolemy in the *Almagest.* The star catalog presented by Ptolemy in book seven of the *Almagest* was originally compiled by Hipparchus.

Hipparchus initiated the basic methodology that would characterize advanced Hellenistic astronomy. One began by devising a geometrical model to explain the motion of a given celestial body. Observations were then used to compute the parameters of the model, that is, the constants that precisely specify the orientation and relative dimensions of the model. For this purpose Hipparchus drew upon the detailed and very accurate ephemeridae produced by the Seleucid Babylonians. The model as so calibrated could then be made the basis for predictive schemes giving the positions of the body for a sequence of future times. These predictive schemes were the basis for a table giving the planet's position as a function of time. It was Ptolemy who systematized and refined the methodology involved in the second stage of this project.

The essential difference between the Babylonians and the Greeks was the fundamental place occupied by geometric models in Greek astronomy. As far as cosmology is concerned, the main question is to understand the relationship between Ptolemy's geometrical models and the physical conception he held of the heavens. As we shall see, this question is not entirely straightforward.

THE HIPPARCHAN SOLAR MODEL

The third book of the *Almagest* is devoted to the study of the motion of the Sun and presents the theory that Ptolemy attributed to Hipparchus. In one year the Sun travels 360 degrees along the zodiac, beginning from a given fixed star, passing through the 12 constellations of the ecliptic, and returning again to the same fixed star. This is known as the sidereal year (from *sidus,* Latin for "star") because the motion is measured with respect to the fixed stars. If the motion of the Sun is measured relative to the first point of Aries—one of the two points of intersection of the ecliptic and the celestial equator—one obtains another measure for the length of the year, the so-called tropical year. The two years are almost the same, differing only by a very small amount. The slight discrepancy is the result of an effect known as the precession of the equinoxes, first detected and measured by Hipparchus.

The primary characteristic of the Sun's annual motion around the ecliptic is that it takes place with variable velocity: in the spring and summer it moves more quickly along the ecliptic than it does in the fall and winter. As we saw in chapter 2, the Babylonians possessed tables that tabulated the variable solar motion using arithmetic functions of some sophistication. For each year and the beginning of each month these tables tabulated the position of the Sun along the ecliptic, given in terms of the constellation of the zodiac and the number of degrees (from 0 to 30) from the beginning point of the constellation. A comparison of the Hipparchan theory (as reported by Ptolemy) with extant Seleucid Babylonian tables dating from around 300 to 500 B.C. establishes beyond doubt that Hipparchus used these tables in the construction of his theory.

Hipparchus explained the variable solar motion by assuming that the Sun moves about the Earth on a circle whose center is displaced slightly with respect to the Earth. The Earth lies close to the center but is not at the center itself. In figure 3.3 the Sun moves on the circle whose center is at E. The observer on Earth is located at Z. The distance ZE (measured as a fraction of the radius) is known as the solar eccentricity. The direction of the line DA gives a second parameter that, together with the eccentricity, fixes the solar model. The Sun moves uniformly along the circle, but because it is observed from the Earth at a point offset slightly from the center, it appears to be moving more quickly when it is at D than when it is at A. The variable solar velocity is therefore explained as an optical phenomenon resulting from the way the moving Sun is observed from the Earth against the celestial sphere.

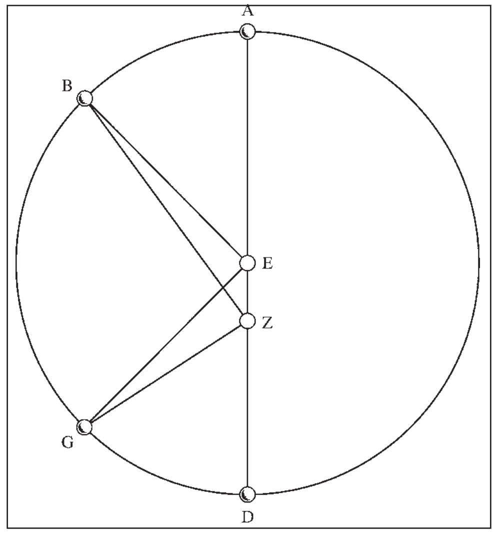

Figure 3.3: Hipparchus's solar model.

Hipparchus was able to compute tables of the Sun's position along the ecliptic as a function of time. He did so using a powerful new mathematical tool, developed by him and later extended by Ptolemy, known as trigonometry. The basic object of study is the triangle, and the central problem is to find a given side or angle, given that one knows the values of two other sides or angles. The angles are measured in degrees, with 360 degrees making up a whole circle and 90 degrees making up a right angle. The most complete exposition of trigonometry is contained in the first book of Ptolemy's *Almagest,* which explains how to construct a table of chords—a version of what we would call a table of sines—that gives the chords of angles from 0 to 180 degrees in one-half-degree increments.

Using Hipparchus's solar model, Ptolemy produced a table that enabled one to go from the mean, or average, position of the Sun to its true position in the sky. For a given one-year period one considers the position of the Sun at equally spaced intervals along the circle ABGDA. For each of these positions B one uses trigonometry to go from the angle AEB to the angle AZB. The first year begins at some specified point in time called the epoch. To find the position of the Sun at any future time, one determines the number of years elapsed since epoch, calculates the time elapsed since the beginning of the current year, determines for this time the position of the mean Sun, and then uses the table to go from this datum to the true position of the Sun.

Scientific cosmology may be said to have begun with the Hipparchan solar model. First, this model is causal, explaining the phenomena in question—the variable motion of the Sun—from two hypotheses: that the motion of the Sun takes place uniformly in a circle and that this motion is observed from Earth at a point that is slightly offset from the center of the circle. Second, in the Hipparchan model the motion of the celestial body is no longer assumed to take place on a two-dimensional manifold embedded in the celestial sphere. Instead, the Sun follows a trajectory in which its distance from the Earth constantly varies, and a crucial third dimension signifying depth in space is implied in this conception. Third, the phenomenon in question and the accompanying cosmological assumptions are combined with suitable mathematical tools to provide a precise quantitative analysis of the motion that allows systematic comparison with observation and the basis for prediction. In summary, the Hipparchan model provides a coherent and causal explanation that yields an empirically adequate account of the Sun's motion. It was the first contribution to cosmology in the sense of what we would today call a science and pointed the way to the later work of Ptolemy.

PTOLEMY'S MODEL FOR THE MOON

The fourth book of the *Almagest* is devoted to a detailed study of the motion of the Moon. This subject presented to Ptolemy one of the most challenging problems of his whole astronomical system because the Moon's motion is subject to several irregularities that are not present in the relatively simple case of the Sun. In addition to the inherent complications of the lunar motion, Ptolemy's observations followed a very particular pattern: the primary set of observations were made when the Moon was at syzygies, that is to say, at new Moon and full Moon, when the Sun, Moon, and Earth lie in a straight line. He developed a model for these observations, noticed that it was in conflict with observations of the Moon at first and third quarters, and modified the model to account for these differences. He noticed that the resulting model was slightly at odds with the measurement of the Moon at octants and modified the model once more to account for this fact.

It is worthwhile to consider what we know today about the motion of the Moon. The Moon is a member of the three-body system of the Moon, Sun, and Earth. Its primary motion takes place in an ellipse of small eccentricity, with the Earth at one focus. This motion is disturbed by the action of the Sun, a disturbance that results in several changes to simple elliptical motion. The most important change that occurs is a rotation in the direct sense of the line of apsides of the Moon—the apogee, or position of minimum lunar velocity, moves in a direct direction along the ecliptic by an amount equal to about three degrees per month. Separate from the rotation of the apogee, there are several other corrections that need to be made to simple elliptical motion. The largest of these changes is called the evection, a periodic correction that reaches its maximum when the Moon is 90 degrees from the Sun and a minimum when the Moon is aligned with the Sun and Earth. Next to the evection, there is a correction called the variation; it is also periodic and determined by the elongation of the Moon from the Sun.

The modern account depends on an analysis using the theory of perturbations, in which the mean, or average, motion is supplemented by a series of terms resulting from the mutual gravitational pull of the Earth and Sun on the Moon. Of course, Ptolemy knew none of this, but he did notice two observational facts:

the motion of the Moon along the ecliptic takes place with variable velocity, and the point of minimum velocity itself is not fixed but moves in a direct sense along the ecliptic; and the Moon's motion is subject to small changes that are connected to the relative position of the Moon and the Sun. To account for the variable velocity, Ptolemy used a deferent-epicycle model, in which the Moon lies on an epicycle whose center lies on a deferent with Earth at the center. The epicycle rotates with a period of one month, and the center of the epicycle revolves on the deferent around the Earth in about the same period. However, to account for the direct movement of the lunar apogee, one lets the period of the epicycle be slightly smaller than the period of its center on the deferent. The resulting model nicely accounts for the first observational fact mentioned above. Ptolemy also identified the evection and modified the model to account for it. He placed the center of the deferent on a small circle with center at the Earth. The center of the deferent rotates on this circle, in the process alternately drawing the epicycle closer to the Earth and extending it farther away in a reciprocating motion. One of the particular features of Ptolemy's lunar model is that the distance from the Moon to the Earth varies by a very considerable degree, with the maximum distance being almost twice the minimum distance. (This fact conflicts with the lack of any observed variation in the angular size of the Moon, a problem we discuss below.) Ptolemy did some further minor tampering with the model, but we will not follow him in this.

PTOLEMY'S PLANETARY MODELS

The challenge posed by the planets consisted in finding models that accounted for their chief motions, most notably their periodic retrograde motions, and could do so in a way that conformed to the best observational data and could serve as the basis for the production of tables of future planetary positions. The basic geometrical device used by Ptolemy is believed to have originated with the mathematician Apollonius of Perga, who lived in the third century B.C. and is best known for a major mathematical work he composed on conic sections. For the sake of simplicity we will concentrate on the case of the superior planets. The model consisted of motion on two circles, a larger deferent and a smaller epicycle. In figure 3.4 the observer G is located at the center of the deferent circle AB, while the planet is located on the epicycle DZH. The center A of the epicycle moves around the deferent, completing a full circuit of the ecliptic in a period that has a characteristic value for each planet, while the epicycle rotates with a period equal

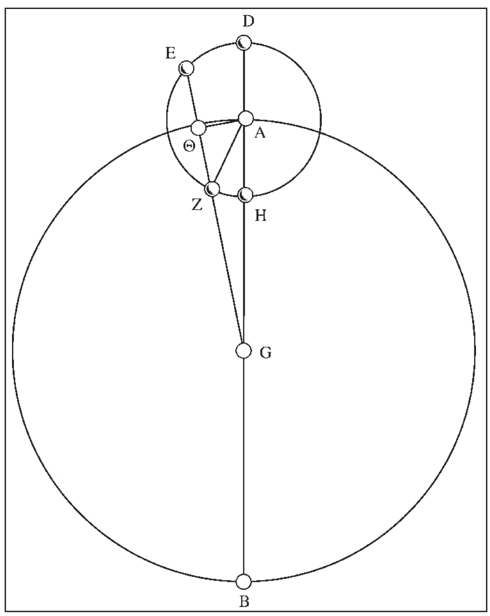

Figure 3.4: Ptolemy's model for the planets.

to one year. The direction of rotation of both the epicycle and the deferent is the same. If in figure 3.4 we are looking down from the north celestial pole, then this direction is counterclockwise. The motion of the planet at E is therefore the sum of its motion around the epicycle and the motion of the center of the epicycle around the deferent. When the planet is on the inside of the epicycle with respect to the Earth, it will, for a time, appear to move backward relative to the celestial sphere, exhibiting the characteristic retrograde motion observed around opposition at H; during this part of its circuit, it is closer to the Earth and also appears to be brighter.

The deferent-epicycle model was the basis for Ptolemy's planetary predictive schemes. However, it was necessary to modify the model slightly to conform with strict accuracy to the observational data. Recall that in order to account for the variable motion of the Sun along the ecliptic, it was necessary to displace the observer slightly from the center of the circle on which the Sun moves. The Sun moves uniformly on its circle, but an observer on Earth sees the Sun from a point offset from the center. Each planet also exhibits a similar variation in motion, although it is less immediately evident because of its periodic retrogradations. In order to account for this variation, or "anomaly," one moves the observer away from the center of the deferent but does so in a way that is slightly more complicated than in the case of the Sun. The Earth is situated at a point E slightly offset from the center Z of the deferent; hence the epicycle center Θ moves on a circle that is eccentric with respect to the Earth (figure 3.5). So far, we have proceeded exactly in the same way as the solar theory. However, we now specify another point D on the diameter HZK of the deferent on the other side of Z from E and at a distance from Z equal to EZ, EZ = ZD. D is called the equant point. The motion of Θ on the deferent is uniform not with respect to the center Z of the deferent but with respect to the equant point D. Thus the angle ADΘ increases uniformly as the center of the epicycle Θ moves around the deferent. Hence the motion of Θ occurs on a circle with center at Z but with uniform angular velocity about D. The model turns out to be a good one and succeeds very well in representing the variations that are observed in the motion of the planets. Many later commentators would object to what they saw as a violation of the Platonic principle implicit in the introduction of the equant since it is no longer the case that the deferent rotates uniformly with respect to its center. Nevertheless, it is still the case that this rotation is uniform with respect to the equant D, and so the Platonic principle may arguably be said to hold. The modified deferent-epicycle model involving the equant is regarded as one of Ptolemy's greatest technical achievements in astronomy.

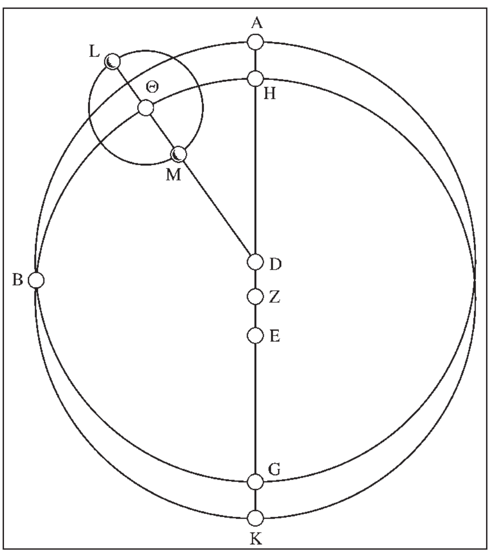

Figure 3.5: The equant.

THE *ALMAGEST* MODELS: INSTRUMENTALISM VERSUS REALISM

A new stage in the historical study of ancient exact science began in the nineteenth century with the preparation of reliable textual editions of the extant Greek scientific classics. Of Ptolemy's two notable works on astronomy, the *Almagest* was a work of positional mathematical astronomy, while the *Planetary Hypotheses* was devoted to an investigation of the physical structure and dimensions of the celestial system. The *Almagest* has survived relatively intact and represents a scientific achievement of the highest order; it has received far and away the most attention from historians of astronomy. The *Planetary Hypotheses* is a comparatively minor work. Only in the 1960s was it realized that an important part of it that is missing in the Greek existed in Arabic translation. A full appreciation of Ptolemy's cosmological conceptions as set forth in this work has been the product of relatively recent historical investigation.

In the nineteenth century the study of ancient Greek astronomy was carried out in a philosophical atmosphere that was strongly influenced by positivism. A prominent writer around 1900 was Pierre Duhem, a leading proponent of positivist physical philosophy and a major contributor to the history of astronomy. Duhem asserted that the primary purpose of ancient Greek astronomy was "to save the phenomena," that is, to devise predictive schemes to account for the motions of the planets. Duhem's point of view was developed by later historians into an interpretation of ancient astronomy that has come to be known as instrumentalism. According to this view, the geometrical models of Greek astronomy were not regarded by their inventors as real material mechanisms in the heavens but were merely mathematical constructions that were effective in prediction.

Some familiarity with the debate over the status of astronomical entities in ancient astronomy is crucial to any understanding of the rational cosmology of the Greeks. If the positivist-instrumentalists are correct, Ptolemy's astronomical theory as set forth in the *Almagest* was of limited explanatory import and should not be viewed as expressing a strong commitment to any particular physical arrangement of the universe. Although the *Planetary Hypotheses* did present an explicit cosmology, it was a minor work in comparison with the *Almagest* and was less influential in the subsequent history of astronomy.

The instrumentalist position is based first on the fact that it is not possible with naked-eye observation to determine the distances to the planets; with the exception of the Moon, they show no observable parallax. All we are able to observe is their direction in the sky, their positions on the celestial sphere. As historian Derek Price (1959, 200) has explained,

> all observations and hence all planetary theory was concerned only with the angular motion of the planets. Indeed it was concerned only with their apparent motion on and about the arbitrary unit circle constituted by the ecliptic.... We must not therefore make the mistake of thinking that the mathematical astronomers regarded the epicyclic loops traced out by the combination of deferent and epicycle as being in any way the real path of the planet in space. The orbit in

> space was not a question which could be resolved from observation alone, only by the importation of cosmological ideas not capable of experimental proof or disproof.

In the case of Ptolemy, evidence for his instrumentalism is found in his presentation of different geometrical models to explain the same motion, models that are clearly incompatible if they are regarded as material mechanisms to produce the motion in question. Consider the case of the Sun. As we saw earlier, Ptolemy developed a successful analysis of the variable speed of the Sun along the ecliptic using an eccentric circle. He also presented a second model for the solar motion using the concept of what is known as a secondary epicycle. In figure 3.6 the observer is located at E at the center of the deferent circle ABGD. The Sun lies on the epicycle ZKΘH, whose center is A. A moves on the deferent, while the Sun moves on the epicycle; the period of the two motions is the same and is equal to one year. The direction of rotation of the deferent is counterclockwise, while the direction of rotation of the epicycle is clockwise.

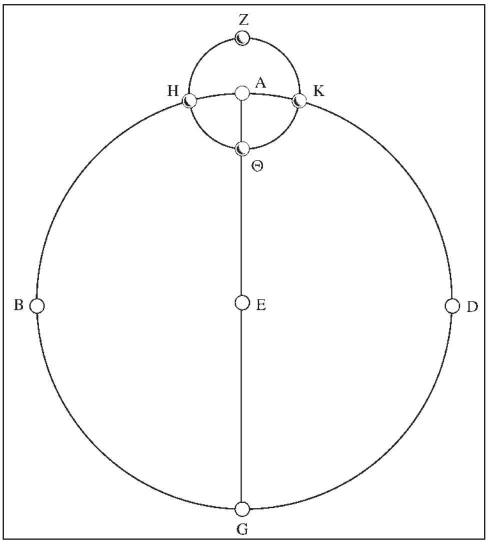

Figure 3.6: Model for solar motion using secondary epicycle.

We now have two geometrical representations of the motion of the Sun, the eccentric-circle model (figure 3.3) and the secondary-deferent-epicycle model (figure 3.6). Ptolemy was able to show by elementary geometry that if the radius of the eccentric circle is equal to the radius of the deferent, and if the eccentricity of the eccentric circle is equal to the radius of the secondary epicycle, then the two models are fully equivalent: the combined motion of the deferent-epicycle results in the same observed solar motion as the eccentric-circle model. The trajectory of the Sun in the two models is identical, although the motion is given by different geometrical constructions in each case.

In the first section of the 12th book of the *Almagest,* Ptolemy also introduced equivalent models for the superior planets, one the standard deferent-epicycle model and another model involving a circle—the eccenter—centered on a point that itself moves on a smaller circle, called the concenter. The planet moves on the eccenter with a period equal to its sidereal period, while the center of the eccenter moves on the concenter with a period equal to one year. The Earth is located at a point slightly offset from the center of the concenter circle. The two models are shown to be equivalent by elementary geometry. Ptolemy's presentation of two geometrical models for the motion of the Sun and for the motion of each planet has led some astronomers of the past and many modern historians of astronomy to conclude that his models were only mathematical devices and should not be interpreted as actual physical mechanisms to produce the planetary motion.

The case of the Moon raises even more direct issues because Ptolemy used a model that seemed to be inconsistent with observation. The variation in lunar distance in this model is at odds with his eclipse theory and also leads to values for the Moon's diurnal parallax that do not accord with observed values. It is also clear that if the Moon's distance varied so greatly, its apparent diameter in the sky should change markedly, and this is not observed to be the case. The English historian of astronomy J.L.E. Dreyer, a contemporary of Duhem's, found in the *Almagest* lunar model compelling evidence for Ptolemy's instrumentalism. Dreyer (1953, 196), writing in 1905, concluded,

> But though Ptolemy cannot have failed to perceive this [change in the apparent size of the Moon that occurs in the model] he takes no notice of it. It had now become a recognized fact, that the epicyclic theory was merely a means of calculating the apparent places of the planets without pretending to represent the true system of the world, and it certainly fulfilled its object satisfactorily, and, from a mathematical point of view, in a very elegant manner.

Although the instrumentalist position was popular among historians during the formative period in the study of ancient Greek astronomy, over the past 40 years, many commentators have shifted to a realist interpretation of Ptolemy's theories. (It is of interest to note that historians of non-Western astronomy still seem very much attracted to an instrumentalist interpretation of Greek astronomy.) While it is certainly true that the considerations raised above should be carefully weighed, these commentators cite counterevidence that is, on balance, compelling. First, and most obviously, there is the extended discussion at the beginning of the *Almagest,* in which Ptolemy attempted to justify a geocentric conception of the heavens. Ptolemy cited reasons traditional within Aristotelian philosophy and discussed the composition of the superlunary world (the world including and beyond the Moon) from the fifth, or perfect, element, ether.

Although it is true that there are alternative geometric constructions to explain the motion of the Sun, the actual trajectory of the Sun in the two models is identical. The existence of multiple models may caution us against assuming material reality for the parts of the model, but it is still meaningful to speak of a definite trajectory of the planet in three-dimensional space.

In book five of the *Almagest* Ptolemy included tables of the diurnal parallax of the Moon, and these values, in combination with eclipse data, were used to obtain estimates of the distances to the Moon and the Sun, measured in Earth diameters. The method in question originated with the astronomer Aristarchus of Samos, who lived in the third century B.C. This concern for the dimensions of the celestial system in the work of ancient Greek astronomers seems inconsistent with a purely instrumentalist understanding of planetary models.

From a realist perspective Ptolemy's lunar model may be seen as only a provisional and imperfect attempt to deal with complex irregularities in the Moon's motion. As historian John North (1994, 113) puts it, "If he [Ptolemy] noticed the variation [in the apparent size of the Moon]—and he could hardly have failed to do so—it must have been a great disappointment to him." According to this view, Ptolemy was striving to produce physically correct models, and his inability to find such a model in the case of the Moon would have been regarded by him as a defect to be remedied in some future revision of the theory.

ASTRONOMY VERSUS COSMOLOGY

Although today the instrumentalist interpretation of ancient astronomy is seldom defended in the strong form advocated by Duhem and Dreyer, there continues to be a vigorous body of historical writing that, while not explicitly embracing the positivist dogma, is nonetheless sympathetic with the older perceptions of Greek astronomy. This position is found in the writings of Otto Neugebauer, perhaps the greatest historian of ancient astronomy, and in the work of Neugebauer's former students and associates, Derek Price, Asger Aaboe, Bernard Goldstein, and others. The views of these historians are influential, in part, because of the preeminent contributions they have made to our technical understanding of ancient astronomy. In their opinion a distinction should be made between astronomy and cosmology, both generally in reference to Greek science and more particularly in the case of Ptolemy. Aaboe (2001, 116) asserts categorically that "a cosmological scheme is nowhere to be found in the *Almagest*." In an essay on the origins of the Copernican system Goldstein (2002, 219) advances the following principle: "it is important to distinguish astronomical issues, such as the use of the equant, from cosmological issues, such as the location of the center of the planetary motion and the order of the planets in space."

Underlying the notion of a rift between astronomy and cosmology is the view, expressed by Aaboe (2001, 71), that "the role of a geometrical model of the motion of, say, a planet is that of serving as a basis for computing the planet's position at a certain time in some relevant coordinate system" and that (116) "the principal aim of the *Almagest* is to enable you to answer the question: Given your location on the Earth, and given the time, in precisely which direction should you look in order to see a given celestial body?" Ancient Greek astronomy is devoted to the calculation of positions of planets as functions of time and includes eclipse theory; it is mathematical and is concerned with prediction. Cosmology attempts to identify the physical arrangement of the heavens; it is qualitative and is concerned with explanation.

One of the important historical discoveries over the past 50 years has been to identify the origin in Seleucid Babylonian astronomical tables of certain basic astronomical parameters used in the *Almagest*. Ptolemy obtained these values from Hipparchus's theory, and historical research has established fairly certainly that Hipparchus relied substantially on Babylonian data to determine the parameters of his solar and lunar models. Although the Babylonians used refined theoretical methods in the construction of their arithmetical tables, their astronomy operated without any underlying cosmology or geometrical model. There is a tendency to view the introduction of geometrical models by the Greeks as a purely mathematical move, an advance over the Babylonians in technical sophistication but essentially in the same vein as the earlier work. According to this view, the logical independence of Greek numerical astronomy from cosmology is a natural concomitant of methodological continuity in the history of ancient mathematical astronomy.

The idea that one should distinguish between astronomy and cosmology is also supported by arguments concerning the observational equivalence of heliocentric and geocentric models of the Sun and the planets. Consider a model in which the Sun moves about the Earth in a circle. (For the sake of discussion we neglect the eccentricity of the solar orbit.) The background celestial sphere to which solar observations are referred is assumed to be very distant, effectively at infinity. Then, the model in question could equally be regarded as a model for the Earth revolving about the Sun, with the Sun as it is observed from a moving Earth. The difference between the two models can be distinguished observationally only with respect to the presence or absence of parallax of more distant objects, and we have assumed the background reference frame is too far away for any effect to be observed. There are then two different cosmologies, one geocentric and one heliocentric, consistent with the same set of Earth-based observations.

There are also observationally equivalent but cosmologically opposed models for the motion of the planets. Figure 3.7 is a key element in the reasoning of advocates of the cosmology-astronomy distinction. Figure 3.7(a) depicts the standard deferent-epicycle model for the superior planet P. P is situated on an epicycle whose center C lies on a deferent with center O at the Earth. Draw a line through P parallel to OC and a line through O parallel to CP (figure 3.7(b)). In the resulting parallelogram, CO = PS and OS = CP. With S as center, draw a circle with radius SP and a circle with radius SO (figure 3.7(c)). We now put the Sun at S, obtaining a heliocentric model for the motion of P: P revolves about S in the larger circle of radius SP, and the Earth revolves about S in the smaller circle of radius SO. In terms of line-of-sight observations made from the Earth, it is simply a matter of mathematical convenience whether one adopts model 3.7(a) or 3.7(c) to calculate the position of P. The only way of distinguishing them would be evidence of the presence or absence of Earth's motion relative to more distant objects, and we have assumed that the stellar sphere is too distant for any parallax to be observed. The two models are equivalent as astronomical schemes for prediction but represent distinct cosmologies—in one the Earth is motionless at the center of the celestial system, while in the other the Sun is motionless at this center.

Advocates of the rift between astronomy and cosmology do not deny that Ptolemy adhered in the *Almagest* to an Earth-centered cosmology. In the first book he gave some of the standard Aristotelian arguments for a motionless Earth and discussed the constitution of the heavens in terms of the element ether. The claim is that these views were not particularly consequential nor were they connected in an essential way to the mathematical models used by him to analyze the positions of the planets. The predictive and mathematical

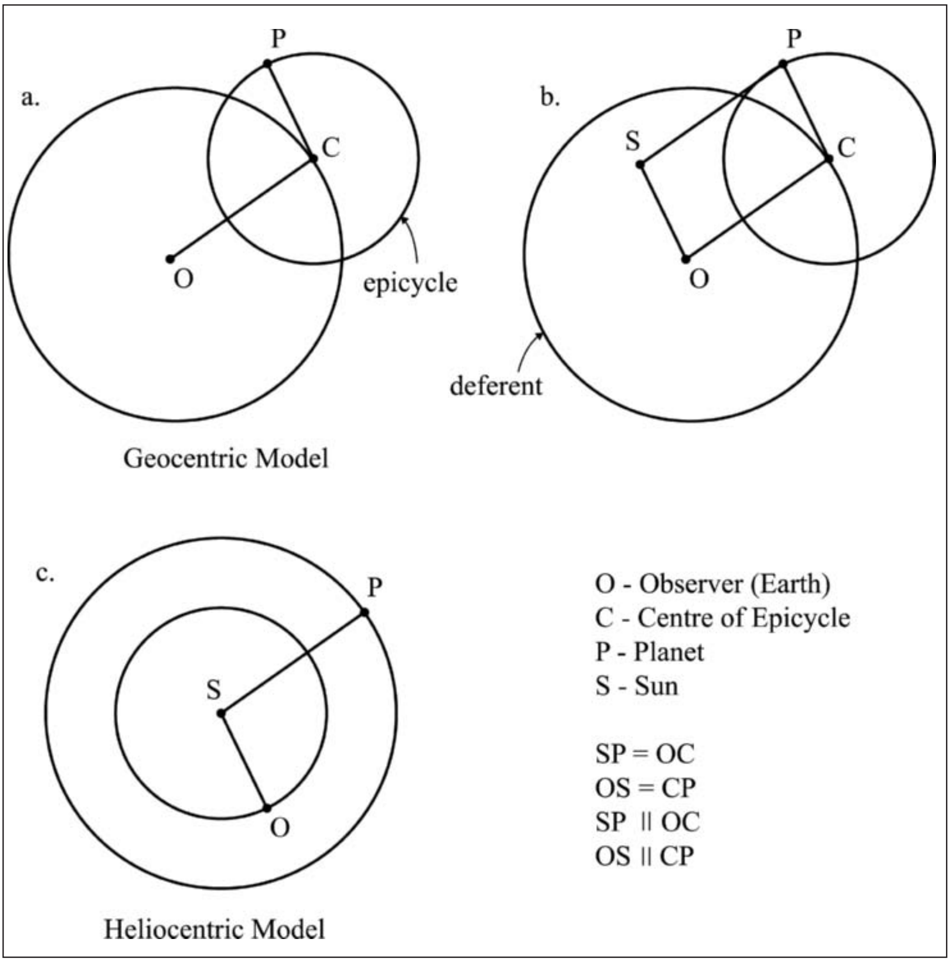

Figure 3.7: Equivalence of geocentric and heliocentric models for planetary motion.

point of view predominates in the *Almagest,* and this work was far more substantial than Ptolemy's other, more explicitly cosmological writings.

Ptolemy's models on a case-by-case basis are, it is argued, mathematically consistent with different and opposing cosmologies. If the essence of Ptolemaic astronomy is simply to predict the positions of the planets in the sky, then there is a definite sense in which one can say that there is no privileged cosmological scheme in the *Almagest.* However, although it may be the case that there is more than one physical arrangement that can correspond to a given model, it does not follow from this fact that the model is independent of cosmology. Indeed, the opposite is the case. The geometrical model can only succeed if it represents the actual planetary motion. While there may be more than one physical configuration for this motion consistent with the available observations and represented by the model, at least one of these configurations must be the true one, and the model must represent it. Each of the Ptolemaic geometric models is cosmological in this sense, and it is this fact that distinguishes the Greek work so fundamentally from its Babylonian antecedents.

It should also be noted that there are differences between Ptolemaic astronomy and heliocentric astronomy that are not simply a matter of the assumed reference frame of observation. The heliocentric system as observed from the Earth against the background of the stellar sphere gives rise to what is called the Tychonic system, a cosmology advanced by the Danish astronomer Tycho Brahe at the end of the sixteenth century. Brahe wanted to preserve the advantages of the then much-discussed Copernican system within a geostatic and geocentric universe, and he did so by assuming that the Sun and Moon revolve around the stationary Earth, while all the other planets revolve around the Sun. (See figure 3.8. We will discuss the Tychonic system in more detail in chapter 5.) It is not a historical accident that the Tychonic system came after Copernicus: it is simply the Copernican system as observed from the vantage point of Earth, but where the absence of annual parallax is attributed to the nonmotion of the Earth. The thesis that the *Almagest* models are independent of cosmology really reduces to the assertion that Ptolemaic astronomy, as it is developed in the *Almagest,* is compatible or mathematically equivalent to the Tychonic system. (Asger Aaboe suggests as much, citing as one piece of evidence Ptolemy's presentation in the twelfth book of the *Almagest* of equivalent models for planetary motion; the second of these models, though not strictly Tychonic, is nonetheless closely related to the Tychonic construction.)

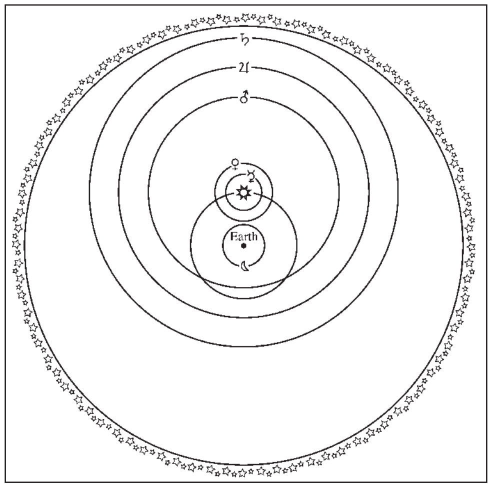

Figure 3.8: The Tychonic system.

Ptolemy, in the *Almagest,* accepts as fundamental the concept of planetary order with respect to the Earth. The planet Mercury is always closer to the Earth than the Sun is to Earth, and the Sun is always closer to the Earth than Mars is to Earth. Although Ptolemy seems to suggest in the ninth book that the actual order of planets may logically be arbitrary, the existence of an order itself is never questioned, either implicitly or explicitly: each planet has a fixed zone within which it alone moves, defined by its maximum and minimum distances from the Earth. One can speculate that Ptolemy believed in material spheres that moved the planets and that this physical conception explains his attachment to the concept of planetary order. Whatever the reason, this concept is fundamental to his account of astronomy in the *Almagest.*

In a Tychonic system, there is no place for such a concept of order: the distances from the Earth to Venus, Mercury, and Mars are constantly being shuffled as the planets move about the Sun and the Sun moves about the Earth. Ptolemy's adherence in the *Almagest* to a definite planetary order indicates his underlying allegiance to a different cosmological scheme, one in which the planets revolve around the Earth within shells formed by successive spheres centered on the Earth. This arrangement produces a definite and fixed order of position and distance with respect to the Earth.

A more technical objection to any cosmology-astronomy distinction that is based on an equivalence argument concerns the latitude theory developed by Ptolemy in the 13th book of the *Almagest.* The path of each planet is inclined slightly to the ecliptic so that the planet is sometimes north and sometimes south of the ecliptic. The angle of inclination varies from planet to planet. The planet's distance from the ecliptic at any instant is called its latitude; the planet reaches its maximum latitude at points 90 degrees from the intersection of the ecliptic and the planetary orbit. To account for motion in latitude, which is subject to a number of variations, Ptolemy assumed that the planetary epicycles are inclined to the deferents and that the epicycles wobble in a prescribed manner to produce the observed motion. Latitude theory is perhaps the most technically complex subject explored in the *Almagest.*

In the Tychonic system the Ptolemaic epicycle for each of the outer planets coincides with the Sun's orbit about the Earth, while the deferents of the inner planets coincide with the Sun's orbit about the Earth. If the Tychonic system is equivalent to the Ptolemaic system, then one would expect to find some common elements in Ptolemy's treatment of the latitude theory for each of the outer planets and similar corresponding common elements for the latitude theory of Venus and Mercury. Among other things, one would expect the epicycles of the outer planets to remain parallel to the plane of the ecliptic and the deferents of the inner planets to remain parallel to this plane. In fact, this is not the case, and Ptolemy, in the *Almagest,* develops a separate and distinct latitude theory for each of the planets. Whatever common elements that are present are necessitated by the need to save the phenomena (that is, to conform with observations) and not by any underlying equivalence of the models. Cosmological assumptions, in particular the assumption that the planetary epicycles are distinct entities characteristic of each individual planet, enter into Ptolemy's latitude theory.

PLANETARY HYPOTHESES

The thesis that the astronomy of the *Almagest* is independent of its ostensible underlying cosmological assumptions is provocative and, in an unqualified form, implausible. In the case of Ptolemy's other astronomical work, the *Planetary Hypotheses,* there can be no question of the logical priority of cosmology in the theory. This work was composed sometime after the *Almagest* and consists of two books, each divided into two parts. The first part of the first book survives in Greek, while the rest of the work is available only in an Arabic translation made in the ninth century. Furthermore, it was only in the 1960s that the second part of the first book came to the notice of modern scholars. The work as a whole was intended to explain the physical cosmology underlying the astronomy of the *Almagest* and to help instrument makers in the construction of models of the planetary system.

Part two of book one is devoted to an investigation of the dimensions of the planetary system. The planetary order adopted is the same as that of the *Almagest.* Moving out from the Earth at the center, one encounters successively the Moon, Mercury, Venus, the Sun, Mars, Jupiter, and Saturn. The dimensions of the planetary system follow from a nesting principle, according to which there can be no gaps, or empty space, between the spherical shells within which the planetary bodies move. Thus the maximum distance of the Moon from the Earth is equal to the minimum distance of Mercury, the maximum distance of Mercury is equal to the minimum distance of Venus, and so on.

Ptolemy's adherence to the nesting principle was based in part on the Aristotelian doctrine that there are no vacua in nature, that space without matter is quite literally inconceivable. More to the point, perhaps, the principle gave him a method for establishing the distance to each of the planets and thereby for determining the dimensions of the planetary system. The first step is to derive the distance to the Moon. The Moon is close enough to the Earth that it exhibits a diurnal parallax. It revolves about the center of the Earth, while it is observed from the surface of the Earth. The positions of the Moon given by tables derived from the lunar model are for a hypothetical observer situated at the Earth's center. The line from the observer to the Moon and the line from the center of the Earth to the Moon make a slight angle, the angle of parallax. This angle will evidently depend on the relative positions of the center, observer, and Moon during each 24-hour revolution of the Moon and is referred to as diurnal (or daily) parallax. The maximum value of this angle is called the horizontal parallax. Measuring the difference between the observed and predicted positions of the Moon, Ptolemy obtained a value of 1 degree, 26 minutes for the horizontal parallax of the Moon. (This value is larger than the true value because of inaccuracies of naked-eye observation and limitations of the theory.) By an elementary trigonometric calculation, he then determined that the distance to the Moon is 39.75 Earth radii. The radius of the Earth was known according to an established method originating several centuries earlier with Eratosthenes, which involved measuring the altitude of the Sun at noon at two points a known distance apart that lie on the same line of longitude. Hence the distance to the Moon was given absolutely in terms of a standard surveyor's unit of distance as defined by an observer stationed in Alexandria.

Beginning with the distance to the Moon, one may use the nesting principle and planetary parameters known from observation to obtain the distance to each of the planets and the width of the shell within which it moves. For example, suppose we have two adjacent planets. In the case of the planet closer to the Earth, suppose we know the radii of the outside and inside spheres making up the shell within which it moves. Consider the planet farther from the Earth, and suppose that the ratio of its epicycle radius r to its deferent radius R is e so that $r = eR$. (The value of this ratio is a parameter given from observation in the Ptolemaic model.) Then, the planet moves within the shell formed by spheres of radii R +eR and $R - eR$. The inner sphere will coincide with the outside sphere of the planet closer to the Earth, and by assumption we know the value R^* of the radius of this sphere. By equating $R - eR$ to R^* we are able to calculate R. Using this value, we proceed to the next planet out from the Earth and calculate the value of its deferent radius. The set of values obtained in this way must be fine-tuned to account for the eccentricities of the deferents. Table 3.1 gives the resulting dimensions of the planetary system in units of Earth radii. Note that the maximum distance of each planet is equal to the minimum distance of the next planet beyond it. The slight discrepancy in the case of Venus and the Sun, which would seem to allow for an empty space between them and a violation of the nesting principle, may be accounted for by rounding errors or slight adjustments that need to be made to the distances to the Moon and Mercury.

The value for the ratio of the distance to the Sun to the distance to the Moon is the same value as the one derived in the *Almagest* using the eclipse method of Aristarchus. (It is also equal to the value Aristarchus obtained using another method known as the method of lunar dichotomy.) The consistency of these results presumably strengthened Ptolemy's confidence in the system set out in the *Planetary Hypotheses.* However, the true value of the ratio of solar and lunar distances is some 16 times Ptolemy's value. The fact that he obtained the same wildly incorrect value by different and independent methods would seem to indicate that he manipulated the data to conform to theoretical desiderata. This tendency of Ptolemy's, which would be considered unacceptable today (in principle, if not in practice), was present in other parts of his astronomical work and is a well-documented aspect of his science.

The second book of the *Planetary Hypotheses* is devoted to a discussion of the physical structure of the planetary system. Each planet revolves within its epicyclic shell, while the epicyclic shell itself revolves within a deferent shell. Ptolemy indicated that it would be possible to replace the spheres composing the shells by tambourine-like disks. Although this would not be possible for the sphere of the fixed stars, which contains stars throughout its surface, the planets move in a fairly narrow disk aligned to the plane of the ecliptic. Whereas in the *Almagest* Ptolemy had presented two different models to describe the same motion, in the *Planetary Hypotheses,* only one of these models is given.

Table 3.1: Ptolemaic Planetary Dimensions. Distances given in Earth radii. From VanItelden (1985, p. 27).

Body	Least distance	Average distance	Least distance
Moon	33	64	48
Mercury	64	166	115
Venus	166	1,079	622.5
Sun	1,160	1,260	1,210
Mars	1,260	8,820	5,040
Jupiter	8,820	14,187	11,503
Saturn	14,187	19,865	17,026
Fixed stars			20,000

For example, the eccentric-circle construction for the Sun's motion is described using a physical model involving an eccentrically placed shell. The alternative description in terms of a concentric deferent and secondary epicycle is not mentioned, either as a mathematical or physical possibility.

Ptolemy rejected the Aristotelian doctrine according to which motion was transferred from the outer parts of the planetary system to the inner parts by means of an intervening set of rolling spheres. In addition to the question of mechanical difficulties in how such a transfer would be made, the conception did not fit easily within the deferent-epicycle theory developed by Ptolemy. The concept of a prime mover as the thing causing the motion of the outermost celestial sphere was also rejected. To explain the motion of each of the planets, Ptolemy instead made reference to the concept of an active planetary soul or intelligence that served to power and guide the planet's somewhat complicated motion. In the flight of a bird, messages in the form of sensations or impressions pass from the mental faculty of the bird through its nerves to its wings. The bird flies without any assistance or interaction with other creatures. Similarly, a planet moves itself. It possesses a soul or intelligence, and instructions on how to move the epicycle, deferent, and other circles involved in its motion are transmitted from the planet to the corresponding spheres.

Today, the idea of a planetary soul that determines or guides the motion of the planet may seem rather farfetched, but it is an idea that has much to offer. In the absence of any physical theory, such as gravitation, it explains how it happens that the planet comes to execute the many different motions that combine in exactly the right way to produce the observed motion of the planet. The idea was entertained seriously by Kepler in his astronomical research and may be regarded as a natural step in the sequence of ideas leading to a physical explanation of planetary motion. Consider the example of the Ptolemaic eccentric-circle model of the Sun's motion. The Sun is revolving in a circle in which the Earth is offset from the center. How is the Sun able to guide its motion in this circle since the center of the latter is an empty point in space? How does one coordinate the motion of a sphere about an empty point? One possibility would be to suppose that the Sun uses the apparent diameter of the Earth as a point of reference to calculate this center. Considerations such as these, and the difficulties and possibilities suggested by them, are the first step of an investigation leading to a study of the physical causes of planetary motion.

THE *TETRABIBLOS*

Ptolemy published another work relevant to cosmology, the *Tetrabiblos,* which deals with the influence of the celestial bodies on events on Earth. Astrology as a subject emerged in the later part of Greek antiquity, during the Roman Imperial period in the second century B.C. Earlier Greek thinkers such as Aristotle evinced no interest at all in astronomical prognostication. The popularity of astrology in the time of Ptolemy was connected to the emergence of Stoic philosophy and the Stoic concept of cosmic sympathy and belief in the interconnectedness of everything in the universe. The *Tetrabiblos* consists of the recitation of conventional beliefs rather than an exposition of reasoned doctrine or the derivation of theses. A prominent topic is what is known as judicial astrology, also known as horoscopic astrology, involving astronomical prognostication for an individual based on the position of the planetary bodies at the time of birth. Although it is difficult to gauge the level of Ptolemy's interest in astrology, the demands of a subject that enjoyed widespread popularity were a strong motive for the compilation of planetary tables. Just as divination from the entrails of animals contributed to the knowledge of physiology, so the interpretation of personality and the forecasting of events using celestial signs contributed to the study of astronomy.

There is one respect in which Ptolemy's conception of the physical influence of celestial bodies on the Earth is connected to his cosmology. According to Ptolemy, the different planets have a physical effect on Earth according to their positions in the planetary ordering, Moon-Mercury-Venus-Sun-Mars-Jupiter-Saturn. The Sun is the primary source of warmth, and the Moon is the primary humidifying agent. Mars and Venus both impart warmth to the Earth because of their closeness to the Sun, while Saturn has a cooling effect because of its distance from the Sun. The Moon serves to humidify, and Mercury also does so because of its proximity to the Moon.

CONCLUSION

Ptolemy believed that there was some mutual influence between the Earth and the heavens. The motions of the Moon and Mercury are the most complicated of all the planets because their spheres lie closest to the Earth, and their motions partake, if only partially, of the complexity of the terrestrial domain. Conversely, celestial bodies exert a physical influence on the Earth and serve to predict and explain terrestrial events. The existence of this mutual influence should not obscure the more fundamental antithesis of the terrestrial and celestial in ancient astronomy. This antithesis was emphasized by Aristotle and was reiterated by Ptolemy in the *Planetary Hypotheses.* Ptolemy called attention to the intrinsic differences between celestial motions on the one hand and the mechanical models constructed by artisans to represent these motions on the other. Although the motions of the planets may seem complicated, this is only because we are used to the friction and complications that are involved in terrestrial models and instruments. The world on Earth and the world beyond the Moon differ both in their physical constitution and laws of motion and constitute two irreducibly different domains.

4

COSMOLOGY FROM ISLAM TO COPERNICUS

ISLAMIC ASTRONOMY AND COSMOLOGY

The emergence and spread of Islam in the seventh and eighth centuries was followed by the establishment of enlightened institutions that actively encouraged the study of mathematics, science, and philosophy. In the Middle East, and later in Sicily, northern Africa, and Moorish Spain, the Arabic language became the medium for scholarship. The pursuit of astronomy fulfilled both astrological and religious purposes. In religion the Islamic calendar required a very accurate lunar-solar calendar, involving tables giving the days of first sighting of the crescent moon, an event that marked the beginning of a new month. An important question in Islam concerns the direction at a given location to Mecca, what is called the *Kibla,* which determines the orientation of morning prayers. The calculation of the *Kibla* required astronomical knowledge. There were also strong rationalist elements in the Arabic scientific tradition, involving the assimilation of Greek philosophy, especially Aristotle, within a more broadly theocratic cultural setting.

Mathematics in classical Greek culture developed into a mature field of study without any concern for astronomy. The three centuries of intense mathematical activity that culminated around 300 B.C. with Euclid's *Elements* constituted a primarily internal line of development centered on arithmetic and geometry. Astronomy in the sense of an exact science, as it was cultivated by Hipparchus and Ptolemy, took hold several centuries after Euclid. In India and the Islamic world, by contrast, astronomy was studied from the beginning in conjunction with mathematics. In India, mathematics was virtually coextensive with astronomy. Although there was some interest in pure mathematical research in Arabic science, the leading Islamic mathematicians were most often leading astronomers; spherical geometry and trigonometry were studied alongside the theory of numbers and Euclidean geometry.

Islamic researchers showed an interest in understanding Ptolemaic astronomy as a physical system of rotating material spheres. Until fairly recently, this interest was cited as evidence of a stronger physical orientation of Islamic astronomy in comparison with its more mathematical Greek antecedents. With the recovery of the complete edition of Ptolemy's *Planetary Hypotheses* in the 1960s it became apparent that much of this Arabic work was simply a continuation of themes from Ptolemaic cosmology. Nevertheless, it is still believed to be the case that Islamic researchers possessed a stronger sense of physical realism than their Greek forebears. Ptolemy's *Almagest* was criticized from a fairly early period for its abstract mathematical presentation of planetary theory. During the twelfth century, Aristotelian philosophy, and with it, physics, became a central concern of Islamic thinkers. Although Aristotle's homocentric cosmology as such led nowhere, it at least focused attention on achieving a coherent physical conception of the natural world. Much more significant developments occurred at Maragha (Iran) in the thirteenth century and Mamaluk, Syria, and Egypt in the fourteenth century, as researchers showed an active interest in modifying Ptolemaic kinematic models in order to produce physically plausible representations of planetary motion.

In the survey that follows we will concentrate on the cosmological views of several of the leading Islamic astronomers. It should be noted as well that Islamic scholars invented astronomical instruments such as the astrolabe and mural quadrant, established major observatories, and carried out important observational work. A substantial weakness of Ptolemy was in the area of observational astronomy. The observations reported in *Almagest* were, in several cases, simply derived by calculation from those made by Hipparchus several centuries earlier. (Ptolemy extrapolated Hipparchus's observations using the value Ptolemy had calculated for precession; because the latter value was slightly low, he obtained data that failed to correspond to what real observations would have yielded.) An important goal of Islamic astronomy throughout its history was the compilation of an astronomical manual with tables, what in Arabic is called a *zij*.

The House of Wisdom under the caliph al-Mamun and successive regimes in Baghdad became the center of ninth-century Islamic science. In its early development this science was influenced by contact with India. In astronomy, Indian influence was primarily in the area of observational data and mathematical technique rather than in the importation of cosmological notions. Al-Khwarizmi (d. 860) adopted some ideas from Indian mathematics—including a base-10 positional numeration system involving special symbols to denote the digits from one to nine and a symbol for zero—and used trigonometric methods of Indian origin to construct astronomical tables. Al-Khwarizmi's *zij* became the basis of several subsequent commentaries and revisions and was translated into Latin by Adelard of Bath in the early twelfth century.

Ishaq ibn Hunain (ca. 850–910) composed what was to become the most influential Arabic translation of Ptolemy's *Syntaxis Mathematicas,* which he named the *Almagest,* or "greatest." Thabit ibn Qurra (836–901) revised Ibn

Hunain's translation and carried out some important observational work. His study of precession led him to a theory of the trepidation of the equinoxes, according to which precession was believed to be a periodic variable in which the equinox oscillates back and forth about a given point on the ecliptic. Although mistaken and rejected by later Islamic astronomers, it remained influential and found supporters in the Latin West. Al-Farghani (d. ca. 850) was an expositor of the *Almagest* at an elementary level whose famous *Elements* became a very influential popular exposition of Ptolemaic astronomy. The most important astronomer of the ninth century, al-Battani (ca. 890), known in Europe as Albategnius, came from Mesopotamia and worked at an observatory in northeast Syria. He composed a major treatise on Ptolemaic astronomy, the *Kitab al-Zij,* a work that was translated into Latin by Plato of Tivoli (ca. 1125) in 1116 and was an important influence on the development of astronomy in the later Middle Ages. Al-Battani established that the solar apogee was not fixed in place, as Ptolemy had stated, but moves along the ecliptic with a steady motion, whose magnitude he was able to determine. He rejected the doctrine of the trepidation of the equinoxes and derived a very accurate value for precession.

The Islamic Ptolemaists adopted the general theoretical framework of the *Almagest* but tried to interpret it as an actual physical system. To this end, the physicist Ibn al-Haitham of Basra and Egypt (965–ca. 1040) wrote a treatise titled *Configuration of the World.* Al-Haitham became known in the Latin West as Alhazen and is best remembered for the book *Optics*, regarded as one of the greatest Islamic scientific works. In *Configuration of the World* he took as his ostensible starting point the astronomical system of the *Almagest,* although in fact the book consisted of an exposition of the subject matter of Ptolemy's other astronomical treatise, the *Planetary Hypotheses.* Scholars disagree on the question of whether he was directly familiar with the *Planetary Hypotheses.* However, there is no question that the contents of this work were known to him from some source, and he referred to it explicitly in his later writings. In his account of the motion of the Sun he described a shell whose outer surface is tangent to Mars's inner sphere and whose inner surface is tangent to Venus's outer sphere. Al-Haitham was here adopting the nesting principle from the *Planetary Hypotheses,* a principle that, like Ptolemy, he also applied to the Moon and the planets.

In a subsequent work, the aptly titled *Doubts about Ptolemy,* al-Haitham embarked on a critique of the Ptolemaic system in both its geometric form in the *Almagest* and its more physical development in the *Planetary Hypotheses.* This treatise was also an implicit criticism of his own earlier cosmological work. One of the things that came under attack was the characteristic Ptolemaic device of the equant. In the Ptolemaic model for a planet the epicycle moves on an eccentric circle, a circle whose center Z is offset slightly from the observer E on Earth. Furthermore, the center of the epicycle moves around the deferent with an angular velocity that is constant not with respect to Z but with respect to a point D, offset from Z on the opposite side of Z from

E (see figure 3.5, chapter 3). It is sometimes argued that the rejection of the equant by Islamic astronomers was a consequence of their strict philosophical adherence to the Platonic principle of uniform circular motion. However, there were also more concrete reasons for this rejection. It is not possible to regard the deferent as a rotating rigid sphere since such a sphere cannot rotate with constant angular velocity about a point not at its center. The position of the equant for some of the planets also raised difficulties. In the case of the planet Saturn its equant lay on the deferent sphere for Mercury, a situation that seemed physically untenable.

Although al-Haitham took exception with specific technical elements of Ptolemaic astronomy, he accepted the broader framework of this system, with its various eccentric circles and epicycles. A more radical critique of Ptolemaic cosmology is found in the writings of several Western Islamic philosophers in the twelfth and thirteenth centuries, who attempted to make Aristotelian philosophy an integral part of natural science. A leading figure here was Ibn Rushd (1126–1198) of Cordoba, who became known in the Medieval West as Averroës the Commentator for his extensive writings on Aristotle. On Aristotelian physical grounds Ibn Rushd boldly repudiated some of the central tenets of the Ptolemaic system. He reasoned that "the body that moves in a circle moves about the center of the universe and not exterior to it" (Arnaldez and Iskandar 1975, 4). Hence the motion of a heavenly body is motion about the center of the world, that is, the Earth. Motion on an epicycle is impossible because the center of the epicycle is located on the deferent and not at the center of the Earth. Similarly, the eccentric circle and the equant circle of uniform angular motion are mathematical constructions with no physical meaning because their centers are located away from the Earth.

Ibn Rushd advocated a return to the Eudoxan-Aristotelian cosmology of concentric spheres. As a project of technical astronomy, this idea was developed in detail by his near-contemporary, al-Bitruji (ca. 1190), the leading astronomer among the Spanish Aristotelians. Each planet exhibits two motions, its daily motion from east to west and a much slower motion along the ecliptic from west to east. The combined motion is therefore a westward daily motion that is smallest in the case of the Moon and greatest in the case of the outer planets, Jupiter and Saturn. According to al-Bitruji, the daily motion of each heavenly body results from the action of a ninth outer sphere, the so-called primum mobile. The action of this sphere weakens as it extends inward so that it is strongest in the case of Saturn and weakest in the case of the Moon, resulting in the pattern of motions observed in the planetary system.

Al-Bitruji conceded that his theory was only qualitative, an adjective that unfortunately can be taken to mean observationally crude. No astronomical tables of any value could come from such a system. The cosmology of Ibn Rushd and al-Bitruji, although of some later influence in the Latin West, was bound to be a complete failure. The Jewish scholar Maimonides (1137–1204) reacted to the extreme physicalism of his fellow Spanish Aristotelians by advocating an instrumentalist approach: the astronomer "does not profess to tell us the existing properties of the spheres, but to suggest, whether correctly or not, a theory in which the motion of the stars and planets is uniform and circular, and in agreement with observation" (Goldstein 1972, 41).

Following the conquest of southwest Asia by the Mogul conqueror Hulugu Khan in the thirteenth century, a major observatory and library were built at Maragha in northeast Iran. The leading astronomer at Maragha was Nasir al-Din al-Tusi (1201–1274). Al-Tusi made fundamental contributions to mathematics, including original work on the foundations of Euclidean geometry. His major contributions to astronomy consisted of a set of astronomical tables, known as the Ilkhani tables, and a major treatise on Ptolemaic astronomy, the *Tadhkirah,* or "treasury of astronomy." Al-Tusi aimed to modify Ptolemy's models to bring them into line with a more physically realistic cosmology, without compromising the predictive value of these models for empirical astronomy. In particular, he devised ingenious methods to avoid the use of the equant, which, following Islamic tradition, he saw as unsatisfactory. For this purpose he introduced what became known as the al-Tusi couple. Assume that a circle of radius *r* rolls on the interior perimeter of a larger circle of radius 2*r* (figure 4.1). During this motion a point on the perimeter of the smaller circle will trace out a straight line. In figure 4.1 the point O will move in a reciprocating motion back and forth on the line XX. By means of this construction, two circular motions are able to generate a straight-line motion, a fact in itself that seemed to challenge conceptually the Aristotelian opposition between rectilinear motion (terrestrial) and circular motion (celestial). Al-Tusi was able to use the couple device in a somewhat complicated way in order to represent the motion of the inferior and superior planets with suitable accuracy, without using the equant.

A sophisticated treatment of Ptolemaic's planetary theory was contained in the work of the fourteenth-century Damascus astronomer Ibn al-Shatir (ca. 1305–ca. 1375). The model al-Shatir devised for lunar motion seems to have been

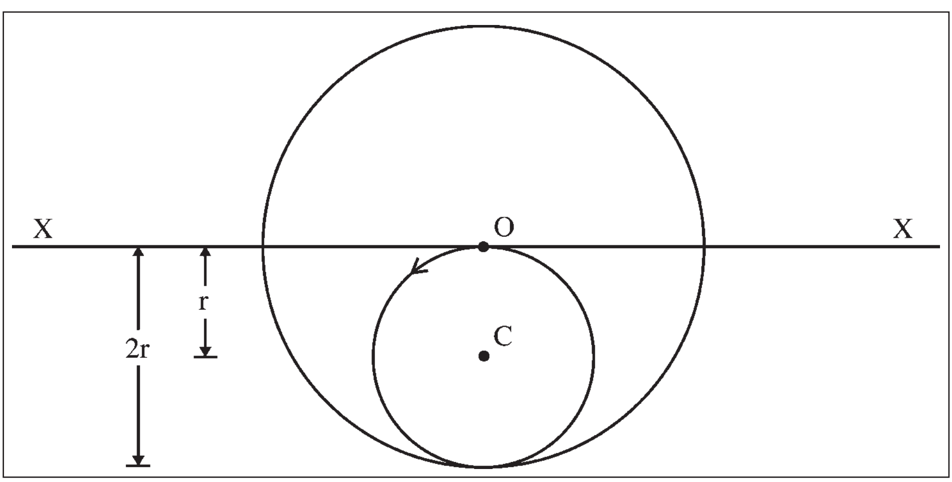

Figure 4.1: The al-Tusi couple.

designed to correct a specific difficulty with Ptolemy's model. In the *Almagest* Ptolemy had introduced a mechanism that drew the lunar epicycle closer to the Earth by having the center of the deferent rotate on a small circle about the Earth. There resulted a considerable change in the distance of the Moon from the Earth that should have been reflected in noticeable changes in the apparent size of the Moon, changes that in fact were not observed. Al-Shatir fixed the center of the deferent but added a secondary epicycle, making the Moon move on the secondary epicycle as the primary epicycle itself moved around the deferent (figure 4.2). The resulting model succeeded in saving the phenomenon but involved only a relatively small variation in the lunar distance. The work of al-Shatir was another indication of the tendency in late medieval Islam to develop physically realistic theories of planetary motion.

MEDIEVAL COSMOLOGY

In the centuries surrounding the decline and fall of the Roman Empire the level of understanding of astronomy and cosmology in Western Europe was very low, hardly rising above the literal interpretation of a few biblical pronouncements. The Earth was believed to be flat and situated at the bottom of the universe. Above the heavens were the upper waters mentioned in the

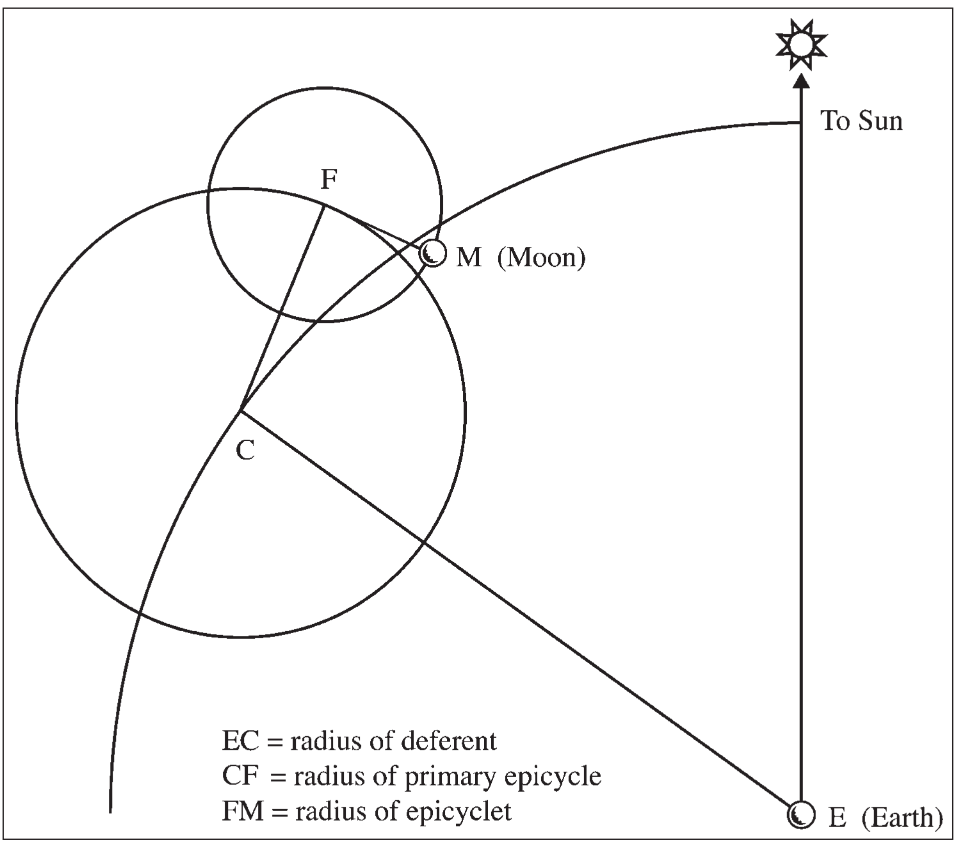

Figure 4.2: Al-Shatir's model of the motion of the Moon.

book of Genesis. Although some of the views of the Greek philosophers were known, they were rejected as inconsistent with biblical authority.

The growth of Christian institutions encouraged the preservation of at least a rudimentary level of Greek and Roman science. The Spanish bishop Isidore of Seville (560–636) composed an encyclopedia that included some basic cosmological and astronomical facts from Latin sources. An early eighth-century English monk, the Venerable Bede (673–735), carried out a detailed study of books brought from Rome to two monasteries in northeast England. Drawing from the writings of the Roman author Pliny, Bede taught that the Earth is a sphere, a fact confirmed by the experience of travelers from the observed variation in the altitude of the noon Sun as one traveled south. Bede also called attention to the sphericity of the heavens and identified the standard classical order of the planets as the Moon, Mercury, Venus, the Sun, Mars, Jupiter, and Saturn. His book *On the Reckoning of Time* became an important manual for chronology and the construction of calendars. The practice of dating events by the number of years that have elapsed since the birth of Christ was begun by Bede.

During the twelfth century a few industrious scholars, notably Gerard of Cremona (1114–1187) in Toledo and Adelard of Bath (1076–1160), produced Latin translations of Arabic scientific works as well as Latin translations of Arabic editions of Greek works. Euclid's *Elements,* Ptolemy's *Almagest,* and Aristotle's major philosophical works circulated in Latin. A translation of al-Farghani's *Elements* by John of Spain (1110–1180) helped to make Ptolemaic astronomy more widely known, and the more advanced work of al-Battani also circulated in the Latin West.

The classical geocentric cosmology of Aristotle and Ptolemy was described to European readers by Johannes de Sacrobosco (1195–1256), also known as John of Hollywood, who lived in Paris in the first half of the thirteenth century and wrote *On the Sphere* in 1220. This book was supplemented later in the century by the anonymous *Theory of the Planets,* which gave a more detailed account of the planetary theory sketched in the last part of *On the Sphere.* The latter work superseded al-Farghani's *Elements* and became the most widely read treatise on astronomy and the principal popular source for European views about cosmology for the next three centuries. The first part of the work contains a description of the Aristotelian physical universe:

> The elementary region, existing subject to continual alternation is divided into four. For there is Earth, placed as it were, as the center in the middle of all, about which is water, about water air, about air fire, which is pure and not turbid there and reaches to the sphere of the moon … and these are called the "four elements"….
>
> Around the elementary region revolves with continuous circular motion the ethereal … of which there are nine spheres, as we have just said: namely, of the Moon, Mercury, Venus, the Sun, Mars, Jupiter, Saturn, and the fixed stars, and the last heaven. Each of these spheres encloses its inferior spherically. (Thorndike 1949, 119)

Sacrobosco's *Sphere* articulated the physical views of Aristotle, while the *Theory of the Planets* provided an elementary account of Ptolemaic astronomy. During the thirteenth century Aristotle's writings on theology and philosophy became established as the main source of theoretical reflection about the world; with the Bible they formed the basis of the medieval scholastic canon. The two most influential interpreters of Aristotle were Albertus Magnus (1200–1280) and his student Thomas Aquinas, both members of the Dominican order and prolific authors. Aquinas formulated the principle concerning the relationship between religion and science that has guided Christianity ever since (with the exception a few fundamentalist sects), namely, that truths of faith and truths of reason will never come into conflict. In the history of astronomy, Aquinas is primarily remembered for his commentary on Aristotle's *On Heavens,* a work in which he brought Peripatetic astronomical doctrines into the compass of Christian medieval thought.

The translation into Latin of the astronomical works of al-Rushd (Averroës) and al-Bitraji (Alpetragius) gave rise in the thirteenth century to the adoption by some thinkers of Aristotle's cosmological scheme involving concentric spheres, in opposition to Ptolemy's system of deferents and epicycles. Albertus Magnus was one such figure, and the English scholar Robert Grosseteste (1168–1253) also rejected Ptolemy on various Aristotelian grounds. Although Aristotle's physics continued to provide the basis for all speculation about the physical world, his system of homocentric spheres was unsatisfactory: it was qualitative and unable to account for planetary motions. The schools of Paris and Oxford rejected Aristotle's cosmology, and during the later Middle Ages the Ptolemaic system was widely adopted, although not universally accepted, by commentators on astronomy. The rivalry (such as it was) between the two systems did become mixed up with a question of genuine interest, namely, the extent to which Ptolemy's models corresponded to physically real mechanisms in the heavens. Aquinas is noteworthy for having explicitly raised this question. According to Aquinas, the fact that Ptolemy's epicycles and deferents save the phenomena, that is, account for the observations, does not imply that they are real because it may turn out that there are other theories that account equally well for the phenomena. By contrast, Aquinas believed that in physics one arrives at physically true principles, for example, the principle that the natural motion of celestial bodies is uniform and circular.

The writings of Aquinas and Sacrobosco provide the background for one of the most famous works of Western literature, Dante Alighieri's (1265–1321) *Divine Comedy.* This poem is often cited for its symbolic literary integration of classical geocentric cosmology and traditional Christian thought. In it Dante is conducted by the Roman poet Virgil on a trip to the center of the Earth. Traversing the successive concentric rings of hell, the travelers encounter the devil in hell at the center of the Earth. They continue their travels to the opposite side of the Earth, where they scale the heights of purgatory, a pyramidal-shaped mountain. At the peak of purgatory Virgil is replaced as Dante's escort by Beatrice, a woman from Dante's youth representing idealized love, who accompanies him as they ascend upward to the successive heavenly spheres. The order of the spheres is the characteristic Ptolemaic order of the planets: the Moon, Mercury, Venus, Earth, Mars, Jupiter, Saturn, and the fixed stars. Beyond the fixed stars, there is a ninth sphere, the primum mobile, or prime mover, of all of the spheres, and beyond it, finally, the empyrean and the throne of God.

In *The Divine Comedy,* physical and moral-religious visions of the universe are intertwined. Man lives on the Earth in the outer part of the sublunary world, near the boundary between the corruptible terrestrial and the sublime celestial realms. Similarly, man is balanced morally between a striving for the good and an inclination to submit to vice and temptation. There is a symmetry between the concentric rings of hell within the Earth and the celestial orbits surrounding the Earth. The universe is a concrete, finite, and morally ordered world, and the central place of man in both a literal and spiritual sense is guaranteed as part of the natural order.

Dante was a literary figure who contributed nothing to astronomy or cosmology itself. In the fourteenth century the French bishop Nicole Oresme (1323–1382) developed new mathematical methods that would prove useful in natural philosophy. Oresme was perhaps the leading representative of late medieval scientific thought and someone whose ideas were to be an important stimulus for the early modern growth of science. In mathematics he invented the graphical representation of qualities, his so-called latitude of forms, which consisted of a mathematization of Aristotelian qualities. In the work of Oresme and his Oxford contemporaries, there was an erosion of the peripatetic distinction between physics, which concerns that which is corruptible and subject to change, and mathematics, which concerns that which is unchanging and eternal.

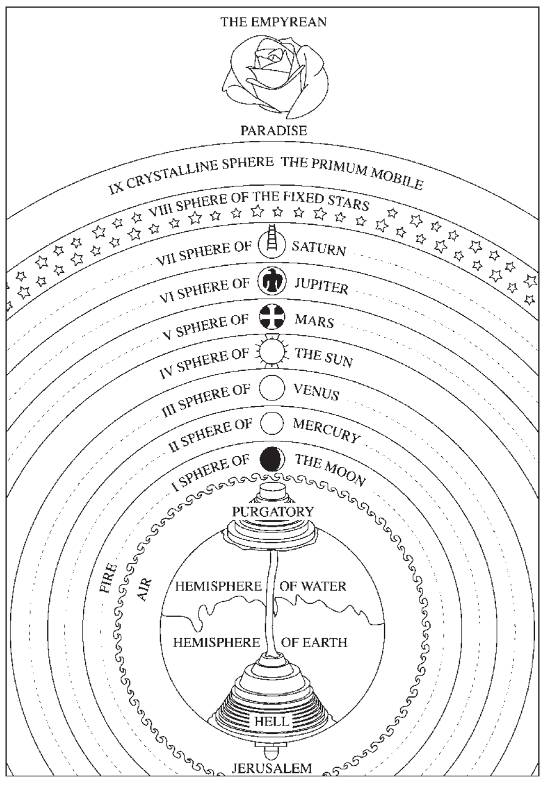

Figure 4.3: Dante's universe.

Although astrology continued to gain in popularity during the Middle Ages, it faced determined opposition from Oresme and other advanced thinkers. In his book *Commensurability or Incommensurability of Celestial Motions* Oresme attempted to counter a notion popular in astrology known as the "great return" or the "great year." This was an event of major significance that would take place when the Sun, Moon, and planets had returned once again to the exact same positions in the sky that they had occupied at some earlier epoch. Oresme considered the periods of rotation of the celestial bodies and examined their ratios or proportions. (The argument that Oresme developed applies to both the Aristotelian and Ptolemaic systems.) The theory of proportions was a central part of Greek mathematics and had, by the fourteenth century, become a subject of interest in theoretical mathematics in Latin Europe. Oresme reasoned that the ratio of the periods of rotation of any two celestial bodies was an incommensurable magnitude, in modern terminology, an irrational number. The ratio of the periods would never be equal to the ratio between two whole numbers. From this fact it followed that there would never be a time when the two bodies have exactly returned to any given earlier configuration, and the concept of the great return was an illusion.

However foreign Oresme's reasoning may seem from the perspective of today's science, it at least involved sophisticated mathematical considerations and showed a healthy aversion to astrological speculation. Oresme also took up subjects directly related to the subsequent history of cosmology. In a translation into French and commentary on Aristotle's *On the Heavens* he considered the question of the rotation of the Earth and the various arguments given by Aristotle for the immobility of the Earth. While finally accepting the general validity of Aristotle's views, his vigorous critique of peripatetic arguments was important in encouraging later thinkers to question the absolute correctness of Aristotelian physical doctrine.

COPERNICUS

In much popular writing on the history of science Copernicus has been portrayed as a heroic genius who overturned 2000 years of prejudice and launched a revolution in science that continues to the present day. The word "Copernicus" has entered the Western vocabulary to refer to that rare individual who initiates a truly fundamental change in some subject or field of investigation. Immanuel Kant was the Copernicus of philosophy for his introduction of the method of critical philosophy; Nikolai Lobachevsky was the Copernicus of geometry for his invention of non-Euclidean geometry in the nineteenth century; and Edwin Hubble was the Copernicus of modern cosmology for his identification of galaxies and universal expansion in the twentieth century.

This picture of Copernicus was largely based on a general understanding of what he had done and not on a particularly close study of his actual writings. As historians over the past 50 years have analyzed the original Copernican texts and documents, a rather different and less admiring portrait has emerged. A sharp critique was advanced by Arthur Koestler (1963, 201–202) in a book on the history of early modern astronomy:

> The figure of Copernicus, seen from the distance, is that of an intrepid, revolutionary hero of thought. As we come closer, it gradually changes into that of a stuffy pedant, without the flair, the sleepwalking intuition of the original genius; who, having got hold of a good idea, expanded it into a bad system, patiently plodding on, piling more epicycles and deferents into the dreariest and most unreadable among the books that made history.

Koestler was primarily a literary biographer rather than a scientist. Historians of science have also been very critical of Copernicus, including scholars who have studied closely the mathematical development of the heliocentric idea. Although he wrote for mathematicians and regarded himself as one of them, Copernicus displayed a lack of technical skill and creativity in comparison with Ptolemy. Some of the technical devices that he introduced were in fact almost certainly obtained from Islamic sources, with no acknowledgment of this fact, and he contributed relatively little to the field of observational astronomy. Although his contributions to cosmology were undeniable, even here it is suggested that his achievement was only the result of "a fortunate philosophical guess," to use a dismissive phrase of historian Derek Price (1959, 256). Copernicus is seen as the last medieval astronomer, someone who was mentally shackled by traditional conceptions of natural philosophy and slow to appreciate the possibilities opened up by his new system. His ideas entered the public domain slowly and through the efforts of friends; his famous book, written in a somewhat emotionless style, did not appear until he was on his deathbed. His reluctance to publish has been attributed variously to a fear of controversy, to his own recognition of technical weaknesses in his system, or simply to a weakness of character and an inability to assert himself as a man and a thinker.

The truth about Copernicus lies somewhere between the heroic portrait of popular history and the critical disparagement of contemporary scholarship. Copernicus worked in relative isolation in the single-minded pursuit of a great idea. Furthermore, unlike in modern cosmology, where the major advances have consisted of discoveries (often fortuitous) by skilled observers using advanced technology, Copernicus's achievement was an intellectual one carried out by a single individual working on the northern periphery of christendom. Whatever his limitations as a technical astronomer, he grasped the essential cogency of the heliocentric hypothesis and possessed the tenacity to pursue this idea to the end. That he did so in the circumstances of his time makes him a truly extraordinary figure of the history of science.

Publication of *On the Revolutions of the Heavenly Spheres*

Copernicus's interest in mathematics and astronomy was aroused at the University of Cracow, where he studied for several years in the 1490s. At the encouragement of his maternal uncle, the Bishop of Ermeland, he moved at the end of 1496 to northern Italy to train for a career in the church. Over the next nine years he studied canon law, medicine, and astronomy, obtaining a degree in canon law from the University of Ferrara in 1503. From 1506 until his death in 1543 he occupied the post of canon of the cathedral of Frauenberg in northern Poland by the Baltic Sea. Copernicus devoted his career to his administrative duties, to the practice of medicine, and to the pursuit of astronomy. A notable moment in his career occurred in 1514, when he was asked to participate in a project to reform the calendar. Although Copernicus declined on the grounds that the current state of knowledge of the motions of the Sun and the Moon was too uncertain to provide a reliable basis for calendrical reform, the request indicated that by this early point in his career he possessed a substantial reputation as an astronomer.

The heliocentric system was set forth by Copernicus in 1543 in his great work *On the Revolutions of the Heavenly Spheres.* It was written in a technical style and was aimed at specialists; as he wrote in the dedication, "Mathematics is written for mathematicians." *Revolutions* was preceded by the unpublished *Commentary* of 1530, an outline of the new system that enjoyed a limited circulation and helped make his ideas known to the community of astronomers. A young Lutheran scholar named Georg Rheticus (1514–1574) studied with Copernicus for two years, from 1539 to 1541, and became an advocate for the new astronomy. In 1540 Rheticus published an expository account of the Copernican system under the title *First Narrative.* It was at Rheticus's instigation and with the encouragement of Copernicus's friend Bishop Giese that the Polish canon carried out the final preparations for the publication of his book.

Copernicus came to the study of astronomy following a period of growing interest in the mathematical and observational work of Ptolemy, al-Battani, al-Biruni, and al-Tusi. In the fifteenth century, there was a general revival of astronomy in Europe, and important accounts of Ptolemaic astronomy were published by Georg Peurbach (1423–1461) and his pupil Johannes Müller (1436–1476), also known as Regiomontanus. The latter wrote the *Epitome of the Almagest* of 1496, a book that moved beyond commentary to the level of original research in astronomical theory and technique. Regiomontanus also compiled a major manual of trigonometry, a work that helped to establish him as the leading mathematician of the fifteenth century.

In the first book of *Revolutions* Copernicus called attention to some anticipations of the new astronomy in ancient Greek and Roman writings. It is fair to say that he was giving prominence to some fairly obscure sources in order to build a rhetorical basis for the presentation of his own new system. Copernicus referred to the Pythagorean Philolaus and to Heraclides (387–312 B.C.), the latter a pupil of Plato, who reputedly posited the rotation of the Earth as an explanation of the diurnal motion of the heavens. The late Roman writer Martianus Capella (ca. 470 A.D.) had suggested an alternative to the traditional ordering of the planets (Moon-Mercury-Venus-Sun-Mars-Jupiter-Saturn), suggesting that Venus and Mercury revolve about the Sun as the Sun revolves about the Earth. The proposed arrangement is an example of what is called a geoheliocentric system, the most famous of which is the more fully developed Tychonic system, considered in the next chapter. In comments echoing the sentiments of contemporary Hermetic and neo-Platonic authors Copernicus emphasized the special significance of the Sun in the universe.

Copernican System

The Copernican system was based on two distinct insights, both of which involved imparting a motion to the Earth: first, the daily 24-hour motion from east to west that all celestial bodies undergo may be attributed to the rotation of the Earth; second, certain striking and apparently unaccountable features of the Ptolemaic system can be explained by placing the Sun at the center and having the planets revolve about the Sun. These astronomical insights, and, in particular, the movement of the Earth that they implied, raised fundamental questions for traditional Aristotelian physics and would lead to new and revolutionary lines of investigation in natural philosophy.

There is a basic difference between ancient astronomy and Copernican cosmology that derives from the different interpretations in the two systems of the daily motion of the heavens. In the Ptolemaic system, each celestial body completes a revolution about the Earth in 24 hours, independently of its distance from the Earth. The fantastic speed with which the planets and sphere of fixed stars move implies a qualitative difference between them and objects found in our terrestrial world. Celestial bodies are composed of a mysterious and perfect fifth element, the ether, a conclusion largely derived from the fact of this daily motion. In ancient cosmology, there was a contrast between the world of the Earth and the world of the heavens based on the kinds of motion characteristic of objects in the two domains. By imparting a rotation to the Earth to explain the apparent daily motion of the heavens, Copernicus logically eliminated an assumption underpinning the traditional antithesis of the terrestrial and celestial domains.

It is worth noting that it was easier for astronomers of Copernicus's time to accept the daily rotation of the Earth than it was to accept its annual motion about the Sun. The Earth's rotation had already been discussed in some detail by al-Biruni and by late medieval writers such as Oresme. Independently of Copernicus, a professor at the University of Ferrara, one Celio Calcagnini (1479–1541), reasoned that it made more sense to suppose that the Earth revolves in 24 hours than to assume the entire heavens complete a revolution in the same period. Francesco Patrizio argued the same point later in the century, suggesting it was implausible to assume that the solid spheres of the planets and fixed stars could move with the incredible velocities required by the daily rotations stipulated in traditional cosmology. Both Calgagnini and Patrizio were otherwise firm believers in a geocentric universe. William Gilbert (1544–1603), the English natural philosopher and author of a seminal work on magnetism, was also convinced of the Earth's rotation, a motion he speculated was produced by the Earth's magnetic energies. On the question of the Earth's annual motion about the Sun Gilbert remained noncommittal.

Putting the Earth in motion about the Sun meant that the Earth was no longer the center of the universe and was just another celestial body. The fact that thinkers were willing to consider the Earth's rotation but not its annual revolution about the Sun indicates that the latter assumption represented a more radical departure from orthodoxy. The annual motion was the cornerstone of Copernicus's new world system. The motivation for this assumption derived from certain special features of the Ptolemaic system, in particular, the curious role in this system occupied by the Sun in relation to the planets. For each of the three superior planets the line joining the center of the epicycle to the planet always remains parallel to the line joining the Earth to the Sun. For the two inferior planets the centers of their epicycles always lie on the line joining the Earth to the Sun. Hence the planets and the Sun move about the Earth in a very specific way, a fact that simply expresses what is seen in nature and has no explanation.

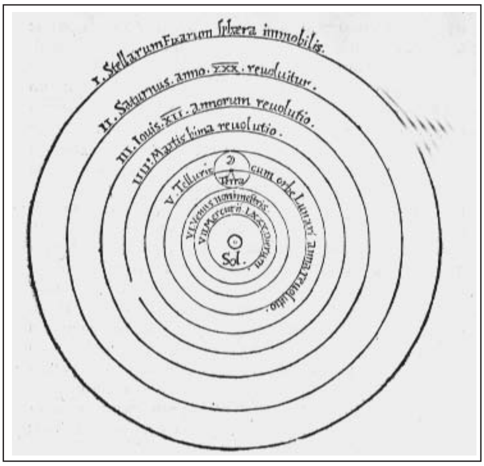

Figure 4.4: The Copernican system. The Thomas Fisher Rare Book Library, University of Toronto.

As we saw in chapter 3, there are simple transformations that relate the Ptolemaic and Copernican models of planetary motion. For each of the superior planets the epicycle of the planet becomes the Earth's orbit about the Sun, while its deferent becomes the planet's orbit about the Sun. In the case of the inferior planets the epicycle becomes the planet's orbit about the Sun, while its deferent becomes the Earth's orbit about the Sun. In the Copernican system the three epicycles for the superior planets are replaced by one circle, the Earth's orbit about the Sun, while the two deferents for the inferior planets are replaced by one circle, once again, the Earth's orbit. The Copernican system, which is depicted in figure 4.4 in an original illustration from *Revolutions,* possesses a definite economy with respect to the Ptolemaic system, having replaced three planetary epicycles and two planetary deferents by the Earth's orbit. There is, in the Copernican system, no longer the curious coincidence concerning the directions of the radii of the superior epicycles and the centers of the inferior deferents. The superior epicyclic radii point in the direction of the Sun because they are simply the radius joining the Earth to the Sun, while the centers of the inferior planets lie on the radius joining the Earth to the Sun because these centers coincide with the Sun, and the radii of the deferents coincide with the line joining the Earth to the Sun.

The naturalness and coherence of the Copernican system provided strong internal evidence in favor of the heliocentric hypothesis. There are also some indications (discussed in Goldstein (2002)) that Copernicus may have been motivated to develop a system in which the planets, as one moves out from the central body, decrease in angular velocity. In the Ptolemaic system this was the case for Mars, Jupiter, and Saturn; however, Mercury, Venus, and the Sun move about the Earth with the same average angular velocity. Although the Sun is farther than Mercury from the Earth, they both complete a circuit around the ecliptic in one year. In the Copernican system the rule of decreasing angular velocity is satisfied by all the planets.

In the Ptolemaic system the distances of the planets were regulated by the nesting principle, according to which there is no empty space between the successive spheres of movement of each of the planets. In the Copernican universe, there is no need to invoke such a principle: once the distance from the Earth to the Sun is set, all the other distances and dimensions of the system are determined. This fact is sometimes expressed by saying that Copernican astronomy has very natural and inherent system-like features.

Viewed purely as a work of geometrical astronomy, the Copernican theory was, in certain respects, characterized by a stronger sense of physical realism than its classical Ptolemaic counterpart. Copernicus seems to have been influenced by Islamic Ptolemaists, who modified some of Ptolemy's technical devices to make them physically more plausible. Of course, the Arabic treatises all supposed that the Earth was at the center of the universe. Because there were already such Islamic geocentric precedents, it is difficult to relate the heliocentric idea in and of itself with any particular push for a physically realistic astronomy. Nevertheless, an emphasis on producing mechanisms that were physically plausible as well as mathematically effective influenced how Copernicus developed his heliocentric scheme.

The Equant and the Earth's Third Motion

Ptolemy had introduced the equant to account for certain irregularities in the motions of the planets. The equant involved uniform angular motion of a circle about a point offset from the center of the circle. Although mathematically useful, it seemed difficult to reconcile with how material spheres actually move. As we saw earlier, Islamic astronomers devised ingenious techniques that allowed one to replace the equant by a combination of circular motions that were strictly uniform about their center. Copernicus was strongly opposed to the equant and attempted within his system of a heliocentric astronomy to produce mechanisms that avoided it. The basic innovation was to introduce a secondary epicycle to account for the small variations in motion that the equant was intended to produce. The end result of this modification of the original idea was a system with a substantially increased number of epicycles. According to some commentators, this fact diminished the essential economy and simplicity of the Copernican system.

In both the Ptolemaic and Copernican systems the motion of the Moon is geocentric, and it might be thought that there would be no significant differences between the two lunar theories. Nevertheless, the realism of Copernicus in comparison to Ptolemy is evident in the theory presented in book four of *Revolutions.* Ptolemy had supposed that the center of the lunar deferent was located on a small circle or eccenter whose center was the Earth so that the lunar epicycle was periodically drawn closer to the Earth by a kind of crank mechanism. Although effective in accounting for the observed variations in lunar position, this model had the disadvantage that the distance of the Moon from the Earth varied by as much as a factor of two, something that was at odds with the observed constancy in the size of the moon and could not be in accord with actual lunar distances. The model of the fourteenth-century Damascus astronomer al-Shatir, described above, had avoided this problem by placing the Moon on a secondary epicyclet. In this model the lunar distances varied within a much smaller range, and the lunar positions were also given with appropriate accuracy. In *Revolutions* Copernicus presented what was essentially al-Shatir's model. Although no mention of his predecessor was made, it is believed that he must have been familiar, if only indirectly from some sources, with al-Shatir's model.

Copernicus believed in the existence of material spheres that carried the planets about the Sun. One of the strongest pieces of evidence for this belief is the third motion Copernicus assigned to the Earth. Assume that the Earth is affixed in some way to a sphere that rotates in one year about the Sun. As the Earth is carried around the Sun, it rotates each day on its axis. It is evident that the direction of this axis of rotation will continuously change: if the axis is initially inclined at an angle to the axis of the ecliptic sphere, then it will, in the course of a year, trace out a circle on the celestial sphere whose center is the north ecliptic pole. If the northern hemisphere were initially inclined toward the Sun, it would stay inclined in this way throughout the year, and we would enjoy perpetual summer. Of course, observation reveals that the direction of the Earth's axis remains fixed on the celestial sphere at a point close to the pole star, a fact which explains the changing elevation of the Sun in the sky during the year and the occurrence of the seasons. Copernicus found it necessary to add a third motion to the Earth, a small conical movement of its axis, which has the effect of causing the axis of the Earth to remain parallel to itself as the Earth revolves about the Sun. In later astronomy the use of material spheres to produce planetary motion was found to be unnecessary, and the parallelism of the Earth's axis was understood to be a natural consequence of the Earth's inertial motion. The Copernican third motion of the Earth was rooted in medieval conceptions about how the planets moved on spheres, conceptions that were present in both Aristotelian cosmology and in the cosmology of Ptolemy's *Planetary Hypotheses*.

Nowhere are Copernicus's technical limitations more apparent than in the latitude theory developed in the final book of *Revolutions*. In reference to this theory Kepler wrote, "Copernicus, ignorant of his own riches, took it upon himself for the most part to represent Ptolemy, not nature, to which he had nevertheless come the closest of all" (Swerdlow and Neugebauer 1984, 483). By collapsing the superior epicycles and inferior deferents to the circle of the Earth's orbit, the Copernican hypothesis should have effected a substantial simplification in the Ptolemaic latitude theory. The orientation of the different planes could be reduced to their relation to one and the same reference plane, the plane of the Earth's orbit. Furthermore, the latter occupies no special conceptual place in the theory other than to act as a reference plane for the analysis of planetary motion. Nevertheless, Copernicus placed the center of the Earth's orbit at the center of each of the planetary orbits and essentially duplicated the Ptolemaic latitude theory for each of these orbits. As Kepler observed, he failed to take advantage of the opportunities and simplifications that his system afforded. In fairness to Copernicus, it should be noted that all observations are of necessity made from the Earth so that his assumption may be viewed as a pragmatic one resulting from the practical needs of observation. Furthermore, latitude theory is the most technically difficult part of planetary astronomy, and a satisfactory treatment of it would challenge not just Copernicus but his most skilled successors.

OSIANDER'S PREFACE TO *REVOLUTIONS*

It is asserted in the preface of *Revolutions* that the aim of the book is to advance a mathematical scheme to aid in the prediction of planetary motions and not to develop a theory of the planetary system as it actually exists in nature. Hypotheses such as that of the Earth's motion "need not be true nor even probable," but provide only "a reliable basis for computation." Although the purpose of the preface was apparently to forestall criticism of the philosophical or irreligious implications of the heliocentric system, it created some uncertainty concerning Copernicus's intentions. The preface was unsigned, and the reader might naturally assume that these were the words of Copernicus himself. In addition, some modern scholars, in particular those who incline to an instrumentalist view of the history of astronomy, have regarded the preface as a reasonable characterization of the contents of *Revolutions.* Thus Neugebauer (1968, 100) writes, "it is hard for me to imagine how a careful reader could reach a different conclusion."

It was later found that the preface had been written by Andreas Osiander (1498–1552), a Lutheran charged by Rheticus with the final stages of the printing of *Revolutions.* In fact, the preface is not at all in line with the organization, contents, and logic of the work. In the first section or book Copernicus presented several arguments to make the motion of Earth seem physically reasonable and thus to counter the traditional common-sense Aristotelian rationale for a stationary Earth. If it were simply a matter of developing an effective mathematical predictive scheme, such an excursion into the domain of natural philosophy would have been unnecessary. As we pointed out above, the technical modifications Copernicus made to traditional models were motivated by a desire to make them physically plausible and more than mere abstract calculational devices. Copernicus's introduction of the third motion of the Earth was motivated by his belief in a physical, material system of spheres, a belief that was grounded in a realistic conception of astronomy with ample medieval precedents.

The point at issue is connected to the origins of Copernicus's system and the place of this system in the history of cosmology. There are several explanations for how Copernicus was originally led to the idea of a moving Earth and a Sun-centered universe, most of them connected to the development of his system as a geometrical and mathematical theory. The first explanation was advanced by Copernicus himself and concerned the role of the equant in Ptolemaic astronomy. Copernicus stated that his purpose was to develop a system that used only motions that were circular and uniform about the center of the circle. Hence a primary concern of *Revolutions* is the elimination of the equant from planetary models. The difficulty with this explanation is that attempts to rid astronomy of the equant were already well established among Islamic astronomers, work with which Copernicus was familiar either directly or indirectly, and this astronomy was geocentric. Indeed, the decision to reject the equant was based primarily on considerations about how a rigid sphere would rotate, considerations that were independent of the choice of cosmology. Despite the prominence Copernicus gave to the question of the equant, it was of secondary importance in the actual decision to adopt a heliocentric system.

The most compelling motivation for heliocentric astronomy has already been discussed above. The Copernican universe is a natural one requiring no coincidences or special assumptions concerning the directions of the various radii of the system. To configure its dimensions as a cosmological system, there is no need to invoke a nesting principle and stipulate a particular planetary order. The distances of all planets are given in terms of the astronomical unit (the distance from the Earth to the Sun), and the order of planets is determined by the fairly natural condition that the angular motion of a planet decreases with its distance from the Sun. The special role of the Sun and the phenomenon of retrograde motion are simple consequences of the assumption that the Earth is a planet and, like the other planets, revolves about the Sun.

Whatever weight we assign to these different factors, there are two things about which we can be certain: the idea for the heliocentric system originated in considerations that were primarily astronomical and mathematical in character, and Copernicus believed that his system described the universe as it actually is, that is, his new system was a cosmology as well as a predictive scheme for saving the phenomena. The characterization of Osiander and such moderns as Neugebauer can only be regarded as incorrect. Given these facts, it was necessary to deal with the problem of developing and justifying an appropriate physics for a moving Earth. The first book of *Revolutions* is an attempt to ground a fact (the movement of the Earth) deduced from theoretical astronomy in the traditional domain of natural philosophy, that is, in Aristotelian physics as it had been elaborated by scholastic thinkers. Although many commentators have found this part of the book to be the least successful, Copernicus's goal at least indicated his primary commitment to the physical truth of the heliocentric system.

RECEPTION OF *REVOLUTIONS*

The invention of printing by Johannes Gutenberg (1398–1468) in 1475 resulted in a large increase in the number of books in circulation and greatly increased the distribution and influence of individual books. *Revolutions* was printed in 1543 in Nuremberg, a second edition followed in 1566 in Basel, and a third in 1617 in Groningen. Questions have been raised by scholars concerning how widely and how closely this very technical treatise was read, and its publication was not accompanied by any particular flurry of work on the kinematics of planetary motion. (Arthur Koestler [1963, 191] referred to *Revolutions* as "the book that nobody read.") Nevertheless, the first two editions, already preceded by Rheticus's expository *First Narrative* of 1540, ensured that the idea of the heliocentric system was disseminated widely throughout Europe in the second half of the sixteenth century.

Copernicus's procedure in *Revolutions* was to recast the Ptolemaic models within a heliocentric framework, and some commentators have suggested that he simply rewrote the *Almagest* from a heliocentric perspective. Copernicus's treatise was used as the basis for the construction of a new set of astronomical tables by Erasmus Reinhold (1511–1553) in 1551. The Prutenic tables superseded the traditional Alphonsine tables based on Ptolemaic theory. The publication of these tables attracted attention to the heliocentric system. Because Copernicus used more accurate parameters than Ptolemy, he obtained better results than those which would result from the *Almagest* theory, but such improvements could also have been achieved within an updated Ptolemaic system. Copernicus's insistence on modeling all motions out of a combination of uniform circular motions also meant that there were definite limits on the degree of observational accuracy that could be reached in his system.

The observational basis of the Copernican system would only be solidified later with the invention of the telescope. There were nevertheless certain observational facts that tended to support the heliocentric hypothesis. For example, Copernicus recognized that there were some problems with the Ptolemaic positioning of Venus below the Sun, with the fact that the distance to the Sun in the Ptolemaic system is supposed to be always greater than the distance to Venus. It was evident from the cycle of the Moon's phases that the light we receive from it is reflected from the Sun. It would be natural to suppose that the planets also shine by reflected light from the Sun, and the changing pattern of brightness of the outer planets is consistent with this hypothesis. In the case of Venus, however, there should, in the Ptolemaic system, be very considerable changes in its brightness. In particular, when it lies on points on its epicycle close to the line from the Earth to the Sun, it should appear very much dimmer than it does when it is at maximum elongation from the Sun. It was known from observation that there was only a fairly small variation in the brightness of Venus. In chapter 10 of book one of *Revolutions* Copernicus pointed out that it was necessary for those who placed Venus below the Sun to suppose that the planets shine by their own light, or at least emit light received from the Sun over their whole surface. Even in this case, there would be substantial changes in the brightness of Venus resulting from the large changes that occur in its distance to the Earth. Overall, the hypothesis of reflected light seemed more likely. The positioning of Venus relative to the Earth in the heliocentric system, where the decreasing brightness of Venus as it moves away from the Earth is compensated for by the greater visibility of its illuminated surface, would account for the constant brightness of Venus. This fact was confirmed observationally by Galileo in 1609, when he observed the phases of Venus through his telescope.

5

COSMOLOGY FROM BRAHE TO NEWTON

INTRODUCTION

It may have been the case that Copernicus's *Revolutions* reached only a limited audience, but it was read by and strongly influenced some important astronomers, the most notable being Tycho Brahe (1546–1601), Michael Maestlin (1550–1631), and Johannes Kepler (1571–1630). Although Brahe developed his own cosmology, the Tychonic system, his main ideas were derived from Copernicus. His historical importance derived from his observational work rather than from his cosmology. Following his teacher Maestlin, Kepler embraced the Copernican system and made the investigation of heliocentric astronomy his lifework. Kepler initiated a new line of reasoning in scientific astronomy by moving away from the traditional clockwork-like conception of planetary motion to a physical understanding of how forces are related to the motions they produce.

Copernican cosmology and the new physics were extended and solidified greatly in the writings of Galileo Galilei (1564–1642). His telescopic discoveries added to the plausibility of heliocentric astronomy, and his analysis of motion laid the foundations for terrestrial dynamics. These different threads of thought were synthesized in Isaac Newton's (1642–1726) great treatise *Principia Mathematica* (1687), a work that capped the early modern revolution in cosmology.

TYCHO BRAHE

Brahe was a Danish nobleman who carried out observations from 1576 to 1597 at his castle observatory Uraniborg on the island of Hveen between Denmark and Sweden. As a result of a dispute with the King of Denmark, he moved, at the end of 1597, to Prague to become court astrologer to the Emperor Rudolph II. He was assisted in the last years of his life by the young astronomer Johannes Kepler, who took possession of Tycho's treasure of observations when the latter died unexpectedly of a bladder infection in 1601.

In November of 1572 a new star appeared in the constellation Cassiopeia in the northern night sky. The star shone brightly for a month and then began to fade, disappearing from view altogether by March. The appearance of the star created a sensation in Europe, where it was regarded as an omen of great import for the affairs of man. Although the new star partook of the daily motion of the heavens, it was not clear whether it was an atmospheric phenomenon located in the sublunary sphere relatively close to the Earth or whether it was indeed a genuine celestial object. According to Aristotelian doctrine, the world beyond the Moon was eternal and changeless, a point of view that implied comets must be atmospheric. Tycho made a careful series of observations of the position of the new star throughout the night in order to determine whether it exhibited any diurnal parallax. The latter is a small shift in the apparent position of an object that occurs within a 24-hour period. (For an explanation of diurnal parallax, see chapter 2. Diurnal parallax is an effect that occurs in both a geocentric and heliocentric system.) The size of diurnal parallax decreases with distance from the Earth; it has the value of about one degree for the Moon and is indiscernible to the naked eye for the planets and Sun. In the case of the star of 1572 Tycho could detect no diurnal parallax nor indeed any change in position whatsoever of the new star with respect to the surrounding stars of Cassiopeia. It followed that the new star was located in the heavens some substantial distance beyond the Moon, an intruder in the celestial realm more like a star than a planet.

Tycho's new star was what became known in later astronomy as a nova, a star that undergoes a large change in brightness over a fairly short time period as a result of internal instability. Tycho suggested that the new star was something that had condensed from the surrounding matter of the Milky Way and even identified a dark spot nearby as its place of origin. Although this explanation was only speculative, the identification of the celestial character of the star was the first step in arriving at a scientific understanding of this class of objects.

Tycho applied the same technique of close observation four years later to a bright comet that blazed across the night sky. Tycho observed the comet during the three months it was visible and concluded that it showed no detectible parallax. This was a more difficult finding than the one for the 1572 nova because the comet actually moved across the celestial sphere, and it was necessary to show that any shift in position was due to this motion alone and not to parallax. Having established the superlunary character of the comet, Tycho charted its position on the celestial sphere and concluded that it was in orbit about the Sun. Over the next decade Tycho pondered the subject of the comet of 1777 and another comet of 1780. He consulted the writings on this subject of other astronomers, most notably Michael Maestlin, who had also derived a heliocentric orbit for the 1777 comet. The results of Tycho's investigation were published in 1588 in the book *Recent Phenomena of the Celestial World*.

In addition to his studies of new stars and comets, Tycho made several other fundamental contributions to observational astronomy. The large instruments that he had built for his observatory at Hveen and his skill in their use permitted a level of accuracy of observation hitherto unknown in astronomy. He showed that precession takes place at a constant rate, revealing the unsoundness of the doctrine of trepidation that had caused so much trouble for earlier astronomers. In his study of the motion of the Moon he identified the irregularity in its motion that is called the variation. He prepared a new star catalog with coordinates of the stars accurate to within three minutes of arc. His detailed observations of the positions of the planets would provide crucial data for Kepler's subsequent research and became the basis for Kepler's Rudolphine tables, published in 1627.

Tycho had used the Prutenic tables at a fairly early stage in his study of astronomy and lectured on Copernican astronomy in 1574 at the University of Copenhagen. Although he praised Copernicus as an astronomer, he could not accept the Copernican system. His opposition to the heliocentric idea was expressed in his correspondence with the German astronomer Christoph Rothmann, the latter himself a firm believer in the Copernican system. Tycho gave the usual common-sense physical objections to a moving Earth such as the change that should occur but is not observed in the range of a canon depending on whether it is fired to the west or to the east on a rotating Earth. He also cited astronomical, cosmological, and religious reasons. If in fact the Earth revolved about the Sun, then one would expect to observe what is known as annual parallax: at two times six months apart, when the Earth is at opposite ends of a line through the Sun, there will be a shift in the apparent direction of the stars. The size of this shift will depend on the distance of the stars relative to the distance of the Earth from the Sun. Tycho, following Ptolemy, placed the fixed stars just beyond the orbit of Saturn, not unreasonably believing that some parallax should be observed. Indeed, speculation since the time of Ptolemy accorded substantial values to the diameters of the stars, a line of thinking in keeping with a belief in their proximity to the planetary system. The observed absence of stellar parallax meant that in a Copernican universe, there must be a very large space between the orbit of Saturn and the celestial sphere. Such a wastage of space was scarcely in keeping with the handicraft of a divine creator, in Tycho's opinion. Finally, Tycho was convinced that the Copernican system was plainly contrary to the teachings of the Bible, a source from which he could cite multiple passages to support his position.

It should be noted that Tycho found much to admire in Copernicus's writings on technical astronomy. His admiration was influenced by what he regarded as the appropriate mathematical methods to be used in astronomy. He objected strongly to Ptolemy's use of the equant, the device in which the angular speed of the planetary epicycle is taken as constant with respect to a point slightly offset from the center of the deferent. He approved of Copernicus's move to eliminate the equant and replace it by alternate mechanisms, which mainly consisted of the introduction of secondary epicycles.

Having decided against Copernican cosmology, Tycho was also unable to uphold the traditional Ptolemaic system. Historians have suggested that his opposition to Ptolemy resulted from a desire to eliminate the equant from technical astronomy and that Tycho was led to his own system as a result of trying to adapt Copernican equant-less models to a geocentric theory. It has been conjectured (Westman 1975, 338) that such a goal motivated a whole generation of astronomers in the years following the publication of the *Revolutions.* A difficulty with this explanation is that Islamic astronomers had shown that it was possible to stay within the confines of Ptolemaic astronomy and do without the equant. There was no necessary connection between the technical goal of eliminating the equant and Tycho's decision to depart from Ptolemy and develop his own heliogeocentric system.

Tycho's comet studies gave him reason to find fault with Ptolemaic physical cosmology. In both the *Almagest* and the *Planetary Hypotheses* Ptolemy adhered in a fundamental way to a principle of planetary order, according to which each planet and that planet alone moved within a definite shell formed by two concentric spheres about the Earth. Ptolemy took Mars to be beyond the Sun in the planetary order, and so it followed that the distance from the Earth to the Sun was *always* less than the distance from the Earth to Mars. Furthermore, Ptolemy took the planetary spheres to be nested in such a way that there was no space between the spherical shells within which the planets moved. The most plausible explanation for Ptolemy's conception was that he believed the planetary spheres to be physical as well as mathematical objects. Tycho's demonstration that comets showed no diurnal parallax revealed that they were located beyond the sphere of the moon, and his study of their trajectories indicated that they were well within the orbit of Saturn. Evidently, comets must move across the zone of motion of the Sun, Moon, or one of the planets. It was impossible to avoid the conclusion that the material spheres of traditional astronomy did not in fact exist.

Tycho's ostensible reason for rejecting Ptolemy was based on difficulties he believed he had found with the relative distances of the Sun and planets in the Ptolemaic system. In both the Ptolemaic and Copernican systems the planet Mars is closest to the Earth when it is in opposition, 180 degrees opposite from the Sun in the sky. In the Ptolemaic system the sphere of Mars lies above the sphere of the Sun so that Mars is always more distant from the Earth than the Sun. Tycho accepted the traditional value for the solar parallax of three minutes, a value that indicated that the Sun was a fairly distant object from the Earth compared to the Moon. (Tycho was, of course, correct, even as he underestimated the size of the solar parallax by a good order of magnitude.) Through his observations in 1582–1583 of Mars while in opposition Tycho concluded that the parallax of Mars was larger than the Sun's, thus establishing that Mars at this time was definitely closer to the Earth than the Sun was. This finding was in plain contradiction with Ptolemy's ordering of the planets.

Brahe began to think about a third system of the world in the late 1570s and continued to work on the idea into the 1580s. He seems to have begun with the Ptolemaic system and then modified it in a series of steps in order to recover within a geocentric cosmology the advantages enjoyed by the Copernican system. He published the result of this investigation in his book of 1588. Like Ptolemy, he placed a stationary Earth at the center of the universe, around which revolve the Moon and the Sun. However, he followed Copernicus in assuming that the five planets revolve about the Sun as center so that each planet goes around the Sun as the Sun goes around the Earth (see figure 5.1). In the resulting geoheliocentric system the analysis of planetary motions is carried out along essentially Copernican lines but within a physical system predicated on a stationary Earth at the center.

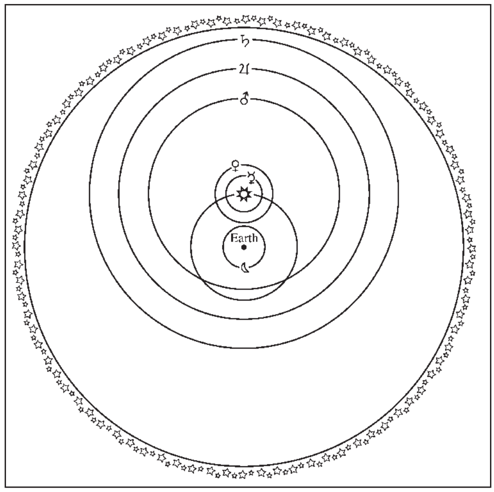

Figure 5.1: The Tychonic system.

In the Tychonic system the orbits of Mars and the Sun intersect, something that could not happen if they moved through the action of actual material spheres. Because Tycho had shown by his comet studies that the spheres do not exist, he was free to ignore this difficulty.

In upholding the tradition of an Earth-centered cosmos Tycho seemed to believe that evidence would be found in the class of new celestial objects that his observations had revealed comets to be. Certainly, the discovery of a sizeable number of Moon-like objects with geocentric orbits would strengthen the case for a geocentric cosmology. The difficulty here was that it was not a straightforward task to determine the orbit of a comet from a series of observations. In cases where Tycho succeeded in doing so, such as the comet of 1577, it was found that the comet's orbit was heliocentric. Tycho suggested that the absence of retrograde motion when the comets are in opposition was consistent with a geocentric orbit, but in the absence of detailed information about the nature of their orbits, such a consideration was hardly conclusive.

Tycho had shown that the motion of the planets cannot take place by means of the rotation of material and impenetrable spheres. Although this was surely his most important contribution to cosmology, he was not altogether consistent in his thinking on the subject. The primary objection to the equant was the fact that it was inconsistent with how a rigid sphere would rotate as it carries the epicycle around the deferent. Because Tycho had shown that such spheres apparently do not exist, this objection was no longer valid, and his own unwavering criticism of the equant seemed to lose much of its force.

The crucial event in Tycho's intellectual journey to the new cosmology was his putative discovery that the parallax of Mars at opposition is smaller than the parallax of the Sun. In fact, both values are an order of magnitude smaller than he believed them to be and are undetectable by even the most sophisticated instruments of naked-eye astronomy. It seems probable that Tycho, in a rather unconscious way, had come to see the plausibility of the Copernican system. For the purposes of positional astronomy the Tychonic system is simply the Copernican system as viewed from an observer on Earth. (When Kepler, a committed heliocentrist, carried out his analysis of the motion of Mars, he used a Tychonic reference system; the observations, after all, were made from the Earth, and the two systems, in this sense, are indistinguishable.) A number of writers of the period pointedly observed that Tycho's scheme was simply an "inverted Copernican" system. The heliocentrist Philip Landsberg suggested in 1632 that Tycho's scheme was "taken more from the diagram of Nicolaus Copernicus than from the heavens themselves" (Schofield 1981, 183). It is not surprising that the geoheliocentric idea occurred independently to several different authors following the publication of *Revolutions* in 1543. Tycho attributed physical meaning to his system and certainly believed in a geostatic cosmology, but this does not diminish the fact that the essential idea was taken from Copernicus.

The derivative character of the Tychonic system is illustrated by a diagram made by the young English astronomer Jeremiah Horrox (1619–1641) (see figure 5. 2). If one takes a diagram of the Copernican system and draws a circle with the Earth as center and the line from the Earth to the Sun as a radius, one is led to a diagram of the Tychonic system (figure 5.2). Another writer of the period, Otto von Guerricke (1602–1686), pointed out that one could take any of the planets as center, thus obtaining a Tychonic-type system corresponding to each planet (Schofield 1981, 408). Considerations of this sort revealed the inherent artificiality of the Tychonic system and its logical status as a frame-of-reference transformation of the Copernican system. Although Tycho believed that the Earth really was at rest at the center of the universe, it remained the case that any advantages his scheme enjoyed over Ptolemy's derived from its exact equivalence to the Copernican system.

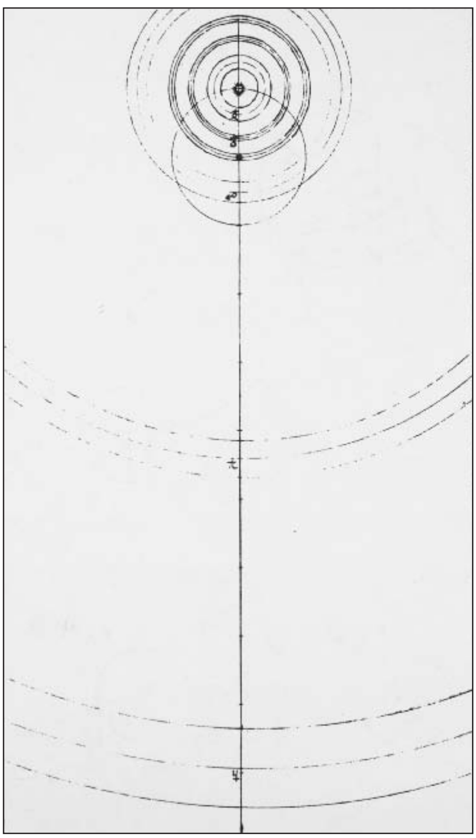

Figure 5.2: Tychonic and Semi-Tychonic World Systems. Figure from Schofield (1981, plate 16).

During the last years of his life Tycho became embroiled in a priority dispute with one Nicholas Reymers Baer (d. 1600), better known by his Latin name, Ursus ("bear"). To the end of his life, Tycho believed that his geoheliocentric cosmology was the most important thing he had done in science. Ursus published a book in which he developed a system very similar to Tycho's, with the exception that the Earth was allowed to rotate, and the orbit of Mars was placed outside the orbit of the Sun.

Ursus's book also contained some intemperate criticism of Tycho and other astronomers of the period. A bitter public dispute between Tycho and Ursus was ended only by the death of the latter in 1600. This episode has to be seen as a fairly odd quarrel given that the essential issue of priority concerned an idea which itself was derived from Copernicus.

Tycho's identification of the celestial character of novae and comets, his discovery of the lunar variation, his analysis of precession, the star catalog he compiled, and his detailed observations of planetary positions established him as one of the greatest observational astronomers of all time. His scientific career shows that it is possible to be a profound observationalist and an indifferent cosmologist. Despite its lack of originality, the Tychonic system was important in the history of cosmology for two reasons. First, Tycho freed cosmological theorizing from a belief in material spheres and the constraints imposed by the associated conception of simple clockwork-like production of planetary motion. Second, the Tychonic system allowed a form of disguised Copernicanism to flourish in an often hostile intellectual environment, allowing for a formal commitment to geocentric astronomy with the adoption of all the system-like advantages of the heliocentric theory.

KEPLER AND PHYSICAL ASTRONOMY

Kepler learned about heliocentric astronomy at the University of Tubingen, where he studied first as a student in the faculty of arts and later as a clergyman in training in the faculty of theology. Tubingen was a leading center of Lutheran theology, and Kepler planned to pursue a career in the church. One of his professors in Tubingen was Michael Maestlin, a supporter of the Copernican system and a prominent astronomer in the last decades of the sixteenth century. From his first encounter with the Copernican system Kepler became an advocate for the new cosmology, about whose correctness he seems never to have had any doubts.

At the recommendation of the Tubingen authorities, Kepler, in 1594, took up a position teaching mathematics in Gradz, Austria, a post that also required him to dispense astrological advice and prepare an annual almanac. In the course of a lecture in mathematics he was struck by what he took to be a deep connection between the heliocentric orbits of the six planets in the Copernican system and the mathematical properties of geometrical solids. He developed this idea in his book *Cosmographic Mystery* (1596), a treatise important for being one of the first astronomical works written from an avowedly Copernican viewpoint. Consider a geometrical solid or polyhedron whose surface is composed of plane polygonal faces. If the polygonal faces are all congruent to each other, then the solid is said to be a regular polyhedron. (We consider only convex polyhedra, that is, those without indentations.) For example, the cube is a regular polyhedron with six faces, each face being a square. It turns out that there exist only five regular solids: the tetrahedron (4 triangular faces), the cube (6 square faces), the octahedron (8 triangular faces), the dodecahedron

(12 pentagonal faces), and the icosahedron (20 triangular faces). The last part of Euclid's great book on geometry, the *Elements,* was devoted to a demonstration of this remarkable fact. Because the philosopher Plato discussed the regular polyhedra, the five solids are sometimes called the Platonic solids.

Consider now the Copernican planetary system, consisting of the six planets revolving about the Sun in circular orbits. For Kepler the fact there were five Platonic solids and six planets was no coincidence, as he attempted to show in his planetary cosmology. In the sphere corresponding to the orbit of Saturn, inscribe a cube. In this cube, inscribe another sphere. It turns out that the latter is a very close fit to the orbital sphere of Jupiter. Within Jupiter's sphere, inscribe a tetrahedron, and within the tetrahedron, inscribe a sphere; doing so, we obtain the orbital sphere of Mars. Within the sphere of Mars, inscribe a dodecahedron, and inside it, inscribe a sphere, thereby obtaining the sphere of Earth. Within Earth's sphere we place an icosahedron, and the sphere inscribed in it contains the orbit of Venus. Finally, within the sphere of Venus, inscribe an octahedron, and within it, inscribe a sphere, obtaining in this final step the sphere of Mercury. Figure 5.3 is taken from Kepler's book and depicts the resulting cosmographic system. For Kepler the nesting of the planetary spheres in terms of the Platonic solids was the key to the mystery of the planetary system.

Kepler spent a good deal of time and effort configuring various possible nestings of solids with spheres until he obtained one that worked. He was very proud of his geometric cosmology and arranged to have actual paper models built of the solids and nested planetary spheres. Although his subsequent research went off in different directions, he continued to believe throughout his career that he had discovered something important in his first book. Today, we are aware of the unsound and even fanciful character of the theory. In fact, there are nine planets and a host of minor planets or asteroids as well as more distant objects ranging beyond Pluto. There is no connection between the five Platonic solids and the planetary orbits, and the fitting that Kepler derived is fortuitous. Even for his own purposes and times, the cosmology could not serve as the basis for an empirical project to produce planetary tables. Johannes Praetorius, an older contemporary of Kepler and a believer in traditional geocentric cosmology, was critical of the *Cosmographic Mystery:*

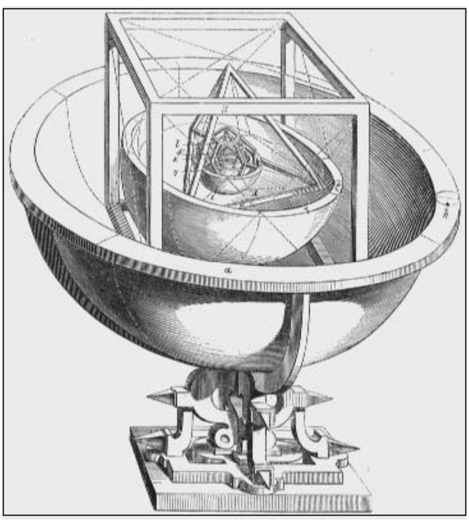

Figure 5.3: Kepler's heliocentric system (1596). The Thomas Fisher Rare Book Library, University of Toronto.

> But that speculation of the regular solids, what, I beg, does it offer to Astronomy? It can (he says) be useful for marking the limits or defining the order or magnitude of the celestial orbs, yet clearly the distance of the orbs are derived from another source, i.e., *a posteriori,* from the observations. And, having defined

> these [distances] and shown that they agree with the regular solids, what does it matter? (Westman 1975, 303)

Despite its speculative character, Kepler's research on geometric cosmology possessed definite scientific value. Through it he gained valuable experience in the numerical analysis of the planetary orbits and dimensions of the Copernican system, experience that would serve him well in his later investigations. The *Cosmographic Mystery* also established Kepler as a young astronomer of promise and served to promote Copernicanism as an astronomical theory.

Kepler's most important book was the *New Astronomy* (1609), one of the great scientific classics of the seventeenth century. In it Kepler set about analyzing Tycho's observations of Mars in order to determine the exact geometric shape of its orbit. The book marked a clear break with medieval thinking about cosmology and helped to launch modern physical theories of the universe. It is written in an autobiographical style in which the various turns and false steps in the investigation are chronicled in lengthy detail. The book offers an unusually revealing study of the process of discovery in mathematical science, something that normally has to be reconstructed from draft notes or various pieces of circumstantial evidence. It is believed that Kepler wrote in this way so that a skeptical reader, critical of the new cosmology and celestial physics, would see for himself how one would be led to the remarkable conclusions of his investigation.

The eccentricity of Mars (the deviation of its orbit from a circle) is, with the exception of Mercury, larger than that of the other planets, and its motion is more difficult to reduce to a combination of circular motions. Kepler had gained access to Tycho's observations of Mars at a time when Tycho had restricted access to his full set of planetary observations. Kepler was fortunate to have found a problem that would lend itself to a solution by new methods and for which there was ample reliable data from which to work.

The *New Astronomy* marked a radical new approach to the analysis of planetary motion. In order to understand the nature of Kepler's innovation, it is helpful to consider earlier planetary models. Take as an example the Copernican model of the Earth's motion. The annual motion of the Sun occurs along a great circle on the celestial sphere, and the speed of the motion varies in a regular pattern. Assuming that the Earth moves on a circle with constant angular velocity, and assuming that the Sun is located at rest at a point offset from the center of the circle by a small distance, the observed solar motion will result. All we have to do is determine two parameters: an angle that gives the orientation of the line joining the center of the circle and the Sun with respect to the celestial sphere and the distance of this line as a fraction of the distance from the Earth to the Sun.

The eccentric-circle model successfully accounts for the observations, in the sense of showing that if the planet moves in the way specified in the model, then we will indeed see what we see: the model has excellent predictive value. The model also provides what would appear to be a very reasonable physical representation of how the planet actually moves in the heavens. However, there is no explanation of why or how the planet moves as it does in the model. There was the older concept of a planetary soul, according to which the planet possesses a sort of intelligent soul that powers and directs the different spheres involved with its motion. Copernicus was silent on this part of traditional cosmology, perhaps because he felt that the underlying conception of a soul contributed little to the primary goal of predictive astronomy. For him the planetary system resembled a rather complicated clockwork mechanism, in which the spring, or weight, and regulator were hidden from view; given that the various spheres were positioned in definite ways and rotated with specified speeds and periods, the observed motions of the parts would follow.

Unlike Copernicus, Kepler was working from a perspective in which the planetary spheres had been shown not to exist. He needed to investigate how it comes about that the planets moved in the way that they did. He seriously considered the possibility of a planetary soul or intelligence as the efficient cause of planetary motion. Each planet would possess both an innate power of motion and an intelligence that would direct this power. In the case of the Earth moving around the Sun the Earth's soul would use the apparent size of the Sun as an indicator, enabling it to maintain the proper distance to the Sun in its eccentric-circle orbit. (For Copernicus, this was unnecessary since the Earth was carried around on the eccentric sphere.) In order to account for Tycho's very accurate observations, Kepler found it necessary to use the equant to model the Earth's motion. This was a novel step since Ptolemy, in the equivalent solar model of the *Almagest,* had not used the equant, the latter being employed only for the five planets. Although Kepler found that a model with an equant was in fact not good enough to account for the observations, it served as a useful working hypothesis in his investigation. As more complicated motions were posited, it became very difficult to understand how a planetary intelligence would be able to perform the difficult calculations required in order to direct the motion.

To model the motion of Mars, Copernicus had replaced Ptolemy's equant by a secondary epicycle that produced a motion about as good in agreement with observation as Ptolemy's had been. In general, Kepler found the use of epicycles objectionable since it was unclear how the planetary intelligence could direct the motion of the center of the epicycle, which was only a mathematical point. He therefore reintroduced the equant, this time for the heliocentric orbit of Mars, and attempted to determine the exact relation between the Sun, the center of the orbit, and the equant. Even here, it was not easy to understand how the planetary intelligence could direct the motion of Mars in such a model.

To supplement the action of the soul, Kepler advanced the idea of a causal physical connection between the Sun and the Earth: the motion of the Earth about the Sun is seen to result from the physical action of the Sun on the Earth. There is a direct causal link between a central physical agent and the orbit described by the planet. As his investigation progressed, Kepler came to dispense with the intelligent soul altogether and cast his analysis solely in terms of the Sun's physical action on the planet. Historian E. J. Dijksterhuis

(1961, 310) sees Kepler's shift from the conception of an intelligent soul animating planetary motion to an inanimate process involving solar force to be one of the most profound moments in the mechanization of the world view that occurred during the Scientific Revolution. Like other scientists of the period, Kepler was influenced by William Gilbert's (1544–1603) *On the Magnet* (1600), a work that was notable for its methodical and experimental investigation of magnetic phenomena. Kepler conjectured that the Sun acts on the planets in a way analogous to the magnetic action of a lodestone on iron. The Sun rotates on an axis perpendicular to the plane of the ecliptic, setting in play a series of concentric, rotating filaments that propel the planets in their orbital trajectories. Through a process of reasoning involving some very questionable steps Kepler arrived at the conclusion that the line joining the Sun to the planet sweeps out equal areas in equal times. Using the area law, he deduced, after a series of attempts, that an ellipse with the Sun at one focus was the best fit to the orbit of Mars. His analysis of this planet contained two facts that he would later formulate as general laws of planetary motion: the area swept out by the radius from the Sun to the planet is a linear function of time; and the orbit of each planet is an ellipse with the Sun at one focus.

By a complicated route involving a mixture of geometric and physical reasoning Kepler had managed to derive facts of the utmost importance about the planetary orbits. Although he did not succeed in developing this idea into a coherent theory of celestial dynamics, he had at least initiated a line of investigation that would eventually lead to a satisfactory dynamical theory of planetary motion.

GALILEO GALILEI AND THE NEW PHYSICS

Galileo was born in Pisa in 1564, the eldest son of a prominent musician and musical theorist. He assumed the chair of mathematics at the University of Pisa in 1589; this was followed three years later by a professorship at the larger and more prestigious university in Padua, where he remained for 18 years. As a result of his astronomical discoveries with the recently invented telescope, Galileo, in 1510, was appointed as "first philosopher and mathematician" to the Grand Duke of Tuscany in Florence, a position he occupied until his death in 1642.

In his student days and first university appointment in Pisa Galileo seems to have been a follower of Ptolemy, although he had become aware of Copernicus and, even at this early stage, was critical of Aristotelian physics. In a letter to Kepler in 1597 Galileo stated that he had long accepted the Copernican hypothesis, and certainly, from this year on he became a public advocate of the new astronomy. However, his primary interests up to 1508 were not in astronomy but in mechanics and the study of the natural motion of falling bodies and pendula.

In 1609 Galileo learned of work in Holland on an observing instrument constructed from a tube and ground lenses that made distant objects appear larger. He set about constructing a telescope of his own and trained it on the heavens. His exciting first telescopic discoveries were reported in the *Starry Messenger* of 1610; important further discoveries would follow in the next few years. Galileo's resolution of the Milky Way into stars and his identification of a large number of faint stars was a significant contribution to the new field of stellar astronomy, a development we take up in the next chapter. In terms of planetary astronomy his telescopic observations provided several pieces of evidence to support the heliocentric hypothesis. This evidence was essentially circumstantial, but taken together, it made for a rather persuasive case. The discovery of craters and mountains on the Moon and spots on the Sun raised questions for the traditional Aristotelian doctrine of the immutability and perfection of the heavens and challenged the supposed fundamental qualitative difference between terrestrial and celestial domains in a geocentric universe. The fact that stars exhibited no appreciable diameters under telescopic magnification indicated that they were very far away, thus explaining why no annual parallax is observed as the Earth revolves about the Sun. The phases of Venus could be naturally explained in the Copernican system and constituted a substantial difficulty for the Ptolemaic system. The discovery of the moons of Jupiter showed that satellites travel with their primary body and thus nullified the geocentrist's objection that if the Earth moved, it would leave the Moon behind.

Galileo's contributions to astronomy were nowhere near as important as those of Copernicus, Tycho, and Kepler, but they did play an important role in advancing the cause of heliocentric cosmology. His most important work consisted of his contributions to the science of mechanics. In his greatest work, *Discourse on Two New Sciences* (1638), Galileo laid the foundation for the mathematical analysis of the motion of bodies moving under the action of gravity on or near the surface of the Earth. He derived laws to describe the natural motion of falling bodies, bodies moving down inclined planes, and the motion of pendula as well as the trajectories of balls fired from canons. His approach was based on the physical concept of inertia and on the use of rigorous mathematical techniques for the analysis of motion. The careful observation of nature and even some degree of experimentation were also characteristic of his investigation. The mathematization of space was carried further by Galileo's contemporary, René Descartes (1596–1650), who devised a form of graphical analysis involving the reduction of geometrical relationships to equations of algebra.

Inertia is the tendency of a body in the absence of any external agents to move indefinitely with constant speed in a straight line. Its formulation by Galileo and Descartes presupposed an idealized and abstract conception of the physical interaction of bodies. The emergence of inertia as a fundamental concept was associated with a change in understanding of how the mechanical interaction of bodies is related to the universe at large. Although this change did not occur in a complete form in Galileo's science, he had begun a conceptual transformation that would be largely complete by the end of the seventeenth century. The finite, closed world of Aristotelian physics was replaced by a physics involving the analysis of bodies moving in Euclidean geometrical space. As historian Alexandre Koyré (1957) has emphasized, a profound conceptual shift had occurred to an open and mathematically (if not physically) infinite universe governed by precise laws.

Because of his advocacy of Copernican astronomy, Galileo, in the last part of his life, came into serious conflict with church authorities in Rome. In the early seventeenth century the Catholic hierarchy was actively concerned with defending church doctrine from Protestant criticism. There were also disputes within the church itself between Dominican and Jesuit orders about the interpretation of scientific facts. During this period the church became more insistent on a literalistic interpretation of the Bible, and it was well known that the heliocentric hypothesis was in open conflict with many statements from the Bible. In the 70 years or so since the publication of Copernicus's *Revolutions,* this work was increasingly seen not just as a treatise on mathematical astronomy but also as an exposition of the true cosmology of the world.

Theological opponents of Copernicus found allies among the university philosophers who wished to defend Aristotelian natural philosophy against the criticisms of Galileo and his followers. In 1616 the church issued an official decree forbidding any books that treated the motion of the Earth or the stability of the Sun as real. *Revolutions* was placed on the Index, the list of books considered dangerous to faith and forbidden to Catholics. Other books of a Copernican bent, such as an outspoken work by the Carmelite father P. A. Foscarini (1565–1616), were also put on the Index. In February of 1616 Galileo was summoned to an audience with Cardinal Bellarmine (1542–1621) to discuss his advocacy of the Copernican hypothesis and his public statements about faith and science. The Council of Trent of 1543 had assigned authority for the interpretation of the Bible to the church fathers, so regardless of the merits of Galileo's arguments, he was on weak legal ground. What exactly took place at the meeting is the subject of controversy. Galileo later stated that it was only stipulated that Copernicanism "cannot be defended or held," while the official church records report that he was issued an injunction not to "hold, teach or defend [Copernicanism] in any way whatsoever, verbally or in writing" (Koestler 1963, 463).

With the arrival of the new pope Urban VIII in 1623 Galileo made an unsuccessful attempt to have the decree of 1616 rescinded. He did secure permission to write a book about the Ptolemaic and Copernican systems, providing that these be presented only as hypothetical schemes to describe planetary motion. In the years that followed he worked on a book dealing with planetary astronomy and the effects of the tides. Galileo was convinced that the tides provided evidence of the rotation of the Earth and developed a detailed explanation of them. His *Dialogue on the Two Chief World Systems,* published in 1632, was written in Italian in the form of a conversation between an Aristotelian, a Copernican, and a third gentleman, who was receptive to reasoned and plausible argumentation. Ostensibly an impartial discussion of the Ptolemaic and Copernican systems, it was composed in a way that underlined the strengths of heliocentric cosmology and the weaknesses of geocentric cosmology. No reader would have any doubts about the author's preferred cosmology. Galileo's engaging literary style and his choice of Italian as language ensured that the book would find an audience beyond the universities and church.

Pope Urban VIII and the ecclesiastical authorities in Rome were angered by the advocacy of the heliocentric hypothesis as physical truth evident in Galileo's book. In one of the most famous episodes in the history of science Galileo was summoned in 1633 to Rome and tried by the Congregation of the Holy Office, or Inquisition. At the trial, much was made of Galileo's meeting with Cardinal Bellarmine in 1616, where he was served with the injunction neither to hold nor teach the Copernican hypothesis. The issue at the trial was not the truth of this hypothesis but whether Galileo had violated the injunction. Galileo was found guilty of heresy, forced to make a humiliating confession of the error of his scientific beliefs, and confined indefinitely to house arrest. Galileo spent the last eight years of his life confined to his residence in Florence, where he worked on his masterpiece of mathematical mechanics, *Discourse on Two New Sciences.*

Galileo's trial is generally viewed as a very regrettable event in the history of the church, although there have been apologists for the Holy Office's position in the affair. In 1992 Pope John Paul II addressed the matter and admitted that errors had been committed by the theological advisors to the church hierarchy of Paul V and Urban VIII. Whether there was, in principle, a conflict between the new cosmology and Christian belief is a matter of debate. The Catholic canon Copernicus had dedicated *Revolutions* to the pope, and Kepler was a devout Christian who held that his astronomical discoveries were evidence of God's design. Although in Revelation 7:1 it is written, "I saw four angels standing on the four corners of the Earth" (King James Version, 1611), the roundness of the Earth was accepted in the seventeenth century by educated Europeans. As Galileo (1957, 196) himself noted, it was not in the interest of the church to take positions on scientific questions for it would be "a still greater detriment to the minds of men … to see a proposition proved that it was heresy to believe." By the middle of the eighteenth century, if not earlier, scientists of every Christian denomination had come to accept the truth that the Earth was a planet like other planets and revolved about the Sun. The challenge posed by heliocentric astronomy to religious faith was much less radical than the questions raised in the nineteenth century about what it means to be human by biological theories of evolution and the ascent of man.

ISAAC NEWTON AND THE NEWTONIAN SYNTHESIS

Newton studied at Cambridge University and was appointed professor of mathematics there in 1669. In 1687 his greatest work was published, the *Mathematical Principles of Mathematical Philosophy,* a work that is commonly known as the *Principia* from the first word of its Latin title. The *Principia* was the crowning achievement of the revolution in cosmology and physics that began one and a half centuries earlier with Copernicus. It contained the "Newtonian synthesis," a mathematical dynamics of forces acting on bodies that unified the inertial physics of Galileo and the heliocentric astronomy of Copernicus and Kepler. Kepler's ideal of a celestial physics was finally realized in the theory of universal gravitation presented in the *Principia*.

The chain of events leading to the publication of the *Principia* was a visit by astronomer Edmund Halley (1656–1742) to Cambridge in 1684. Halley asked Newton what would be the magnitude of the force exerted by the Sun on the planets, given that the planets revolved about the Sun in elliptical orbits. Newton immediately replied that such a force would vary inversely as the square of the distance to the planet. At Halley's encouragement he began to compose a systematic mathematical analysis of the action of a central force (a force that originates in a point) upon one or more particles. The resulting tract would become the core of book one of the *Principia*.

In the third book of the *Principia*, titled "System of the World," Newton took the mathematical theory from the first book and applied it to the solar system. The planets were considered as a system of point masses acting on each other by the force of gravity. The fundamental law that all bodies satisfy was the universal law of gravitation: every two bodies attract each other by a force that is proportional to the product of their masses and inversely proportional to the square of the distance between them. Although Newton did not think that gravitation was an intrinsic property of matter, he also did not attempt to speculate about the underlying physical process by which it acts. He would "feign no hypotheses" (Jammer 1969, 98) concerning the nature of gravity but simply investigate its action according to the inverse-square law.

Some of Newton's contemporaries objected to the idea of action at a distance across empty space and sought a mechanical explanation of gravity in terms of particles of matter interacting by contact or collision. Proponents of this point of view tended to be followers of Descartes. The fluid-dynamical concept of a vortex was used to model the motion of a planet about the Sun. The solar system was filled with a very fine fluid that rotated rather like a whirlpool about the Sun, propelling the planets in their orbits. The Cartesian mathematical physicist Christiaan Huygens (1629–1695) further elaborated the vortex mechanism in an attempt to explain the action of gravity near the surface of the Earth.

Newton believed that a fluid-dynamical explanation of planetary motion was untenable, a fact he attempted to document in his study of the motion of bodies in a resisting fluid in book two of the *Principia*. During the eighteenth century, there was considerable interest in the vortex theory of planetary motion, but it was eventually abandoned. It proved difficult to derive mathematical laws that described the vortex action, and some of the main predictions of the theory were found to be false. By 1750 Newton's theory of universal gravitation based on the inverse-square law had triumphed throughout European scientific circles.

Gravity is fundamental to cosmology because it acts between any two bodies anywhere in the universe. Among the different fundamental forces of physics, gravity is the only one that acts over the great distances of interest in cosmology. Newton wrote that gravity "must proceed from a cause that penetrates to the very centres of the Sun and planets, without suffering the least diminution of its force … and propagates its virtue on all sides to immense distances, decreasing always as the inverse square of the distances" (Koyré 1957, 228). A theory of gravitation underpins any attempt to describe the universe as a whole and therefore is basic to all attempts to produce a scientific cosmology.

With the consolidation and acceptance of Newton's theory the main unresolved problem was to show that the solar system as governed by the inverse-square law of gravitation was in fact a stable dynamical system. It was known from a comparison of ancient and modern observations that no disturbance increasing indefinitely with time had occurred in this system. During the eighteenth century, researchers developed methods of increasing mathematical sophistication to analyze the gravitational interactions of systems of three and more bodies. By the 1790s Simon Laplace (1749–1827) was able to apply this theory to the three-body system consisting of the Sun, Jupiter, and Saturn and proved that there were no perturbations of the system that increased with time over the long term.

FOUNDATIONS OF DYNAMICS

Near the beginning of the *Principia* Newton interrupted his presentation of the mathematical theory to discuss at some length the basic concepts of his new dynamics. He was writing a pioneering treatise on "natural philosophy," something that was still not quite mathematical physics in the modern sense, and it was not surprising to find such explanatory remarks about the foundations of the subject. Newton's intent was to provide compelling evidence for the absolute character of space and time.

A fundamental law of Newton's mechanics asserts the proportionality of the force acting on a body to its acceleration. If a body is moving with constant velocity, then its acceleration is zero, and the total force acting on it is zero. This fact is related to something called the restricted principle of relativity, also called the Galilean principle of relativity because a version of it was formulated by Galileo. In studying motion we measure the velocity of a body with respect to some time taken as given and with respect to some object taken at rest. This velocity is the relative velocity of the body measured with respect to the rest object. If we consider a system of bodies interacting in any way and impose a uniform velocity of translation on the whole system, then the acceleration and therefore the force of this motion is zero, and it follows that the interactions of the bodies remain unchanged. For example, on a ship sailing with a steady velocity on the open sea the objects within the ship interact mechanically just as they would if they were situated on land.

Motion measured with respect to a reference object is relative motion. Newton held that there is also such a thing as absolute motion, motion measured with respect to absolute time and absolute space. Given two objects moving with a nonzero velocity relative to each other, it follows that the motion of at least one of the bodies is true and absolute. The existence of absolute motion follows because forces are real, and forces are proportional to acceleration. Because the acceleration of a body is associated with a real tangible force, this acceleration must occur as part of an absolute or true motion and cannot be something that is only measured according to some convention with respect to a group of reference objects.

The principle of the relativity of motion does not hold for acceleration because acceleration brings with it dynamical effects. Newton illustrated this fact using examples of circular motion, where centripetal forces act within the system. Because Jupiter rotates on its axis, it is flattened at the poles; it has the shape of what is known as an oblate spheroid. The rotation gives rise to centripetal forces that vary differentially over the surface of the planet, producing the distortion from a sphere. If one were to assume that Jupiter were at rest and the world system were revolving about it, no forces would act on Jupiter, and it would have the shape of a pure sphere. The accelerative motion of Jupiter is something that is absolute, taking place with respect to absolute time and space and producing real forces acting on the planet. Newton also illustrated this point with the example of two spheres joined by a string, in rotation about their common center of mass. In such a system, there is a force of tension in the string; if all the rest of the matter in the universe were removed, there would still be this force acting. It follows that the circular motion of the spheres is an absolute motion. Accelerations are connected to forces, and forces are real and not conventional. Although accelerative forces indicate the existence of absolute motion, it is nonetheless difficult to determine the precise motion of any given object with respect to absolute space and time. In "System of the World," book three of the *Principia,* Newton formulated the hypothesis that the center of this system was at rest with respect to absolute space. He took this center to be the center of gravity of the solar system. In effect, Newton identified the whole universe with the solar system, the fixed stars, being uniformly distributed and being very distant, not significantly affecting the position of this center. His belief in an absolute point of rest has been seen as a sign of the influence of residual geocentrism, of an inability to follow some of the implications of his new theory to their logical conclusion.

Newton's views on absolute motion were criticized by his philosophical contemporaries Gottfried Leibniz (1646–1716) and George Berkeley (1685–1753) and have since become the subject of an extensive philosophical literature. In terms of the reception and success of Newtonian mechanics in the eighteenth and nineteenth centuries, these views did not play an important role. In using mechanics to investigate planetary motion, the strength of beams, the vibration of elastic strings, or the behavior of flowing fluids it is unimportant whether one supposes the motion to take place with respect to something called absolute space. Newton seems to have emphasized the absoluteness of space and time partly for psychological reasons. The conception of motion in a universe of absolute time and space was the natural development of the new picture of the world developed by Copernicus and Galileo. An open and potentially infinite universe based mathematically on absolute scales of space and time stood in contrast to the closed relational universe of Aristotle and Ptolemy. The two cosmologies represented two understandings of how God orders the universe: on the one hand, the intimate and literal world depicted in Dante, and on the other, the austere, absolute, and mathematical universe of Newton.

Analysis of such fundamental concepts as space and time would become a subject of interest to physicists in the late nineteenth century, and Newton's conception of absolute motion would be severely criticized, particularly by the physicist Ernst Mach (1838–1916) in his 1883 book *The Science of Mechanics; A Critical and Historical Account of Its Development*. These criticisms would influence Einstein in his development of the general theory of relativity. In this way the philosophical issues raised by Newton in his discussion of absolute space and time would come to play a role in the evolution of modern theories of the universe.

6

STELLAR ASTRONOMY: THE UNIVERSE BEYOND THE SOLAR SYSTEM

DIMENSIONS OF THE UNIVERSE

Prior to the sixteenth century, conceptions of the dimensions of the universe referred primarily to the dimensions of the planetary system, and thinking about this was largely based on the calculations contained in Ptolemy's *Planetary Hypotheses.* Although this book itself may not have been widely available in later history, the cosmology presented in it was disseminated indirectly through Arabic and Latin sources. The Ptolemaic dimensions began with calculation of the radius of the Earth using Eratosthenes's method, combined with a measurement of the distance to the Moon obtained from the value of its diurnal parallax. The distances to the Sun and planets were derived from Ptolemy's nesting principle, which stipulated that the outer sphere of a given planet's shell was at the same distance as the inner sphere of the shell of the next planet. The order of the celestial bodies going outward from the Earth was the Moon, Mercury, Venus, the Sun, Mars, Jupiter, Saturn, and the sphere of the fixed stars. On various grounds Ptolemy placed the sphere of the fixed stars only slightly beyond Saturn, at a distance of 20,000 Earth radii. (The closest star to the Earth is today known to be at a distance of over 6,000,000,000 Earth radii.)

The advent of the Copernican system resulted in a complete overhaul of planetary dimensions since the dimensions of the entire system were determined by geometry once the distance from the Earth to the Sun was given. Table 6.1 compares the Copernican values with modern distances, showing that the original Copernican system provided a very good fit to the actual dimensions of the solar system. It should be noted that the scale of the Copernican system was roughly comparable to the Ptolemaic system as far as planetary distances were concerned. However, in heliocentric astronomy the fixed stars have to be placed at a very large distance beyond Saturn in

Table 6.1: Planetary distances according to Copernicus and modern theory. Distances are given in astronomical units (one astronomical unit equals the distance from the Earth to the Sun). All distances are averages. (Copernican distances are from van Helden (1985), and modern distances are from Olcott (1954).)

Planet	Copernicus	Modern
Mercury	0.3763	0.3875
Venus	0.7193	0.7234
Earth	1.0000	1.0000
Mars	1.5198	1.5231
Jupiter	5.2192	5.2024
Saturn	9.1743	9.5371

order to account for the absence of annual parallax. The universe became an immeasurably larger place in Copernican astronomy.

Although the relative dimensions of the Copernican system were in good agreement with modern values, this was not the case for the absolute distances, which were off by a good order of magnitude. For example, the average distance from the Sun to the Earth in Earth radii was calculated by Copernicus to be 1,142; by comparison, the true value is 23,466. In order to obtain absolute distances, it would be necessary to take measurements using telescopes, micrometers, and quadrants with telescopic sights. When this was done, it was found that the apparent planetary diameters were much smaller than had traditionally been assumed, implying that the planets were at much greater distances. A large amount of effort was devoted to determining the astronomical unit, or Earth-Sun distance. A preliminary and reasonably accurate value was obtained by an expedition in 1672 under the administrative direction of Giovanni Cassini (1625–1712), which involved measuring the parallax of Mars at opposition at two points on the Earth, one in Paris and one in Cayenne, near the equator in South America. A better and more reliable value was derived in the middle of the eighteenth century from data collected at expeditions to observe the transit of Venus across the Sun.

Because the annual stellar parallax of even one star was not detected until the nineteenth century, no specific values were available for the distances to the fixed stars. However, as the precision of telescopic observation increased, the null results of parallax measurements implied fairly large lower bounds on stellar distances. By the middle of the eighteenth century it seemed clear that the stars had to be at distances greater than 1,200 astronomical units since the parallax at that distance is about three seconds of arc, and this was within the scope of telescopic measurement. The trend from Copernicus on was to expand the dimensions of the solar system and to vastly increase the size of the universe of stars.

THE BEGINNINGS OF STELLAR ASTRONOMY

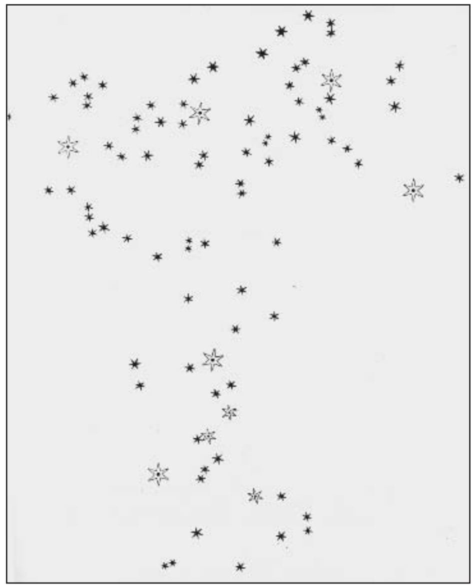

Figure 6.1: A drawing of stars from Galileo's *Starry Messenger* (1610). The Thomas Fisher Rare Book Library, University of Toronto.

The appearance in November of 1572 of a new star in the constellation of Cassiopeia attracted much attention in the scientific circles of Europe. The nova changed in brightness and color, passing from blue to yellow to red as it weakened in brightness, disappearing from view altogether in 1574. Tycho Brahe established that the nova displayed no diurnal parallax and therefore could not be an atmospheric phenomenon; it was definitely a celestial object located somewhere beyond the Moon. He conjectured that the new star might have formed through condensation from thin matter of the Milky Way and even identified a dark spot nearby as its place of origin. The nova of 1572 as well as another one in 1604 contradicted Aristotle's doctrine about the immutability and perfection of the superlunary region of stars, a point emphasized by Galileo in a commentary he wrote on Tycho's book.

In 1610 Galileo published the *Starry Messenger,* an account of his discoveries with the newly invented telescope. Stars were a prominent topic in Galileo's book. He found that there were a vastly larger number of stars than were visible to the naked eye. The Milky Way was shown to consist of stars, the Pleiades were resolved into a group of 36 stars, and the belt of Orion was found to contain countless stars not visible to the naked eye (see figure 6.1). A very important discovery made by Galileo concerned the relative telescopic appearance of the stars and the planets. In ancient astronomy, both the planets and stars were understood to be kindred objects made of the same substance, the fifth element quintessence, or ether. The planets were "wandering stars," and the sphere of the fixed stars was situated just beyond Saturn. Galileo found that the planets viewed through his telescope were rounded disks, while the stars retained the same appearance they possessed in naked-eye observation. The planets seemed to be objects similar to the Moon, the latter having been found to be similar to the Earth, with its mountains and craters. By contrast, the stars remained points of light under telescopic magnification, a fact that implied that they were located at a great distance and consistent with the absence in them of any measurable annual parallax.

From antiquity to the sixteenth century the stars were understood to be fixed, belonging to constellations whose shape and configuration remained unchanged throughout history. Unlike the planets, the stars were subject to no movement with respect to each other. They were regarded as points on a great sphere with the Earth at the center and hence were all at the same distance from the

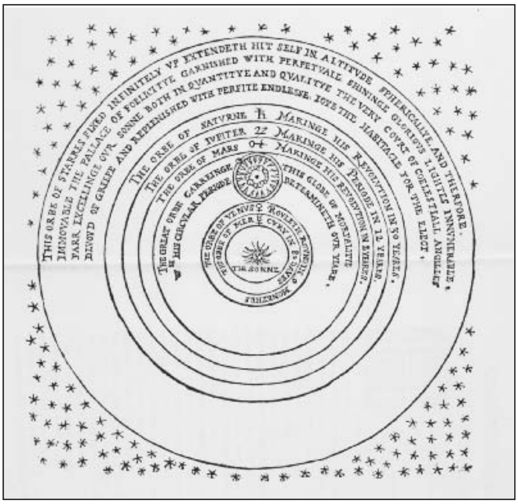

Figure 6.2: Leonard Digges's universe (1576). The Thomas Fisher Rare Book Library, University of Toronto.

Earth. In the century and a half following the publication of Copernicus's *Revolutions,* the stars came to be conceived in a very different way, as luminous objects distributed in space, objects like the Sun, except much more distant. This conception is evident in a book published in 1576 by the English Copernican Leonard Digges (1546–1595), from which figure 6.2 is taken, where the stars are shown to be scattered throughout space beyond the solar system.

In the early eighteenth century the English astronomer Edmund Halley (1656–1742) reasoned that if in fact stars are Sun-like objects, then they should move with respect to the solar system, in the same way that the Sun and the planets move with respect to each other. Over time one should be able to observe changes in the position of the stars on the celestial sphere. Any such motion would be most visible in stars that are closest to us. Halley reasoned that the brightest stars are very likely the closest and therefore were good candidates to exhibit change in position over time. Halley concentrated on the three first-magnitude stars, Arcturus, Sirius, and Aldebaran. Using very accurate values for precession and change in the obliquity in the ecliptic, he was able to show that the coordinates of these stars in ancient star catalogs were significantly different from their coordinates in 1718. These stars exhibit motion with respect to the solar system that is called proper motion. The observed proper motion of a given star is obtained as a combination of its own motion and the motion of the Sun. It is apparent that we will observe only the component of this motion that lies in a direction perpendicular to the line from us to the star.

Among stars visible to the naked eye, the one with the largest proper motion is 61 Cygni in the constellation of Cygnus. The star with the largest proper motion ever observed is known as Barnard's Arrow and was discovered by the American astronomer Edward Barnard (1857–1923) in 1916. It is a faint dwarf star that exhibits an annual motion of a little over 10 seconds of arc per year, at this rate covering in 180 years an arc on the celestial sphere equal to the angular diameter of the Moon. Barnard's Arrow is the second closest star to the Sun, just under six light-years away.

The discovery of proper motion would later become the basis of a method called statistical parallax for determining distances to stars. On average, the size of the proper motion observed in a star will be proportional to its distance from us. In a group of stars the average of their proper motions will provide an estimate of their distance. This method really only became effective in the late nineteenth and twentieth centuries, when data on the proper motions of a large number of stars had been accumulated.

Beginning at the end of the sixteenth century, astronomers identified stars that vary in brightness in a regular pattern. Unlike the novae, these objects were known stars in standard constellations. The star Mira in the constellation of Cetus was shown by David Fabricius (1564–1617) in 1596 to be fading in brightness and was later found to vary in brightness with a period of about 333 days. It was hypothesized that the star possessed a large dark spot and that the change in brightness arose as the star rotated and periodically revealed the spotted and less bright surface to the Earth. In the eighteenth century an important amateur observer who made notable contributions to the study of variable stars was the English deaf-mute John Goodricke (1764–1786). In the 1780s Goodricke showed that the known variable Algol in the constellation of Perseus exhibited an extremely regular and short-period pattern of variation in brightness. He conjectured that a secondary body revolved about Algol and that its changes in brightness occurred as the secondary body passed in front of and around it. It was later confirmed by spectroscopic methods that Algol is indeed an eclipsing binary star; it is now regarded as the prototype for this family of variable stars.

THE TRANSMISSION OF LIGHT: ROEMER AND BRADLEY

During the 1670s the Danish astronomer Olaf Roemer (1644–1710) made repeated observations of the satellites of Jupiter. As the satellites revolve about Jupiter, they pass behind this planet for a time, reappearing a few hours later. The time from one eclipse, or passage into the shadow, to the beginning of the next eclipse gives the period of revolution of each satellite about the primary body. It was hoped that accurate tables giving the times of the eclipses of the satellites could be used as a clock, enabling one through their observation to determine the longitude of points on the Earth far from Europe. Roemer found that the period of revolution of a satellite varies in a systematic way connected to the relative position of the Earth and Jupiter in their orbits about the Sun. During the time when the Earth is moving away from Jupiter the period is longer than it is on average, while it is shorter than average when the Earth is moving toward Jupiter. Roemer realized that this effect could be explained if one assumes that light propagates with finite velocity through space. If the Earth is moving away from Jupiter, the two planets will have separated a certain distance during the time the satellite is in Jupiter's shadow; when it reemerges from the shadow, its light will have to travel this additional distance to reach the Earth, leading to a slightly increased value for its measured period of revolution. Similarly, when the two planets are closing on each other, the period from one eclipse to the next is shorter than it is on average.

During the seventeenth century, there was disagreement over whether light is transmitted instantaneously or with a finite velocity. René Descartes likened the transmission of light from a source to the eye to the transmission of sensation from the end of a cane to the hand holding the cane. He believed that this transmission of sensation occurred instantaneously and reasoned by analogy that the transmission of light was also instantaneous. Roemer's observations of the Jovian satellites not only contradicted this hypothesis but gave a simple way to estimate light's velocity, which Roemer found to be about 200,000 kilometers per second. This conclusion was not immediately accepted by all scientists, and some Cartesian astronomers in Paris maintained with some justification that the effect detected by Roemer was within the margin of error present in the Jovian satellite eclipse data. As a further detailed study of the Jovian satellites was completed, it became clear that Roemer had indeed found something genuine, and the hypothesis of the finite propagation of light won general acceptance.

An important astronomical effect arising from the finite propagation of light was detected by James Bradley (1693–1762) in 1727 and concerned the reception of starlight in the course of Earth's annual orbit about the Sun. Bradley made repeated, very accurate observations of the position of the star δ Draconis in the constellation Draco. He hoped to detect the annual parallax of this star that should exist if the Earth is revolving about the Sun, something everyone accepted by the 1720s. He detected a cyclical shift in the position of the star with an amplitude of 15 seconds of arc and a period equal to one year. After puzzling over this phenomenon for some time he arrived at a simple explanation for it. The star δ Draconis is located very close to the north pole of the plane of the ecliptic, and so its light is falling vertically downward on the plane of the Earth's orbit. As Earth moves in its orbit about the Sun, it is necessary for an observer to tilt the telescope a small amount forward in the direction of the Earth's motion in order to capture the image of the star; the position of the latter is displaced a small amount forward from the vertical. The effect is similar to the one experienced as a person walks on a calm day in the rain; the apparent direction of the vertically falling rain shifts slightly forward from the vertical, and it is necessary to tilt one's umbrella forward.

The phenomenon discovered by Bradley is called stellar aberration. The amount of aberration observed in a star is determined by four quantities: the speed of light, the speed of the Earth in its orbit, the position of the star with respect to the plane of the Earth's orbit, and the position of the Earth in its orbit. After Bradley it was recognized that all stellar observations had to be corrected for aberration. Given the position of the star and the time of year, one applies a formula to determine the amount of aberration or consults a table listing these values.

The finite propagation of light implies that as we look out in space, we are looking backward in time. If a star is 10 light-years away, then we are seeing this star as it was 10 years ago. In cosmology, where we are considering the universe as a whole and are interested in the most distant objects that exist in space, this fact is of the utmost concern. The universe becomes younger the farther we look out. This fact would take on fundamental significance much later, in the twentieth century, when the universe was found to be evolutionary and of finite age.

HERSCHEL AND THE EMERGENCE OF STELLAR ASTRONOMY

When Newton discussed questions pertaining to the whole universe, he considered the solar system and took the center of gravity of this system as his primary point of reference. As historian Max Jammer (1969, 103) has noted, the scope of Newton's cosmological conceptions was quite limited. Astronomy in the eighteenth century was dominated by the same concern with the planetary system that is evident in Newton's writings. The primary focus of observational work was the preparation of accurate planetary tables, and theoretical research was a highly mathematical endeavor devoted to questions of gravitational stability and interaction among the various bodies of the solar system. While it was true that there was some interest in variable stars and the cataloging of nebulae, these activities were very intermittent and never constituted a systematic program of research.

Stellar astronomy as a serious subject may be said to have begun with the German-English astronomer William Herschel (1738–1822) in the last few decades of the eighteenth century. Herschel's career began as a church organist in Bath, where he built telescopes in his spare time and scanned the heavens with them. He constructed a series of large reflecting telescopes, the first at Bath and later ones at Slough near Windsor Castle. The most successful of his instruments was an 18-inch (45 centimeter) reflector of 20 feet (6 meter) focal length, depicted in figure 6.3. Although Herschel's reputation as an astronomer was cemented by his discovery of the planet Uranus in 1781, the greater part of his scientific efforts were devoted to stellar astronomy. Reflecting telescopes with their large light-gathering power were well suited to the observation of faint objects of interest in stellar astronomy. In his detailed surveys of the heavens Herschel has been likened to a "celestial naturalist." He often used imagery from natural history

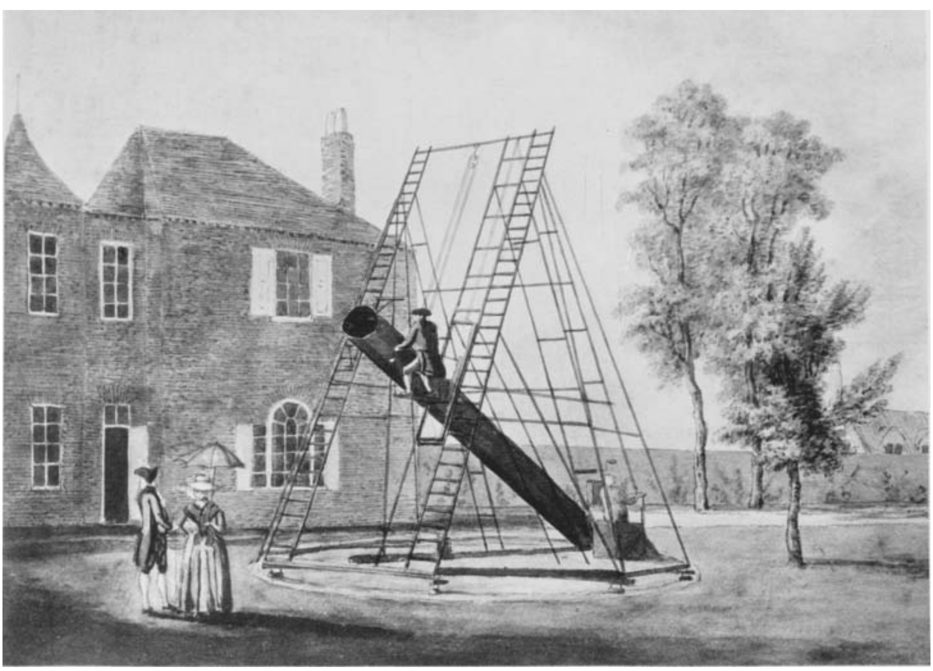

Figure 6.3: Herschel's 20-foot reflector, completed in 1783. The Thomas Fisher Rare Book Library, University of Toronto.

to describe his work, writing for example that "[the heavens] are seen to resemble a luxuriant garden, which contains the greatest variety of productions, in different flourishing beds" (Crowe 1994, 123–24). Herschel was joined by his sister Caroline (1750–1848) in 1773, who proved to be an indefatigable assistant and today is regarded by historians as having been an important observer in her own right.

Telescopic observation had revealed a substantial number of multiple stars, stars that appear as a single object to the naked eye but are resolved into two, three, or more stars by the telescope. It was believed that a typical double star consisted of two unrelated stars at very different distances from us, aligned by chance in the same direction as seen by an observer on Earth. Such doubles were a subject of interest because they seemed to offer an effective way to measure stellar parallax. Observation of the double over the course of the year should reveal some shift in the position of the nearer member of the double with respect to the farther member. Observations of a double star would allow for the measurement of extremely small shifts of position of one member with respect to the other. The use of double stars to measure stellar distances motivated Herschel to compile catalogs of these objects, two of which were produced by him in 1782 and 1784.

Attempts at detecting parallax in double stars failed. Quite apart from the problem of measuring small shifts in position, it was realized that there was a fundamental flaw in this method of detecting parallax. In 1767 John Michell (ca. 1742–1793) reasoned on the basis of statistical arguments that the number of multiple stars in the sky was far too large to be accounted for by chance alignments of stars at different distances. In the case of the Pleiades he argued that the odds were 500,000 to 1 that the six stars just happened to lie on the same line of sight from the observer. These arguments were developed in more detail in 1789 by Herschel, who asserted that most double stars and globular clusters consisted of physically connected systems, composed of stars at approximately the same distance from us. When he returned some years later to reexamine the double stars in his catalogs, he found that certain of them had shifted slightly in position, reflecting the motion over the years of one member with respect to the other. He concluded that doubles and clusters were gravitationally interacting systems of bodies analogous to the solar system.

During his career Herschel increased the number of nebulae known to exist from the 101 of Charles Messier's (1730–1817) catalog to more than 2,500. His views on the nature of the nebulae changed during his career. In the 1780s he believed that the large majority of nebulae were resolvable into a multitude of individual stars, their milky or nebulous appearance being a result of the blending of light from many very distant sources. He developed a cosmology based on the idea that the universe originally consisted of stars scattered throughout space and that these stars had come together to form the clusters and the areas of increased stellar density observed by astronomers. Later discoveries led him to reject this theory. In 1790 he identified a nebula in Taurus consisting of a central star surrounded by a small nebulous region and concluded that it was

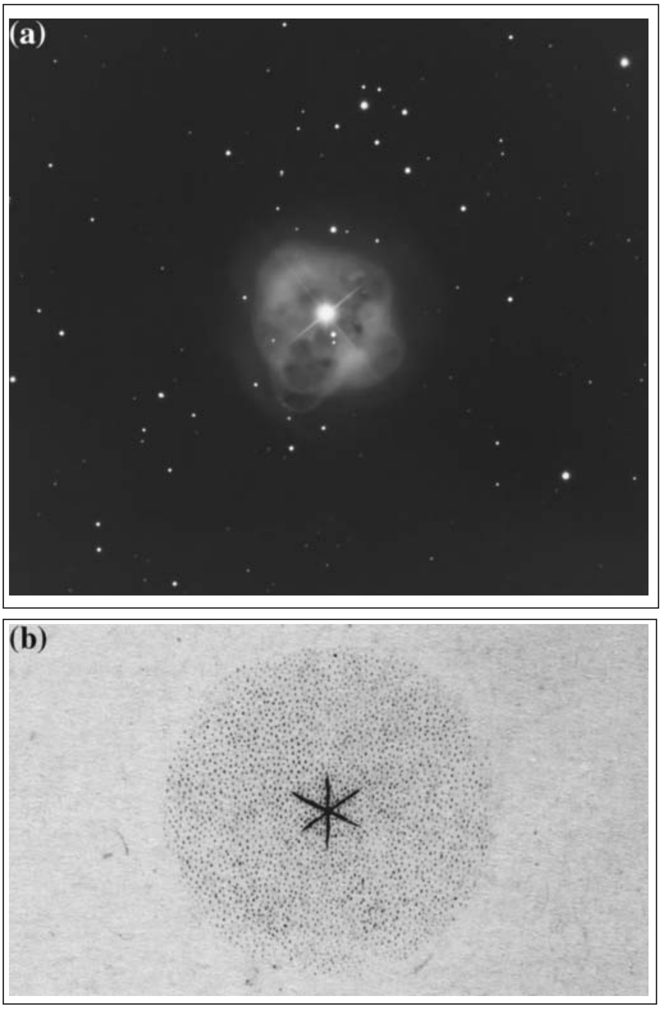

Figure 6.4: NGC 1514: (a) Modern photograph and (b) Herschel's original sketch. Credit: Adam Block/NOAO/AURA/NSF.

an instance of "true nebulosity." This object is today designated NGC 1514, or the 1514th object in the *New General Catalogue* published by John Dreyer (1852–1926) in 1888. (Dreyer's catalog was based in part on the observations of Herschel.) Figure 6.4 shows this nebula as it was drawn by Herschel in 1814 and as it appears in a modern telescope. NGC 1514 belongs to a class of objects Herschel called "planetary nebulae" because of their general resemblance to the disk of a planet. The existence of this class of objects showed that many nebulae were not composed of an aggregate of stars and called into question his theory about "the construction of the heavens" (Crowe 1994, 113). Undeterred, Herschel would take the planetary nebulae and develop a theory of stellar evolution in which they occupied a central place. He hypothesized that NGC 1514 consisted of a star in the process of formation; one was observing nebular matter as it condensed to form a central star. He envisaged a succession of stages in which the nebular matter slowly dissipated as the central star intensified in magnitude. In 1814 he published an article presenting a selection of planetary nebulae from his catalog of nebulae that he believed represented the different stages in a star's development.

In 1783 Herschel calculated the proper motions of seven stars and concluded that these motions were only apparent and resulted from the solar system's motion with respect to the stars. The solar apex is the direction on the celestial sphere of the Sun's motion with respect to the stars in its neighborhood. Herschel located this apex at a point in the constellation Hercules very close to its true position. He also completed several star gauges, or counts of the number of stars in different regions of the sky, and verified the disk theory of the Milky Way system.

Although Herschel's cosmological hypotheses and theories about stellar evolution were speculative and turned out to be false, his contributions to stellar astronomy nonetheless established his reputation as a major scientific innovator. His catalogs of double stars and nebulae and his sky surveys created a systematic basis for further investigation and essentially created stellar astronomy as a field of investigation. He pioneered the use of statistical methods and recognized double and multiple stars as physical systems bound together by gravitation. In his attempts to determine stellar distances, the two-dimensional vault of the celestial sphere was expanded to the third dimension, opening up for study a world of objects distributed in distant space. Herschel's son, John Herschel (1792–1871), would extend his father's observational program to the southern skies and enjoy a career as a leading mathematical scientist in the first half of the nineteenth century.

Herschel's adoption of the reflecting telescope as the instrument of choice for stellar astronomy was ahead of its time. A century following his death, the reflector would achieve dominance in stellar astronomy, its superior light-gathering capacity being crucial in the investigation of objects that were increasingly more distant and faint. In his own time the choice of reflector seemed to confirm Herschel's status as an amateur original, an innovator, and even a genius, but someone who worked at the periphery of professional science.

STELLAR DISTANCES

The first attempts to estimate distances to stars proceeded on the assumption that the stars were roughly the same brightness as the Sun so that their distance could be determined by comparing their brightness with the brightness of the Sun. The task of determining the relative brightness of a star and the Sun was not completely straightforward, and the assumption that all stars possessed the same brightness as the Sun was later shown to be false. Hence the estimates that were obtained were quite crude, although they produced values that formed the basis for the first stage in charting the dimensions of the stellar universe.

Accurate distances would only be achieved with the measurement of trigonometric annual parallax. In the first few decades of the nineteenth century, advances in the grinding of lenses and the mounting of the telescope tubes permitted much greater accuracy in the determination of stellar positions. The refracting telescope was better suited to such work than the reflector and throughout the century was the instrument of choice for research at the leading observatories of Europe and America. A special type of refractor, called a heliometer, was built by the Munich instrument maker Joseph Fraunhofer (1787–1826). This instrument was used to measure the diameter of the Sun—hence its name—and could also be used to find the diameters of planets as well as the angular separation between stars. The objective lens of the refracting telescope was split in half, and the two semicircular halves were allowed to move with respect to each other. Two neighboring stars were sighted in the telescope, the first in one half of the lens and the second in the other half. By measuring the amount the two halves had to be moved in order to make the two images coincide, one obtained a very accurate value for the angular separation between the stars. The Konigsberg astronomer Friedrich Bessel (1784–1846) selected the star 61 Cygni as a candidate to measure parallax because its large proper motion indicated that it must be fairly close to the Sun. Working at his observatory with a 6.25-inch-aperture Fraunhofer heliometer, Bessel observed 61 Cygni over an 18-month period and documented small shifts in its position relative to two nearby faint and more distant stars. In 1838 he obtained a value for its parallax of approximately 0.314 seconds of arc, indicating that the star was at a distance of 10.3 light-years. (Later observation would increase this value to 11.2 light-years.) Subsequently, Thomas Henderson (1798–1844) of the Cape Observatory in South Africa obtained a parallax for α Centauri that was twice as large as 61 Cygni's, indicating that it was twice as close to the Sun.

In 1834 small variations in the proper motion of the dog star Sirius led Bessel to conclude that it had a companion star revolving about it. A similar conclusion followed from close observation of Procyon, Sirius's neighbor in the Little Dog. In 1862 Alvan Graham Clark (1832–1897) visually sighted Sirius's companion using an 18-inch refractor, and astronomers at the end of the century at the Lick Observatory sighted Procyon's companion. It was possible to calculate the masses of the companions, and these were found to be in the range of 10 percent of the mass of the primary star. It was clear that while the brightness of two stars may differ by a factor of thousands, their masses need differ only by only a factor of two or three.

The method of trigonometric parallax only gave distances for stars that were relatively close to the Sun, out to a distance of about 50 light-years. By the time photographic astrometry was introduced in the 1880s the parallax of 90 stars had been determined. Although limited in scope, the trigonometric method provides the essential base line for all further estimations of distance, enabling one to bootstrap from nearby stars to groups of stars more distant, whose distances are obtained by some other method. Distance methods involving variable stars that were developed in the twentieth century would prove to be crucial in the modern revolution in cosmology and are described in more detail in chapter 7.

THE NEW ASTRONOMY

During the nineteenth century, physics expanded beyond the traditional domain of mechanics to include the mathematical theory of optics, heat flow, and electromagnetism. Physics came to encompass a range of phenomena—infrared rays, the polarization of light, electromagnetic induction, and others—which had not even been known to exist in earlier science. In France in the first part of the century, and later, in Britain, Germany, and Italy, physicists devised sophisticated theories based on new physical concepts and powerful mathematical methods. Physics divided into experimental and theoretical branches and was increasingly applied in engineering and technology. Thermodynamics aided the analysis of heat engines, and the theory of electromagnetism was applied to the generation of energy and the transmission of waves. Physics also developed connections with chemistry. In electrolysis, chemical compounds were decomposed by an electrical current, while in spectral analysis the light emitted by burning elements was analyzed by optical methods.

The rapid growth of physics strongly influenced work in astronomy, particularly in the second half of the century. The term "the new astronomy" (self-consciously recalling Kepler's book of 1604) came to designate the study of the heavens by methods and theories of contemporary physics. Writers on astronomy understood themselves to be living in a new stage in the history of the subject, one in which the physical processes of the whole universe were opened up to detailed investigation. Astrophysics as a scientific discipline was well established by the end of the century.

Beginning in 1814 Frauhofer observed and cataloged a large number of dark lines in the spectrum of light from the Sun. In 1849 Léon Foucault (1819–1868) discovered that emission lines seen in the spectrum of a carbon arc in his laboratory appeared as absorption lines when sunlight was passed through a diffraction grating. Eleven years later, Gustav Robert Kirchoff (1824–1887) found that the solar D-lines seen in absorption coincided with the bright lines in the laboratory spectrum of sodium. The origin of the solar absorption lines was hypothesized by him to arise from a cooler layer of gas surrounding the Sun and containing sodium, through which the Sun's rays passed. Kirchoff and his fellow researcher Robert Bunsen (1811–1899) recorded the emission spectral lines of several common elements and, by matching these with absorption lines in the solar spectrum, were able to carry out the first chemical analysis of a celestial body.

Solar spectroscopy would lead to the discovery of a new element unknown on the Earth. Solar prominences are gaseous eruptions from the Sun first observed in the eighteenth century during total eclipses of the Sun. In the eclipse of 1868 Norman Lockyer (1836–1920) and Jules Janssen (1824–1907) identified a new spectral line in a prominence, which Lockyer called D3 because its wavelength was very close to that of sodium (D). Lockyer attributed this line to a new element, which he named helium (from the Greek *helios,* for "Sun.") Some 30 years later, helium gas was produced in the laboratory by burning a mineral, and much later yet, helium would be shown to play a crucial role in the energy processes in the centers of stars.

In the 1860s, two pioneers in the application of spectroscopy to the analysis of starlight were William Huggins (1824–1910) in England and Father Angelo Secchi (1818–1878) in Italy. Huggins attached a diffraction grating to his 15-inch refractor and fitted his observatory with chemical equipment for the burning of compounds and the recording of spectra. Although observation of stellar spectra was difficult, Huggins was able to show that different elements showed up in different stars and that the same elements found in stars were present on Earth. One of his most notable discoveries was the detection of emission lines in a planetary nebula in Draco, indicating that the nebula was gaseous in character and not an unresolved conglomeration of stars. Secchi pioneered the classification of stellar spectra, identifying four main types, and produced a catalog of over 4,000 stellar spectra.

The invention of photography in the 1830s would result in major advances in astronomy. Although in the first stage of its history, astronomical photography was primarily the preserve of amateurs and a few isolated professionals, these pioneers were important innovators in the development of the new technology. For example, much of the terminology of photography, including the word itself, was introduced by the astronomer John Herschel. Technical improvements such as the invention of dry plates and increases in photosensitivity brought photography by the 1880s into the scientific mainstream. Faint images invisible by ordinary optical methods could be captured by long photographic exposures. In 1887 the French Academy of Sciences initiated a large international project, the *Carte du Ciel,* to produce a comprehensive photographic map of the heavens. Photography transformed astrometry, the study of the positions of stars, and photometry, the study of the brightness of stars. The spectra of stars were also photographed and analyzed by Sechhi, who used spectral-line characteristics to classify stars according to their color and surface temperature. At Harvard College Observatory the modern classification system (O-B-A-F-G-K-M-R-N-S, from blue to red, hotter to cooler) was developed to classify stellar spectra, a photographic project that would culminate in the publication in 1924 of over 200,000 stellar spectra. The increasing use of the reflecting telescope coincided with the spread of photography, permitting ever more faint and distant objects to be observed and analyzed.

CONCLUSION

At the beginning of the Copernican Revolution the universe consisted of the Sun and the planets revolving around it. The celestial sphere remained largely as it was in ancient geocentric astronomy, except much more distant, a vault of stars forming a spherical surface far beyond Saturn. By the end of the nineteenth century the Sun had become only one of a countless number of stars in space, and the universe beyond the solar system was the seat of physical and chemical processes that were analyzable in terms of the theory and concepts of the terrestrial laboratory. William Herschel had shown that most double stars are physical systems interacting by gravitation, and spectroscopy had revealed the chemical constitution of celestial bodies. As Huggins wrote of his spectroscopic investigations, the question of "whether the same chemical elements as those of our earth are present throughout the universe was most satisfactorily settled in the affirmative; a common chemistry, it was shown, exists throughout the universe" (1909, 7). Investigations in the twentieth century would produce further evidence for the chemical uniformity of the stars and the uniformity of chemical processes throughout the universe.

The dimensions of the stellar neighborhood of the Sun were reasonably well known by 1900, although large questions remained about the distances to more far-flung stars and to such objects as globular clusters and spiral nebulae. The Sun was known to be a part of the flat disk of stars making up the Milky Way, although even such a basic question as the position of the Sun with respect to the center of the disk was unresolved. It was widely believed that the Milky Way and environs probably constituted the whole universe, although there was by no means any scientific consensus on this point. An essential element of modern cosmology is a well-developed theory of stellar evolution. This was completely lacking in the nineteenth century and would only become possible with the advent of nuclear physics and a satisfactory theory of energy processes within stars.

7

A UNIVERSE OF GALAXIES: THE TRIUMPH OF THE ISLAND-UNIVERSE THEORY

INTRODUCTION

Observational cosmology is the study of the universe as a whole based on observations of stars and nebulae using all of the instrumental resources available to astronomy. Around 1900 these resources consisted of refracting and reflecting telescopes, photography, spectroscopy, and instruments for the measurement of the brightness, spectral distribution, and position of celestial objects. Observational cosmology may be contrasted with theoretical cosmology, which considers various physical and mathematical issues connected with the idea of the universe as a whole. A typical theoretical exercise is to construct a geometric model to describe the universe beginning with a few principles and assumptions. In the present chapter we examine the development of observational cosmology up to the mid-1920s. In the next chapter we continue this examination and consider as well some of the theoretical work carried out during the period.

THE ISLAND-UNIVERSE HYPOTHESIS

With the advent of telescopic astronomy in the seventeenth century, observers began to detect many small cloudy or fuzzy objects in the sky called nebulae, so named from the Latin word *nebula*, for "cloud." As we saw in the last chapter, successively more detailed catalogs of the stars and nebulae were made, culminating around 1800 in the work of William Herschel with his large reflecting telescopes. In 1755 Immanuel Kant (1724–1804) published *Universal Natural History and the Theory of the Heavens.* Following a suggestion of his English contemporary Thomas Wright (1711–1786), Kant speculated that the Sun and the other stars in the sky make up a connected system bound by gravity, in much the same way (except on a much larger scale) as the Sun and the planets form a system. It was known that the band of the Milky Way consisted of a myriad of stars, with the number of stars in the sky increasing as one approached the band of the Milky Way and decreasing as one moved away from it. According to Kant, the Sun and the other stars formed a flat, thin disk lying in the plane defined by the great circle of the Milky Way. As we look out from the solar system within the disk, we observe many stars in the plane of the disk, while progressively fewer stars are seen as we direct our gaze in a direction perpendicular to the plane. Kant further suggested that the white nebulae were themselves conglomerations of stars similar to the Milky Way and situated throughout very distant space. He suggested the observed oval shape of many of the nebulae was a perspective effect resulting from their disk structure. The theory that the nebulae populate space as so many island systems of stars, our Milky Way system being just one instance, would become known as the island-universe theory. Although Kant's writings were only speculative, he provided the first clear statement of the island-universe theory, and the idea of extragalactic nebulae was implanted in the minds of astronomers.

Kant's book contained many other ideas and speculations. He believed that it was probable that the other planets of the solar system were populated by intelligent beings and gave detailed descriptions of their personalities and temperaments. These fantastic suggestions were typical of Enlightenment speculation and were influenced by exotic facts about traditional societies revealed by European voyages of exploration. More important for the subsequent development of science, Kant outlined a theory of the origin of the solar system. The solar system was supposed to have begun as a nebulous mass of swirling gas and dust. As the mass contracted under gravitational attraction and its rotational speed increased, a dense central object formed, while a series of smaller objects were cast off from the center. The center coalesced into the Sun, and the cast-off bodies became the planets. The theory explained why the planets all revolve around the Sun in the same direction and why their orbits all lie in a thin plane with the Sun as center. A conception similar to Kant's was formulated by the great French mathematician and physicist Simon Laplace at the end of the century. This explanation of the solar system became known as the nebular hypothesis and was important historically for introducing ideas of evolution and development in time into scientific thinking about the natural world.

Johann Lambert (1728–1777) was a Swiss mathematician, who, in 1761, published a theory of the Milky Way similar to Kant's. Lambert conceived of the galaxy as a collection of smaller star systems that physically interacted through the action of gravity. The galaxy in turn was only one of a much larger collection of island universes. In his initial investigations of nebulae William Herschel subscribed to the island-universe hypothesis, believing that since many of the nebulae were resolvable into stars, this would be true of all of them. However, his discovery of the planetary nebula in Taurus (chapter 6, figure 6.4) led to an important change in his thinking. It as well as other similar nebula continued to appear diffuse as they were examined under larger telescopes. Herschel openly questioned the hypothesis of island universes and instead tried to explain the planetary nebulae as the first stage in the formation of new stars.

The construction of ever larger telescopes in the nineteenth century brought new information about the detailed structure of nebulae. In 1850 the Irish astronomer and aristocrat Lord Rosse (1800–1867) discerned with his large reflecting telescope, the Leviathan of Parsonstown, that many of the white nebulae possessed a definite spiral structure. One example is an object close to the Big Dipper, numbered 51 in the Messier catalog (figure 7.1). This object is observed from the Earth face on, and M 51 is sometimes called the Whirlpool nebula. Figures 7.1(a) and 7.1(b) provide a comparison of Rosse's original sketch and a modern photograph of M 51. Another very bright nebula that is observed more obliquely is M 31 in the constellation of Andromeda, often referred to as the Great Andromeda nebula. It may be sighted easily with the naked eye in the autumn from locations in the northern hemisphere. It turned out that the class of nebulae possessing an oval or spiral shape was very large indeed, involving myriad objects distributed throughout the sky in regions away from the band of the Milky Way.

Although the island-universe hypothesis continued to attract occasional adherents, it largely lost favor among astronomers as the nineteenth century

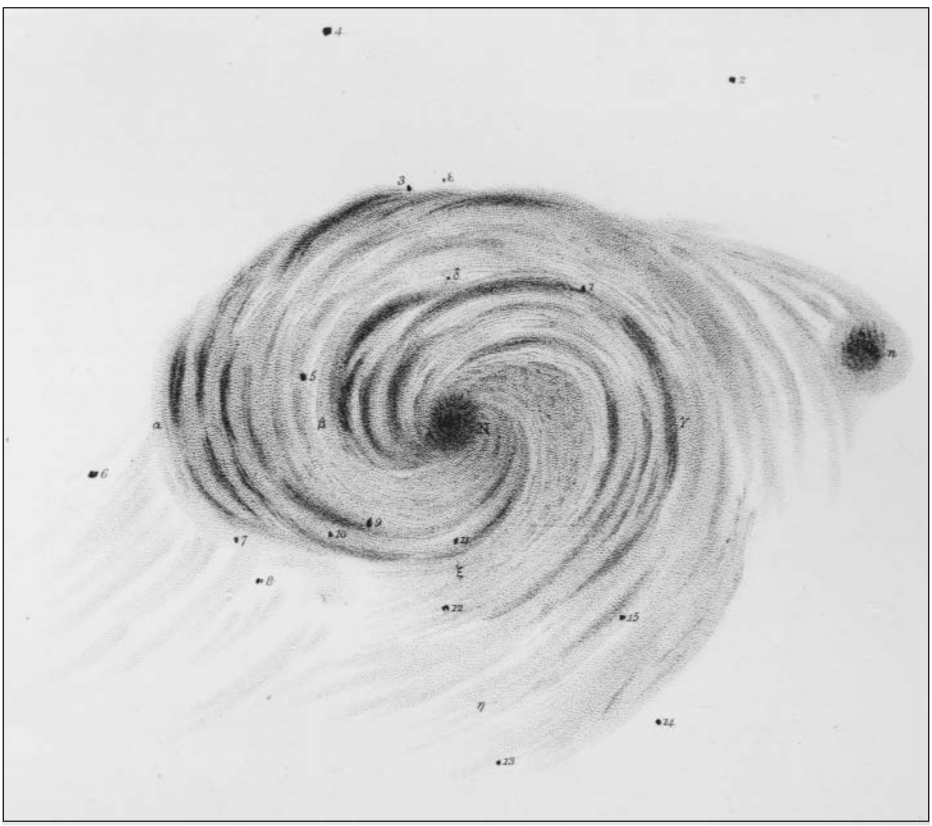

Figure 7.1a: Messier nebula M 51, the Whirlpool nebula. Lord Rosse's sketch. The Thomas Fisher Rare Book Library, University of Toronto.

Figure 7.1b: Messier nebula M 51, the Whirlpool nebula. Modern photo. The Thomas Fisher Rare Book Library, University of Toronto.

came to a close. Several pieces of evidence counted against the hypothesis. First, as was noted above, the nebulae are not distributed randomly in the sky but congregate in regions away from the band of the Milky Way. This "zone of avoidance" seemed to indicate that the nebulae were systemically connected to the Milky Way galaxy and were not independent objects distributed in distant space. In 1885 a new star or nova in the Andromeda nebula for a short period of time outshone in brightness the entire nebula. It was later realized that this star was a supernova, an incredibly energetic and short-lived event in which a massive star explodes. At the time, astronomers reasoned that the brightness of the Andromeda nova meant that it must be nearby, celestially speaking, certainly within the vicinity of the Milky Way system. A final piece of evidence against the island-universe hypothesis emerged with the invention of stellar spectroscopy and the discovery that several of the nebulae classified by Herschel as "planetary" showed only emission lines, indicating that they were composed primarily of gas. This finding confirmed Herschel's own conclusion about such objects. It was later determined that planetary nebulae are of a special sort, being in fact the gaseous remnants of exploded stars within the galaxy. However, it was not apparent at the time that the nebulae were of such radically different types, and the existence of emission lines in some of them was regarded as evidence that the nebulae as a class were not distant objects composed of an immense number of stars.

In 1887 Agnes Clerke (1842–1907), an influential American writer on astronomy, published *A Popular History of Astronomy during the Nineteenth Century*, in which she confidently rejected the island-universe hypothesis: "There is no maintaining nebulae to be simply remote systems of stars … it becomes impossible to resist the conclusion that both nebular and stellar systems are parts of a single scheme" (Crowe 1994, 196). This conclusion was reiterated in 1907 by the German astronomer Max Wolf (1863–1932), who wrote that "the nebulae and clusters of stars represent an essential part of our star-island and perhaps lie relatively close to us. They all form, together with the stars of the Milky Way, an organic whole. Distant, isolated Milky Ways have never been sighted by man" (Crowe 1994, 197).

HENRIETTA SWAN LEAVITT AND CEPHEID VARIABLES

In the period 1880–1910 a group of assistants at the Harvard Observatory, under the direction of Edward Pickering (1846–1919), carried out an extensive program of photometry, involving the measurement of the brightness of stars from the images they produced on photographic plates. Included in this survey were stars in the Magellanic Clouds. The latter are visible as two patches of light in the southern sky and are resolved by telescope into two systems of stars. Pickering's assistants, most of whom were women, measured the brightness of Magellanic stars recorded on photographic plates exposed at regular intervals using the 24-inch refractor at the Harvard field station in Peru. A distinct class of these stars consisted of what were known as Cepheid variable stars, so named after the prototype δ Cephei in the northern constellation of Cepheus. Such stars vary in brightness by about one-half to two magnitudes, with a period of anywhere from 1 to 15 days. The part of the cycle where the star is declining in brightness is longer than the part where it is increasing in brightness. Henrietta Swann Leavitt (1868–1921), a supervisor of the Harvard assistants, noticed that there was a relationship between the brightness of a Cepheid-type star in the Small Magellanic Cloud and its period of variability—the brighter the star, the longer its period of variation. The publication by Pickering of Leavitt's result in the Harvard Observatory circular in 1912 ranks as one of the signal discoveries in the history of observational astronomy.

The stars in the Small Magellanic Cloud belong to a localized group that is quite distant from Earth. Viewed from Earth, the relative differences in their distances are very minor so that the apparent brightness of a given star in the cloud compared to other stars in the cloud is also an indication of its real brightness compared to the real brightness of the other stars. It follows that the relationship between period and observed brightness detected by Leavitt is also a relationship between period and absolute or intrinsic luminosity. This fact was explicitly stated in the 1912 paper of Pickering and Leavitt.

The work of American astronomers was closely followed by the Danish astrophysicist Enjar Hertzsprung (1873–1967). Hertzsprung was very interested in how the different characteristics of a star were related to its luminosity. He found that spectral type, a measure of the surface temperature of a star, was related to luminosity, the total energy emitted by a star. As the spectral type of a star moves from red to blue, there is an increase in its luminosity; there is also a special exceptional group of very bright, red stars. When these facts are depicted graphically, one obtains what is known as the Hertszprung-Russell diagram, named for Hertszprung and the American astronomer Henry Russell (1877–1957), who independently arrived at the relation a few years after his Danish colleague. The H-R diagram would become a key to the study of the evolution of stars and is a basic theoretical element in modern stellar physics.

Hertzsprung knew that information about the intrinsic luminosity of a star could be used to determine its distance and immediately recognized the value for this purpose of Leavitt's result. (Leavitt and Pickering also recognized this fact, but as observationalists, refrained from following up on its theoretical implications.) In principle, the period-luminosity relation could be used to determine the distance to Cepheid variables located anywhere in the universe. From observations over time of a given Cepheid variable star one could measure its period, and by means of the period-luminosity relationship one then knew the star's absolute luminosity. Knowing this quantity, and knowing the observed brightness of the star, one could compute its distance. Cepheid variables therefore provided a yardstick to measure stellar distances. It was necessary to calibrate this yardstick, which meant knowing the distance to at least one Cepheid variable. In addition, it was later revealed that there are different types of Cepheid variables, and it was necessary to ensure that the variable in question belonged to the class for which the yardstick was calibrated. Despite the sizeable uncertainties introduced by these considerations, the discovery of the Cepheid distance method was a major leap forward in the scientific project of mapping the universe.

THE GREAT DEBATE

Observational discoveries in the first two decades of the twentieth century contributed to a revival of interest in the island-universe hypothesis. Spectral analysis of the Andromeda nebula revealed the presence of absorption lines similar to those found in the Sun, suggesting that it was composed of stars. Examination of the spectra of other spirals also turned up dark lines. The Dutch astronomer Cornelius Easton's (1864–1929) "A New Theory of the Milky Way" of 1900 contained speculation about the spiral structure of the Milky Way. Such structure would indicate an obvious family similarity between our galaxy and the spiral nebulae that were turning up in increasing numbers in the big telescopes. The British astronomer David Gill (1843–1914), in 1911, compared the Milky Way to such spirals as M 51 and became an influential spokesman for the island-universe theory.

Heber Curtis (1872–1942) was an American classicist-turned-astronomer, who carried out stellar research at the California Lick Observatory in the mountains east of San Francisco. In studying photographs of spiral nebulae he noticed that many of the ones that were observed edge on possessed an equatorial band of obscuring matter. If the Milky Way were such a system and possessed this obscuring matter, then this fact would explain why no distant nebulae were observed in the band of the galaxy. The zone of avoidance would then be a simple optical effect and would not imply any physical connection between the nebulae and the galaxy. Curtis's study of novae in photographic plates of spiral nebulae seemed to indicate that they were similar in kind to the novae in the Milky Way. Their faintness and apparent similarity to galactic novae would mean that the spirals were at a very great distance. The 1885 nova in Andromeda was interpreted as an exceptionally bright event not typical of the class of novae as a whole. The discovery by Vesto Slipher (1875–1969) in 1916 of very large radial velocities of spiral nebulae (discussed in more detail in the next chapter) indicated to some astronomers that these objects were unlike anything within the Milky Way galaxy and must be extragalactic in nature.

Harlow Shapley (1885–1972) was a prominent American astronomer, who became a strong opponent of the island-universe hypothesis. He developed a method for determining distances based on statistical parallax, on the use of the Hertzsprung-Russell diagram to determine luminosity, and on the luminosity-period relation for Cepheid variables. An important type of celestial object is a globular cluster, a compact conglomeration of thousands of stars. One such cluster, Messier 13 in Hercules, is visible to the naked eye in the summer sky from the northern hemisphere. Shapley calculated distances to the globular clusters and found them to be tenfold larger than estimates based on established theory. He concluded that the Milky Way galaxy was a very large object, as much as 300,000 light-years across. He became a champion of what was called the big-galaxy theory of the universe, according to which the universe was believed to be dominated by our large galaxy, with globular clusters and spiral nebulae being considered part of the galaxy or satellite objects to it. The zone of avoidance was explained as resulting from radiation pressure from regions of increased star density that propelled globular clusters and spiral nebulae away from the equatorial plane of the galaxy.

Although Shapley's theory of the galaxy was later shown to be false, he did make one fundamental and enduring discovery. He reasoned on various grounds that the galaxy is framed symmetrically by globular clusters. Since most of these clusters are in fact observed on one side of the sky, it follows that the Earth is not at the center of the Milky Way system but is located a considerable distance away from this center. This result, which turned out to be correct, was a striking confirmation of the "Copernican" principle that the Earth occupies no special place in the universe.

At a meeting of the American Astronomical Society in April of 1920 Shapley and Curtis engaged in a debate on the island-universe hypothesis and the question of the status of the nebulae. Curtis staked out the claim that the white nebulae were autonomous, distant galaxies of stars, much like our own galaxy. In opposing this view Shapley appealed to his big-galaxy conception. If the Milky

Way system possessed the very large dimensions ascribed to it by this theory, and if in fact the spiral nebulae were comparable galaxies to the Milky Way, then it would follow that these nebulae were at distances of a magnitude hitherto unimagined in astronomy. The novae that were sometimes observed in spiral nebulae would possess a level of luminosity that seemed to defy the laws of physics.

Shapley also appealed to work of the Mount Wilson astronomer Adriaan van Maanen (1884–1946). The latter had investigated spiral nebulae that we view face on. Prominent examples are Messier 51 and Messier 101, both in the constellation of Ursa Major. These were among the very first spiral nebulae identified by Lord Rosse. Van Maanen took photographs of M 101 over a period of time and superimposed the resulting images. He was an expert in precision photographic measurement and believed that he had detected a rotational motion in the spiral arms, a result that would certainly only be possible if these objects were relatively close and relatively small. A similar motion was found to be present in other spiral nebulae. It would later be shown that the motions that van Maanen identified were illusory, a mistake in observation that arose because he was working at the very limits of measurement. Nevertheless, he was a prominent astronomer and a friend of Shapley's, and his conclusions were cited at the time as evidence against the island-universe theory.

Curtis responded by challenging the big-galaxy theory, maintaining that Shapley had overestimated the size of the galaxy by a factor as large as 10. In the first part he raised a number of technical objections to the distance indicators used by Shapley and called attention to the bold and unsubstantiated character of the big-galaxy theory. As evidence for the extragalactic character of the spirals, he cited the large radial velocities that seemed to distinguish them from objects in the galaxy. He cited the band of obscuring matter observed in the equatorial regions of the spirals and suggested that such a band in the Milky Way would explain why spiral nebulae were not observed along the galactic equator. This explanation refuted the argument concerning the zone of avoidance and the hypothesis that the spirals were systemically connected to the galaxy. There were also some grounds for believing that the Milky Way system possessed a spiral structure, and this fact, if verified, would be in keeping with the galaxy's status as a system comparable to the spiral nebulae. The spectra of the nebulae indicated that they were composed of stars and were not mainly gaseous, as Shapley seemed to believe. Finally, the novae in the spirals suggested very large distances for them that would imply they were of a size comparable to our galaxy. It is also of note that Curtis challenged Shapley's hypothesis concerning the position of the solar system within the galaxy, holding that the Sun is located very close to its center. Later, astronomy would affirm the correctness of Shapley on this point of the debate.

LARGE TELESCOPES

Conclusive evidence that would settle the great debate was obtained through observations made with a new generation of powerful telescopes constructed on high stations in the American West. Astronomy had entered a stage where advanced technology and large-scale science would dominate the frontiers of research. Although theorists continued to play an important role, they were overshadowed by the continuous stream of stunning findings coming from the mountaintop observatories. The primary place of advanced technology in astronomical research has continued up to the present and appears to have become a permanent feature of the science of astronomy.

During the nineteenth century the refracting telescope was the instrument of choice for astronomical investigation at the professional observatory. Although the refractor provided very good resolution and was excellent for precision measurement, it was subject to several limitations. The passage of light through a large, primary, objective lens resulted in a loss in the blue end of the spectrum. The objective lens needed to be supported around its perimeter, and there was a tendency for it to sag, a fact that limited its size to less than 100 centimeters in diameter. The largest refractor ever to be built was a 40-inch (100 centimeter) telescope installed at the Yerkes Observatory in Wisconsin. A telescope that would play a critical role in the history of modern cosmology was the 24-inch (60 centimeter) refractor at the Lowell Observatory in Flagstaff, Arizona. Perhaps more than anything else, the establishment of the Lowell Observatory indicated a recognition by researchers that high-altitude and dry locations were necessary to secure ideal seeing conditions.

Until the end of the nineteenth century the reflecting telescope was viewed as an instrument for amateurs, limited by difficulties in keeping the primary speculum-metal mirror polished and properly figured. However, reflecting telescopes could be built much larger than refractors since one entire surface of the primary mirror could be supported structurally within the telescope assembly. With advances in casting technology, plate glass replaced speculum metal in the construction of the primary mirror. The invention of improved mountings and the shorter focal length of reflectors also contributed to their competitiveness.

In the period 1900–1920, large reflectors were built on mountaintops in California and several other locations around the globe. The 36-inch (90 centimeter) Crossley reflector was installed in 1898 at Lick Observatory on Mount Hamilton and proved to be very effective for stellar photography. James Keeler (1857–1900) initiated a program to photograph many thousands of spiral nebulae. His successor, William Campbell (1862–1938), became a vigorous advocate of the island-universe theory. Curtis cited observations of nebulae made with the Lick Crossley reflector in his debate with Shapley.

George Ellery Hale (1868–1938) was an American astronomer who arranged to have three powerful solar telescopes built on Mount Wilson overlooking Pasadena, California. Hale was also very active in promoting the construction of large telescopes for stellar and nebular astrophysical work. He recognized the superiority of the reflector for such work and was able to convince philanthropists and scientific foundations to support some very ambitious projects. In 1908 a 60-inch (150 centimeter) reflector was built on Mount Wilson; this

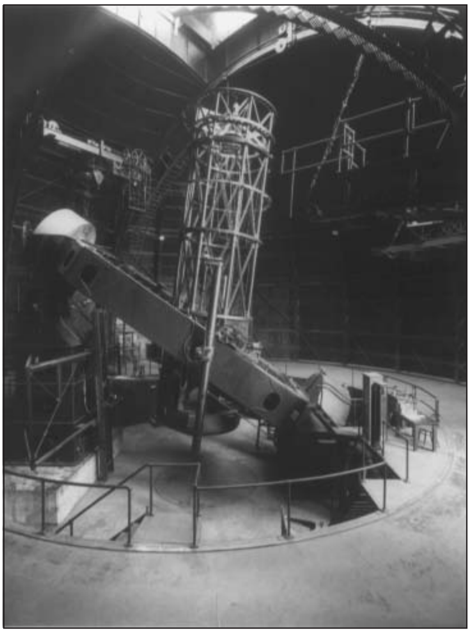

Figure 7.2: The Hooker 100-inch telescope at Mount Wilson. Courtesy of Library of Congress.

was followed in 1917 at the same location by a larger 100-inch (250 centimeter) reflector, the Hooker telescope (figure 7.2), an instrument of unprecedented size and power. Financial support for the Hooker telescope was provided by the philanthropist J. D. Hooker and by the Carnegie Institution. The construction of large reflectors was not limited to California. In 1918 a 74-inch (190 centimeter) reflector was built at the Dominion Astrophysical Laboratory near Victoria, British Columbia, in Canada. The large reflectors were well suited to deep-space observation, in which the objects are very faint, and by the middle of the twentieth century had achieved complete dominance in astrophysical research. This dominance was furthered by the Hamburg astronomer Bernhard Schmidt's (1879–1935) invention in 1931 of a special reflector and camera with a wide field of observation.

HUBBLE AND EXTRAGALACTIC NEBULAE

Many people were impressed by Curtis's arguments in the debate with Shapley, and opponents of the island-universe theory were increasingly put on the defensive. Still, there was no definite winner, and the question of the island-universe hypothesis was an open one in the early 1920s. As an advocate of the big-galaxy model, Shapley continued to question evidence for the hypothesis. Van Maanen's work on the rotation of the spirals also found an influential supporter in the British astrophysicist James Jeans (1877–1946). Although Jeans was skeptical of Shapley's big-galaxy model, he was even more critical of island universes and attempted to describe the dynamical motion of spiral nebulae using van Maanen's measurements. For a time Jeans developed an evolutionary theory of nebulae, suggesting that the spirals, as they evolved, turned into globular clusters. The English astronomer F. H. Reynolds (1874–1949) was another vigorous opponent of the island-universe hypothesis, as was the young Canadian researcher Harry Plaskett (1893–1980).

Support for the island-universe theory was stimulated by growing evidence for the very large distances to the spirals and, in particular, for the distance to the brightest and most prominent spiral, the Andromeda nebula M 31. On the basis of observations of novae in M 31, several authorities concluded that it was much too distant to be part of our galaxy. The Swedish astronomer Knut Lundmark (1889–1958) emerged as a vigorous advocate of island universes, using his study at Mount Wilson of bright stars in the nebula M 33 (a neighbor to M 31 in the constellation Triangulum) to obtain a distance for it of over one million light-years. Lundmark was also dismissive of Shapley's big-galaxy conception, an attitude that resulted in some tension between the two men.

Edwin Hubble (1889–1953) was born into a middle-class American family, the son of a lawyer who worked in the insurance business. Upon completing high school in Chicago he entered the University of Chicago, where he studied mathematics and astronomy. One of his professors was Hale, whose efforts had led, in 1897, to the establishment of the Yerkes Observatory at Williams Bay, Wisconsin. Upon graduation Hubble went to Oxford as a Rhodes scholar, where he studied law, excelling as well as a heavyweight boxer and track-and-field athlete. He returned to the United States in 1913, and after a brief period as a practicing lawyer, entered graduate studies in astronomy at the University of Chicago, where he carried out research at Yerkes Observatory. His doctoral thesis in 1917 was titled "Photographic Investigations of Faint Nebulae."

After service in World War I Hubble, in 1919, was offered a position as staff astronomer at Mount Wilson. In the years that followed he trained the 100-inch reflector on the Andromeda nebula (Messier 31) and was able to resolve a multitude of stars within it (see figure 7.3). Such was the power of this telescope that Hubble was able to identify a group of Cepheid variables within M 31 and accurately measure their periods of variation. In estimating their brightness he made use of photographic magnitude scales that had been prepared by Frederick Seares (1873–1964), a Mount Wilson astronomer whose work in this area had also been essential to Shapley's earlier investigation of Cepheid distance indicators. The stars under study at Mount Wilson were too faint to be included in any of the existing magnitude scales. Seares was an expert on stellar photometric methods and successfully applied them to the 60-inch and 100-inch reflectors. Hubble's Cepheid data immediately provided an indication of the distance of M 31 relative to nearby galactic Cepheid stars. By late 1924 Hubble had established that M 31 was some 285,000 parsecs distant, an object outside our Milky Way and undisputedly an extragalactic nebula. His discovery was reported by Henry Russell at a historic meeting of the American Astronomical Association in early January of 1925.

Figure 7.3: Andromeda nebula M 31. Credit: Bill Schoenign, Vanessa Harvey/REU program/NOAO/AURA/NSF.

With Hubble's result the astronomical community was largely won over to the island-universe hypothesis. The galaxy was named from the Greek word for "milky way." With Hubble's discovery the word galaxy was extended to any of the large star systems external to the Milky Way system on the grounds that these were objects similar to the Milky Way. Throughout his career Hubble himself preferred to retain the theory-neutral term “extragalactic nebula” for such external star systems.

Continued opposition to the island-universe theory was largely based on the evidence supplied by van Maanen for the rotation of spiral nebulae. Van Maanen himself stubbornly continued to defend his measurements and to oppose the island-universe theory well into the 1930s. Hubble was forced to expend considerable effort in exposing the errors of his senior Mount Wilson colleague, well after the time when the reality of external galaxies was accepted by most astronomers. Today, van Maanen is regarded as an example of how an experienced scientist can sometimes allow preconceptions to distort his interpretation of the observational data.

There were two respects in which the conception of external galaxies in the 1920s and 1930s differed from the traditional island-universe theory and from the views that prevail today. The measurement of distances to the spirals implied that they were distinctly smaller than the Milky Way galaxy. This fact provided some support for a modified version of Shapley’s big-galaxy theory, and even those who objected to this theory tended to believe in an anomalously large size for the galaxy. The solution to this riddle would emerge from the work of Robert Trumpler (1886–1956) and his study of the distances to objects within the galaxy known as open star clusters. For comparing such clusters the apparent angular diameter provided an estimate of distance, with smaller clusters being farther away and larger clusters being closer. On the other hand, by examining the spectral characteristics of the stars in the cluster one could use the Hertzsprung-Russell diagram to obtain a measure of the stars’ absolute luminosity, and this datum could be used to calculate the cluster’s distance. When the two measures were compared, it was found that the distance obtained using the H-R diagram was always larger than the distance based on angular size. Trumpler concluded that a significant amount of gas existed in the galaxy and that this led to a decrease in the brightness of star clusters. Because of this, they looked fainter and farther away than they actually were, and the H-R method was overestimating their distance.

The H-R method had been used extensively by earlier researchers. When allowance was taken for interstellar absorption, the dimensions of the galaxy were found to be only a fraction of the size proposed by Shapley. Moreover, in the 1950s it was shown that the distances to the spiral galaxies had been underestimated by at least a factor of two. (We consider this development in the next chapter.) When the revised distance scales were taken into account, it was clear that the galaxy was comparable in size to the spirals, being in fact approximately the same size as the Andromeda nebula.

The second difference concerned the structure of the spirals and the relative balance of stars and nebulous matter to be found in them. Hubble and Jeans believed that the center of a spiral consists of pure nebulosity, while the outer parts of the spiral are composed primarily of stars. Elliptical galaxies were believed to be composed entirely of nebulosity and were thought to be the earliest stage in the development of a spiral galaxy. As the nebula evolved, it expanded outward, resulting in the formation of spiral arms and regions of star formation that grew in size over time. Hubble's views about an evolutionary sequence and the composite structure of spiral nebulae are now believed to be mistaken. However, his scheme for the classification of galaxies using the idea of evolutionary development is maintained today, largely for historical reasons.

One hypothesis that was confirmed during the period through a detailed study of star motions was Shapley's supposition that the Sun occupies a position offset from the center of the galaxy. The statistical studies of Jan Oort (1900–1992) and Bertil Lindblad (1895–1965) in the late 1920s indicated certain systematic patterns of stellar motion implying a differential axial rotation about a point some 5,000 parsecs from the Sun in the direction of the constellation Sagittarius. This center of rotation coincided with the center inferred by Shapley from the conjectured symmetrical distribution of globular clusters around the galaxy. Much later, in the 1950s, Oort would be one of the researchers who identified the spiral structure of the galaxy through a radio astronomical study of interstellar hydrogen.

CONCLUSION

Between 1915 and 1930 a major shift had occurred in the way in which the basic astronomical constituents of the universe were conceptualized. In 1915 the fundamental object of interest to cosmologists was the star, and the aim of cosmology was to elucidate the nature of the sidereal universe. In 1935 the galaxy had replaced the star as the primary unit in the interpretation of the large-scale structure of the universe. A focus on the relationship of the Sun within the Milky Way system was replaced by an interest in the place of the Milky Way among the multitude of galaxies that populate space in every direction in the sky. Questions about the universe as a whole now involved the distribution and relative motion of galaxies, their physical structure and probable evolution.

8

THE EXPANSION OF THE UNIVERSE

INTRODUCTION

By the late 1920s the large-scale structure of the universe had been clarified: distributed throughout distant space are countless white nebulae, galaxies like our own Milky Way galaxy, each consisting of hundreds of millions of stars. The pivotal event in the history of modern cosmology was Hubble's discovery, in 1929, that these nebulae exhibit red shifts that vary in a systematic way with distance, the size of the red shift of a galaxy being proportional to its distance from us. The findings of Hubble and other astronomers in the period from 1912 to 1929 were a result of technological advances in instrumentation and improved observing conditions afforded by mountaintop observatories. In one of history's great coincidences the work of the astronomers occurred at just the same time as Albert Einstein's (1879–1955) seminal mathematical investigations in the general theory of relativity. Einstein's theory laid the groundwork for a systematic mathematical approach to cosmology and led to the formation of a vigorous group of researchers in the new field of relativistic cosmology. These researchers were ready and able to interpret the sensational discoveries in nebular astronomy as they came along. The relativists succeeded in making the concept of an expanding universe central to all modern thinking about the universe.

COSMOLOGY BEFORE EINSTEIN

Discoveries in observational astronomy would provide the basis for the revolution in cosmology that took place after 1900. It should nevertheless be noted that the study of cosmology existed before this time, largely as the study of general questions about the universe as a whole. These investigations reduced, in some cases, to speculative intellectual exercises, but they prepared the way for new mathematical descriptions of the universe and would become important when the revolutionary observational findings of the 1920s were in place. It was out of this speculative, older tradition that analysis of the universe based on Einstein's general theory of relativity emerged, and the latter would eventually provide the theoretical framework for all of modern work in cosmology.

Olbers' Paradox

In the eighteenth century a general picture of the universe emerged in which space is populated by stars, self-luminous objects similar to the Sun. The notion of an extended, possibly infinite distribution of stars in space raised certain puzzles about what we should expect to observe in the sky. In particular, there seemed to be a question about the background level of brightness in the sky that would result from such a distribution of light sources. The fact that the sky is dark at night emerged as something that required explanation. The question was discussed by the Swiss astronomer Philippe Loys de Chéseaux (1718–1751) in 1744 and again independently in 1826 by Heinrich Wilhelm Olbers (1758–1840), a German physician and amateur astronomer. The puzzle is known today as Olber's paradox, a name introduced by the British cosmologist Herman Bondi (1919–) in 1958.

Let us assume that stars are evenly distributed so that their average density in space is constant. Consider an imaginary, thin, spherical shell centered on the Earth. Our goal is to calculate the total intensity of the light reaching the Earth from all the stars contained within this shell. The number of stars will be proportional to the area of the surface of the shell multiplied by the small thickness of the shell. This area is proportional to the square of the radius. On the other hand, the intensity of light from each star will be proportional to one over the square of the radius. It follows that the intensity of light coming from all of the stars in the shell is equal to a constant times the thickness of the shell. As we go deeper into space, we encounter spheres of ever-increasing radius. If all of the starlight reached us, then the total intensity of the starlight from all of the stars in a given sphere would be proportional to its radius. However, some of the light from the more distant stars will be blocked by closer stars. The total amount of starlight from all the stars in a sphere will not increase indefinitely with the radius but will eventually reach a limiting value. Under any reasonable assumptions about the average size and brightness of stars and their density in space it turns out that this value is very high. The sky should be ablaze with light of great intensity. The paradox evident in the darkness of the night sky seems to follow from simple and plausible assumptions about the universe as a whole.

There have been two classes of solution to Olbers' paradox. The first, advanced by Chéseaux and Olbers, involves an explanation in terms of the physical process of light transmission. It is suggested that as light travels through space, some of it is absorbed by intervening matter present in the form of dust, fluid, or gas, leading to a reduction in its intensity. This explanation has been severely criticized on thermodynamical grounds since any energy absorbed by the intervening matter would result in heating and the reemission of radiation. The second explanation involves cosmological assumptions about the distribution of matter throughout the universe. One such argument involves what is called a hierarchic universe. It is supposed that matter is arranged in the form of a sequence of hierarchies so that the overall density of luminous sources of radiation decreases as one moves outward in such a way as to compensate for the increased number of sources. An explanation of this sort was presented by Carl Charlier (1862–1934) early in the twentieth century.

Explanations of the dark night sky using the concept of a hierarchic universe tended to be highly theoretical. Some astronomers suggested solutions that were simpler and supported by what was then known about the universe. Like many, if not most, of his scientific contemporaries, Harlow Shapley in 1915 believed that the whole universe consisted of the Milky Way galaxy as well as possibly a few satellite objects about the galaxy. In this conception the universe is like an island suspended in infinite empty space. Such a universe avoids the puzzle of the night sky since matter is not distributed indefinitely but is circumscribed by the boundaries of the galaxy. Of course, this explanation became unsatisfactory when it was recognized that the galaxy is just one of countless nebular stellar systems scattered throughout distant space.

Gravity Paradox

A puzzle related to Olbers' paradox concerns the question of the gravitational field exerted at a given point by an indefinitely extended distribution of matter in space. Given a system of bodies, the mathematical function that specifies the strength of the gravitational field at each point in space resulting from this system is known as the gravitational potential function of the system. If we let the system of bodies be the whole universe, we are confronted with the problem of how to define a universal potential function. In the late nineteenth century Carl Neumann (1832–1925) pointed out that it was not clear, given standard Newtonian gravitational theory, how one would obtain a mathematical function that is well defined. Even the slightest variation in the density of matter could lead, on a cosmological scale, to singularities in the potential function.

Proposed resolutions of the gravity potential problem followed the same two lines of reasoning used to explain Olbers' paradox: modifying physical processes on the one hand, and on the other, making suitable assumptions about the large-scale distribution of matter in the universe. The solution given by Neumann, and later, by Hugo von Seeliger (1849–1924), was to modify the physical law of gravitation. The gravitational force acting between two bodies separated by a distance r is multiplied by a factor proportional to a quantity of the form $\exp(-r)$. For small values of r the law reduces to the usual Newtonian law, while for larger values of r the force drops off to a value close to zero. The operation of the modified law is equivalent to supposing that over very long distances a repulsive force acts, counteracting the force of gravity and leading to general equilibrium on a large scale.

A different resolution of the potential problem would be to suppose, as Shapley did, that all of the matter of the universe is located within circumscribed boundaries. A more abstract approach is to use the concept of hierarchical ordering, where the density of distant matter is supposed to decrease in such a way as to make negligible the contribution of this matter to the gravitational potential function.

Non-Euclidean Geometry

In the early nineteenth century, mathematicians showed that there are geometries different from the time-honored geometry of Euclid. There is no single absolute and true geometry but many different and mutually inconsistent geometries. Pioneers in this revolutionary field of study were Nikolai Lobachevsky (1793–1856) in Russia and Janos Bolyai (1802–1860) in Hungary. Later in the century, important further work was done by the great German mathematician Georg Riemann (1826–1866), who adapted methods from analysis and calculus to the abstract study of geometrical spaces.

Consider a line and a point not on the line, both lying in a plane. In Euclidean geometry, there is a unique line through the point parallel to the given line. This fact can be shown to be equivalent to the assertion that the sum of the angles of a triangle is equal to two right angles. This property of Euclidean geometry defines its character at the most fundamental level. Lobachevsky and Bolyai considered geometrical systems in which it was possible, through a point not on a line, to draw an infinite number of lines parallel to the given line. This property is equivalent to supposing that the angles of a triangle sum to less than two right angles. A geometrical system with this property is known as hyperbolic geometry because the relations between angles and lengths in it are described using the hyperbolic trigonometric functions. Riemann showed that one could also obtain a geometry different from both Euclidean and hyperbolic geometry by supposing that there are no parallel lines: for any point not on a given line, every line through this point intersects the given line. In the resulting geometry, known as elliptical, or Riemannian, geometry, the angles of a triangle sum to a value greater than two right angles.

It was tacitly assumed throughout history that physical space and the universe itself are described by Euclidean geometry. In the late nineteenth and early twentieth centuries, astronomers questioned this assumption and considered the possibility of spatial relations based on hyperbolic or elliptical geometry. Observational tests were proposed to decide the geometry of our universe and, if this geometry was elliptical or hyperbolic, to determine the radius of curvature of space. Such tests, which would have had fundamental implications for cosmology, were inconclusive, the margin of error being too large to distinguish between Euclidean and non-Euclidean cases. Around 1900 Simon Newcomb (1832–1925) and Karl Schwarzschild (1873–1916) discussed these questions in quantitative terms. Speculation about the mathematician's "fairyland of geometry" and its applicability to the physical world were widespread during the period.

EINSTEIN AND THE THEORY OF RELATIVITY

Special Theory of Relativity

By the middle of the nineteenth century the wave theory had become the accepted explanation for such optical phenomena as the reflection, refraction, and transmission of light. It was believed that space was filled with a universal "luminiferous ether," through which light waves propagated, in the same way that waves from a stone dropped in a pond propagated through water. In the last decades of the century, physicists devised experiments to detect the motion of the Earth as it moved through the ether in its annual revolution about the Sun. The measured velocity of light emitted from a source on the Earth should depend on the direction of the light with respect to the motion of the Earth in the ether. One of the most famous experiments to detect such variations in velocity was designed by two American physicists, Albert Michelson (1852–1931) and Edward Morley (1838–1923). Very small changes were thought to be detectible by means of an optical device known as an interferometer, which recorded the interference produced by two sets of rays moving with slightly different velocities in mutually perpendicular directions. Another method to detect the Earth's motion through the ether was based on the analysis of a phenomenon known as stellar aberration.

Despite the best efforts of experimenters, no evidence was found of the motion of the Earth through the ether. Various attempts were made to account for this failure, including the hypothesis that the length of a measuring rod contracted in the direction of motion in such a way as to exactly counteract the variations in light velocity that should otherwise be observed. Another possibility was that the Earth dragged the ether as it moved through space. In the theory of electrodynamics, there were also puzzles concerning the relative motion of bodies.

Albert Einstein, a young physicist working in a patent office in Switzerland, developed a radical revision of classical Newtonian mechanics that explained the null results of the ether experiments. Einstein's special theory of relativity of 1905 was based on incorporating the observer into the description of a physical system and recognizing that all physical events were witnessed relative to a reference frame. The special theory contains two fundamental postulates:

1. The speed of light in a vacuum is independent of the motion of its source.
2. The laws of physics are the same in all inertial reference frames.

The first postulate was true in traditional ether physics, where light, once it left the source, traveled as a disturbance in the ether at the characteristic speed of c equal to 299,792 kilometers per second. The second postulate was the revolutionary one since it implied that the velocity of light in a vacuum as measured in any inertial reference frame will have one and the same value. In particular, the velocity of light as the Earth moves through space will be found to be exactly the same in all directions, just as was found to be the case in the

Michelson-Morley experiment. The two postulates implied together that such basic concepts as space, time, and mass must be understood in relation to a given inertial reference frame. Events that are simultaneous in one reference frame will not be so as viewed in another frame moving with a nonzero velocity with respect to the first frame. Newton's belief that there were such things as absolute space and time had to be rejected altogether and a new set of equations introduced to transform space, time, and mass between reference frames. Finally, the special theory of relativity showed that however light and other electromagnetic disturbances travel through space, it is not as a result of the simple mechanical motion of a disturbance in a material ether.

General Theory of Relativity

In the years following 1905 Einstein became interested in extending the special theory of relativity to encompass gravitational phenomena, with the eventual goal of developing a comprehensive theory for all physical forces. The force of gravity acting between two bodies is proportional to the product of the masses of the bodies. The mass of a body that appears in this relation is known as the body's gravitational mass. The mass that appears in Newton's second law asserting the proportionality of force to mass times acceleration is the inertial mass and can be regarded as a measure of the body's resistance to change of motion. A series of experiments beginning with Newton had shown that the gravitational and inertial mass of a body were equal. Einstein took this fact and generalized it into something called the principle of equivalence, according to which a system of bodies in a uniform gravitational field may be regarded as equivalent to the same system in which no forces act and in which the system is subjected to a uniformly accelerated motion. The force of gravity is replaced by the acceleration of the given reference frame. The formulation of the principle of equivalence was the first step in the development of what would become known as the general theory of relativity.

Einstein was influenced by the brilliant Goettingen mathematician Hermann Minkowski (1864–1909), who had devised a geometrical interpretation of the special theory of relativity. Einstein's ultimate goal was to describe the effects of gravity in terms of the geometrical structure of space and time. To do this, he used techniques from a branch of geometry known as the absolute differential calculus, a field of research pioneered by Italian mathematicians at the end of the nineteenth century. He was introduced to the subject by his Swiss colleague Marcel Grossmann (1878–1936), who with him wrote several papers on the mathematics of gravitational theory. It was Einstein and Grossmann who gave the now standard name "tensor analysis" to the absolute differential calculus. After considerable effort Einstein finally succeeded in producing tensorial formulations of the field equations of gravitation: the action of gravity acting on a unit mass at any point in space was given in terms of equations containing energy and curvature tensors. The equations connected a physical quantity, gravitation, with a geometrical quantity, the curvature of space and time. The general theory of relativity was published in late 1915, well into World War I, in the leading physics journal of the time, the *Annalen der Physik*.

When Einstein published his 1915 paper, there was, of course, the already established Newtonian theory of gravitation, and the predictions of the two theories were somewhat different. Both theories held that light should bend near a massive body as a result of the gravitational attraction of the body. In particular, light coming from a star observed near the edge of the Sun will experience a small deflection as a result of the Sun's gravitation. The deflection predicted by relativity theory is about twice as large as the value given by the Newtonian theory. Observations of stars near the Sun should therefore provide a crucial test to distinguish between the two theories. Unfortunately, the only time it is possible to see stars close to the Sun is during a total eclipse.

In 1919 the English astronomer Arthur Stanley Eddington (1882–1944) led a solar eclipse expedition to test Einstein's prediction. The path of totality of the eclipse on May 29 passed from West Africa southwest to South America. Eddington and a colleague voyaged to the island of Principe off of Africa, while another team of scientists traveled to Sobral in northern Brazil. Photographic plates were exposed during the eclipse and compared to nighttime plates of the same star field taken at a different time of the year. By comparing the relative positions of the stars on the two plates, Eddington obtained an estimate of the deflection resulting from the Sun's gravitation.

At a historic joint meeting of the Royal Society and the Royal Astronomical Society in November 1919 Eddington reported that the results of the expedition confirmed Einstein's theory. Alfred North Whitehead described the mood of the meeting as follows:

> The whole atmosphere of tense interest was exactly that of the Greek drama. We were the chorus commenting on the decree of destiny as disclosed in the development of a supreme incident. There was dramatic quality in the very staging—the traditional ceremonial, and in the background the picture of Newton to remind us that the greatest of scientific generalizations was now, after more than two centuries, to receive its first modification. (Bernstein 1973, 119)

Eddington's confirmation was reported widely in the press, and Einstein became a famous figure in Britain and North America.

Although Eddington's result seemed conclusive at the time, it was later subject to criticism by scientists and historians and is today seen as unreliable. It was Eddington's authority as a scientist rather than the observations themselves that led the 1919 eclipse expedition to be perceived as a decisive confirmation of the theory of general relativity. (Detailed evidence for this assertion is presented by Collins and Pinch (1993).) In fact, confirmation of the theory would not take place for over 40 years. With the development of radar technology after World War II, researchers were able to bounce radar beams off nearby planets and measure the influence of the planetary and solar gravitational fields on the radar trajectories. Pioneers in this investigation were the Lincoln Laboratory in Massachusetts, the Arecibo facility in Puerto Rico, and the Jet Propulsion Laboratory in California, which first made radar contact with the planets in the early 1960s. The predictions of general relativity could be subjected to direct experimental tests, and these have largely confirmed the theory. General relativity has provided a very successful mathematical formalism to describe the universe as a whole. In addition, a range of astronomical phenomena, from extremely dense stellar objects to the gravitational lensing of distant galaxies, have been explained using Einstein's theory of gravity.

RELATIVISTIC COSMOLOGY

Gravity is the only one of the fundamental forces of physics that acts over the large distances between the stars and nebulae that are of interest in astronomy. From the initial formulation of Newton's theory of gravity, there were efforts to analyze the problem of the gravitational interaction of the whole universe, with results of some note being achieved in the late nineteenth and early twentieth centuries by Neumann, von Seeliger, Charlier, and Schwarszchild. Einstein's general relativity offered a new and mathematically sophisticated theory of gravitation, and it was not surprising that attempts were made to investigate the implications of this theory for cosmology. In 1917 Einstein published an important paper titled "Cosmological considerations concerning General Relativity," in which he described the universe as a whole in terms of the relativistic field equations. This paper stimulated much further work on theoretical cosmology, a field that was dominated for the next two decades by the perspective of general relativity.

Einstein advanced his cosmological solution without any particular attention to contemporary discussions by astronomers about the large-scale character of the universe. The one observation that he cited in his paper was the apparent fact of the small velocities of the stars. He took this to mean that the universe was roughly static in the sense that its disparate parts were not subject to any large systematic velocities. An innovative feature of his approach was his use of Riemannian geometry to describe the universe. In doing this he was following the precedent of such mathematically inclined researchers as Newcomb and Schwarszchild. In a Riemannian universe the world is taken to be a finite but unbounded geometrical structure, with constant positive curvature. The analogy that is usually given to help visualize such a geometry is the surface of a sphere, which is finite in area but unbounded in the sense that it is possible to go forever on the surface in any given direction. The radius of the sphere is a measure of how much it curves at a given point (the radius being inversely proportional to the curvature), and this radius has the same value for every point, being also constant with respect to time.

Researchers before Einstein had determined that in a static unbounded universe the Newtonian gravitational potential had to be modified so that the force of gravity at large distances dropped off more quickly than the square of the distance. Otherwise, even the slightest variations in density would accumulate over cosmic distances and produce large velocities in parts of the universe. The introduction of a modified gravitational force law reduced to assuming that at large distances, there was a force acting against gravity, a kind of cosmic repulsion. Einstein found it necessary to introduce a similar modification into his theory. In a static relativistic universe the field equations needed to be supplemented by the addition of a term seeming to correspond to a universal force that acted in an opposite sense to gravitational attraction. The lambda term, as it has since been called because Einstein used the Greek letter λ to denote it, was an essential part of the equations used by him to describe the universe. (In later physics, λ would be referred to as "the cosmological constant.") In Einstein's static closed solution the universe possessed a finite volume and mass, and lambda was proportional to the average density of the universe. The radius of curvature of the universe was a constant that was inversely proportional to the average density of matter in space.

In the decade following 1917, several mathematical physicists produced cosmological solutions of the field equations that were different from Einstein's. At the end of 1917 the Dutch astronomer Willem de Sitter (1872–1934) proposed a rather strange solution of the field equations for the case of a static universe that contained no mass. In principle, such a solution might describe a universe that has a very low average mass density. As before, the field equations were assumed to contain a lambda term. In a de Sitter universe the spectra of light from distant sources will show an increase in wavelength. This is a relativistic effect that results from the slowing down of clocks at large distances and a concomitant slowing of the frequency of atomic vibrations. This pattern of spectral shifts to lower frequencies became known as the de Sitter effect and was widely discussed in the 1920s.

In a 1922 book on general relativity Arthur Eddington offered a physical interpretation of the de Sitter effect. In addition to the slowing of the atomic vibrations in distant sources, there was another process at work. There was no mass in de Sitter's universe. Hence if a test particle is introduced into it, the particle will be acted upon only by the cosmic repulsion corresponding to lambda, a repulsion that will be stronger the farther the particle is from the observer. Hence a distant source will appear to be scattering or receding from the observer. Eddington cited observations sent to him by the American astronomer Vesto Slipher indicating a pattern of spectral red shifts in spiral nebulae, although he admitted that there were too many uncertainties in the data to allow for any definite conclusions. (We examine Slipher's work in more detail in the next section.) During the 1920s, other theoreticians challenged Eddington's analysis, pointing out that there should also be a tendency in de Sitter's solution for test particles to approach the observer.

Alexander Friedmann (1888–1925) was a Russian meteorologist who published two articles on relativistic cosmology in 1922 and 1924 in the German physics journal *Zeitschrift für Physik*. His most significant finding was that there were relativistic solutions in which the distance between any two points in the universe was steadily increasing or steadily decreasing with time. In such dynamic solutions, there was a function $R(t)$ of time that increased or decreased with increasing t. Suppose at a given time t_0 that the distance between two particles is one. Then, the distance between the particles at any subsequent time t will be $R(t)$. In later cosmology the scale function R would become a basic element of the theory. In the first paper Friedmann derived Einstein's and de Sitter's universes as special cases and also exhibited a class of solutions of finite positive curvature in which the curvature changed with time. In the latter case the function $R(t)$ may be understood as the radius of curvature of the universe at time t. In the second paper Friedmann considered universes of negative curvature and obtained some interesting results. He showed that there was no static solution, that is, no counterpart to the Einstein universe for negatively curved spaced. He also showed that negatively curved worlds may be spatially infinite.

In Friedmann's dynamical solution the whole universe was subject to a unitary change of position and motion with time. The existence of the scale function $R(t)$ dependent only on time assumes that there is no privileged point in the universe, that the change in distance between any two points that occurs as a result of the dynamical evolution of the universe is given by one and the same function $R(t)$. This assumption would later be called the cosmological principle and was also a basic postulate of Einstein's original theory. In Friedmann's world the dependence of R solely on time implied that the universe was isotropic, that is to say, that it evolved about any given point in the same way in all directions. In a dynamical solution in which R is increasing with time it also makes sense to speak of a moment of creation and of the age of the universe, in Friedmann's own words, "The time since the creation of the world is the time which has passed from the instant when space was a point until the present situation" (Kragh 1996, 24). Friedmann succeeded in deriving a very simple differential equation describing how R changes with time in terms of the parameter λ.

It is of interest that Einstein at first believed the Friedmann solutions were in error, although he quickly acknowledged the mathematical correctness of Friedmann's derivation. Throughout the 1920s Einstein continued to doubt that dynamical solutions were of much interest. Friedmann, who died of typhus in 1925, seems himself to have been fairly unconcerned with the observational import of his theoretical work.

As the 1920s progressed, there were increasingly strong indications from the field of nebular astronomy of observational findings that were relevant to the geometric solutions being devised by mathematical cosmologists. In 1927 the Belgian physicist Georges Lemaître, independently of Friedmann, hit upon a dynamical solution consisting of a closed universe of positive curvature. Lemaître's study was more restricted than Friedmann's since he considered only universes with finite radii of curvature. However, unlike Friedmann, Lemaître was aware of the results coming from the large American observatories. During a trip to the United States in the mid-1920s he had been present at the meeting of the American Astronomical Society where Hubble had announced his study of Cepheid variables in M 31. He also visited the Lowell

Observatory and the Mount Wilson Observatory in California, perhaps the two most important scientific stations anywhere involved in investigating the large-scale structure of the universe. Lemaître noted near the beginning of his paper that we could suppose the universe to be similar to a rarefied gas in which the molecules were galaxies. He consciously interpreted his relativistic dynamical solution as a real physical description of the universe, stating that the "receding velocities of extragalactic nebulae [i.e., galaxies] are a cosmical effect of the expansion of the universe" (Kragh 1996, 30).

Although Lemaître would later be heralded as the father of modern cosmology, it is important to appreciate how different his point of view was from today's conceptions. A question he raised in the conclusion was the possible appearance of ghost images of stars and nebulae. In a closed universe a light ray emitted by a source in a direction opposite to the direction to the Earth should travel through space and curve around until it reaches the Earth from a direction 180 degrees from the source. Hence we should observe a ghost image of the source opposite to it in the sky. Lemaître suggested that the spectral shift associated with universal expansion would shift the light circling the universe to the infrared during its long journey, and it is for this reason that no ghost images are observed. (The usual explanation at this time for the nonobservance of ghost images in a closed universe was the absorption of light as it travels through space.) He also thought that in a static universe the circuit of light around the curved universe and back to its starting point would lead to an accumulation of radiation. It is this surplus of radiation that is the cause of the expansion of the universe.

In 1928 the young American researcher Howard Robertson (1903–1961) published some important mathematical refinements of relativistic cosmology. Robertson had been a graduate student at Caltech in Pasadena, the academic base for the Mount Wilson Observatory, and enjoyed contact with the astronomers there. He and Caltech scientist Richard Tolman (1881–1948) were in a better position than other theorists to be informed of ongoing developments in nebular astronomy during the 1920s. In the 1928 paper he considered a dynamic universe and derived a relation giving the velocity of recession as a linear function of distance. He suggested that the astronomical data (gathered by Slipher and Hubble, whom we consider in the next section) led to a "rough verification" of this relation. However, Robertson was primarily interested not in this fact but in such mathematical questions as what the data revealed about the numerical value of the radius of curvature of the universe in a cosmological solution based on elliptical geometry.

DOPPLER SHIFTS AND HUBBLE'S LAW

With the rise of the Industrial Age in the nineteenth century, rapidly moving trains became a common sight in the countryside of Europe. The sound of a whistle emitted by a train and heard by someone standing by the side of the tracks rises in pitch as the train approaches the person and lowers in pitch as the train recedes. In 1842 the German physicist Christian Doppler (1803–1853) gave an explanation for this phenomenon in terms of the periodic nature of sound. Sound travels in a wave form as a succession of compressions and rarefactions in the air. The pitch or frequency of the sound is a function of the distance between two successive compressions; the shorter the distance, the higher the pitch. If the source of the sound is moving toward the observer, then the wavelength will be shortened since the source moves forward a small distance during the time interval between consecutive emissions of a compression. Similarly, the wavelength will be lengthened if the source is moving away from the observer.

The consolidation of the wave theory of light led, in the nineteenth century, to the investigation of phenomena connected to the periodic nature of light emission. The transmission of light is analogous to the transmission of sound waves. Although light waves vibrate in a direction transverse to the direction of their motion, whereas sound waves vibrate in the direction of their motion, both forms of propagation involve the successive emission of wave crests and troughs from a source. In the case of light, pitch corresponds to frequency, with blue light possessing a higher frequency than red light. Doppler reasoned that if a star was moving away from us, we should see some reddening of its light, whereas if it was moving toward us, we should see some blueing of its light. He advanced this effect as an explanation for the changes in brightness observed in variable stars.

It was later realized that the shifts in light frequency that occur because of the motion of variable stars are much too small to account for their changes in brightness. Nevertheless, the effect Doppler had predicted was a genuine one and would become the basis of an important method in stellar astronomy. With the development of techniques to represent the spectrum of the Sun and stars as a sequence of emission and absorption lines, it became possible to determine very precisely the wavelength of light from these objects and thus to identify very small changes in wavelength that occur as a result of the star's motions. Because these motions are along the radius or line joining the observer to the star, the velocities are known as radial velocities. In 1868 William Huggins obtained a spectrum for the star Sirius in Canis Major and found a radial velocity of 29.4 miles per second (48 kilometers per second). In 1871 Hermann Vogel (1834–1898) examined the spectral lines on the east and west limbs of the Sun and identified a shift in their position resulting from the Sun's rotation.

At the end of the nineteenth century the use of stellar spectra to analyze radial velocities became an active field of research. Velocity shifts turned out to be a key to understanding the behavior of an important class of variable stars known as eclipsing-binary stars, objects which exhibit short-term and highly regular changes in brightness. The best known representative of this class of variables is the star Algol in the constellation of Perseus. Spectral analysis of Algol revealed that it consisted of a pair of stars in rapid rotation about their common center of gravity and that the changes in brightness resulted from one member of the pair being eclipsed by the other. Doppler shifts also proved to be useful in determining the motion of the solar system relative to stars in the solar neighborhood. It was through a statistical study of stellar radial velocities that the rotation of the galaxy was first detected.

The Lowell Observatory in Flagstaff was a major American observatory established in 1894 by Percival Lowell (1855–1916). The clear skies and high-altitude conditions of northern Arizona were well suited to the observation of faint objects. The main instrument at Flagstaff was a 24-inch (60 centimeter) refracting telescope built by the firm of Alvan Clark and Sons. The Lowell Observatory was best known for work in planetary astronomy and, more particularly, for its observations of Mars. Lowell's sensational findings of possible advanced engineering structures on the surface of Mars had attracted worldwide attention. In 1912 the Flagstaff astronomer Vesto Slipher embarked on a project to study the spectra of the white, or spiral nebulae, as they were then called. Some astronomers of the period held that in spiral nebulae we were actually observing the formation of new solar systems in space, so Slipher's study of these nebulae was in keeping with Lowell's emphasis on planetary astronomy. The Clark refractor was fitted with a small dispersion spectroscope and a fast-exposure camera that recorded with precision the absorption lines of the faint extended nebular images.

Much to his surprise, Slipher discovered that the radial velocities of the nebulae were a good order of magnitude larger than any of the velocities observed in stars of the galaxy. He found both positive and negative velocities, indicating that some of the nebulae were approaching the Sun and some were receding from it. He initially interpreted this finding in terms of the nebular hypothesis as an effect resulting from the process of star formation, but he soon abandoned this explanation. He conjectured instead that the Sun was moving through space among the spiral nebulae, and the radial velocities resulted from this motion. Although on a much larger scale, this drift effect was similar to the one observed in stars in the neighborhood of the Sun, indicating the motion of the Sun relative to these nearby stars.

The first publication of Slipher's results took place in 1914. For the next decade Slipher worked virtually alone on the problem of nebular spectra. By 1922 he had accumulated spectral data on 41 spirals, which was sent to Eddington and published in the latter's book *General Theory of Relativity*. The data showed a preponderance of red shifts (35 red shifts versus 6 blue shifts), indicating a clear pattern of recession. This seemed to refute the explanation for the radial velocities as resulting from solar motion relative to the spirals since according to this hypothesis, one should observe red and blue shifts in approximately equal numbers. Throughout the 1920s, observationalists stuck to a modified version of this hypothesis, holding that the equation for solar motion relative to the spirals needed to be supplemented by a special term to account for the systematic occurrence of red shifts. For a given spiral with outward radial velocity V and coordinates α (right ascension) and δ (declination) this equation takes the form

$$* \; V = X \cos\alpha \cos\delta + Y \sin\alpha \cos\delta + Z \sin\delta + K,$$

where the K term represents the additional component of velocity that is needed to produce the observed red shift.

By 1925 Hubble had shown that the white nebulae are objects similar to our own galaxy scattered throughout distant space. With this finding the large-scale structure of the universe was clarified and attention shifted to a more detailed study of the properties of nebulae. The very large radial velocities of these nebulae were certainly consistent with their position in the universe far beyond the gravitational range of the galaxy. Milton Humason (1891–1972) at Mount Wilson took over from Slipher the project of measuring and analyzing nebular spectra. The Hooker 100-inch telescope was perhaps the only instrument in the world with sufficient light-gathering power to allow for a really systematic investigation of this sort. Humason was an expert in the measurement of the size of the shifts, while Hubble concentrated on obtaining reliable distance indicators to the nebulae. The difficult and more theoretical part of the investigation was Hubble's. There is a very considerable degree of variation in the intrinsic brightness of nebulae. It does not follow that a dimmer galaxy is necessarily farther away than a brighter one. One must rely on statistical methods and reasonable supposition. For example, it is evident, given a uniform distribution of galaxies in space, that there will be a correlation between apparent and intrinsic luminosity: on average, the fainter a galaxy is, the farther away it is.

Hubble took the available data and tried to fit them to a group of equations of the form (*). He did so fairly cautiously because there had been earlier attempts to do this for globular clusters that had turned out to be premature and incorrect. The German astronomer Carl Wirtz (1876–1939) had conjectured in 1924 that a relationship between distance and red shift held for spiral nebulae, although the data were too meager to permit any definite conclusion. Hubble mostly worked from Slipher's spectrographic measurements, supplemented by a few observations of Humason's. By 1929 he was confident enough to be able to deduce that the K term in (*) had the form kr, where r was the distance to the spiral nebula. The first term on the right side of (*) resulted from the solar motion, which would become negligible in comparison to kr as r became larger. Neglecting the solar motion, the recessional velocity was given by the positive quantity kr, implying that the nebulae were moving away from the solar system with velocities proportional to their distances from us. In recognition of Hubble's work, later astronomers substituted the constant H for k, and the relationship

$$(**) \; v = Hr$$

became known as Hubble's law, giving the red shift as a function of distance. The velocity-distance diagram from Hubble's original paper is presented in figure 8.1. This law is independent of the direction of the nebulae in the sky, thus implying that all nebulae at a given distance have the same red shift. Although Hubble followed convention and expressed these shifts in terms of

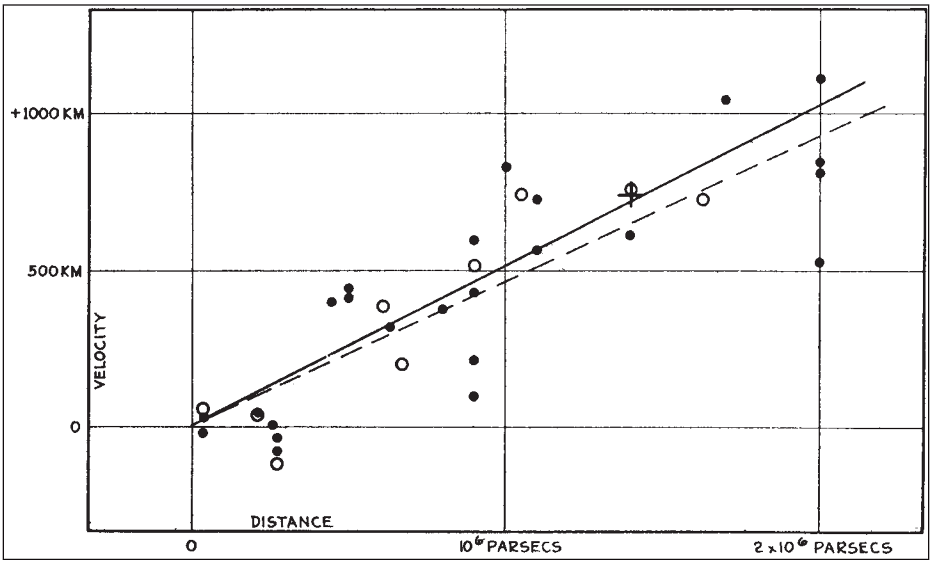

Figure 8.1: Hubble's historic red shift–distance graph (1929). The Thomas Fisher Rare Book Library, University of Toronto.

radial velocities, he believed that they might well be due to some cause other than recessional motion and the Doppler effect. In subsequent writings he tended to refer to the velocities in his law as "apparent" radial velocities to emphasize this point.

In the 1929 paper Hubble verified (**) up to a distance of about two megaparsecs, according to the distance scale for galaxies then in use. (A star one parsec from the sun exhibits an annual parallax of one second of arc; a parsec is approximately 3.26 light-years. A megaparsec is a million parsecs.) Humason and Hubble continued their work and two years later showed that (**) is valid for galaxies out to 32 megaparsecs. The approximate correctness of Hubble's law has since been established for galaxies at any distance, and Hubble's law is now regarded as the fundamental law of modern cosmology. An immediate and pressing issue facing Hubble was to determine as precisely as possible the value of the constant *H*. This constant is today given in terms of units of kilometers per second per megaparsec, abbreviated to km s^{-1} Mpc^{-1}. Here, radial velocity is measured in kilometers per second and distance in megaparsecs. Hubble estimated *H* to be around 520 km s^{-1} Mpc^{-1}. Hence for every megaparsec increase in distance, the recessional velocity of a galaxy increases by 520 kilometers per second . As we shall see, the very high value for *H* derived by Hubble and other researchers created problems for cosmologists for decades to come.

GENERAL RELATIVISTS AND THE EXPANDING UNIVERSE

Hubble's finding that nebular red shifts are linearly proportional to distance was interpreted by researchers working in general relativity as conclusive evidence that the universe is expanding. The red shifts, which were, for the sake of convention, nominally listed by Hubble as radial velocities, were regarded as true velocities resulting from the motions of galaxies relative to the solar system. Theoreticians also accepted the cosmological principle, which stipulated that there is no special or privileged vantage point in the universe. What we observe from one place in the universe is on a large scale the same as what is observed from any other place in the universe. Hence it follows from Hubble's law that every two galaxies in the universe are moving away from each other with a velocity that is proportional to the distance between them. An analogy to help understand the concept of the expanding universe is a balloon with dots distributed over its surface. The surface is the universe, and the dots are the galaxies. As the balloon inflates, the surface expands, and each dot moves away from every other dot with a speed proportional to the distance along the surface separating them.

The original static Einstein solution and the de Sitter solution were not able to describe an expanding universe. The de Sitter solution predicts a general reddening at large distances, but this model contains no matter, and the spectral shifts do not represent real velocities. It was apparent by the late 1920s that the density of matter in the universe was nowhere near small enough for de Sitter's solution to be valid. However, soon after the publication of Hubble's law, researchers in general relativity such as Eddington and de Sitter himself called attention to Lemaître's 1927 paper and his dynamical solution, which did describe an actual expanding universe. This paper was translated into English and published by the Royal Society in 1931, thereby ensuring the wide dissemination of Lemaître's results. Friedmann's and Robertson's work also became the subject of renewed interest among relativists. Cosmological research within the theory of relativity during the 1920s was seen by scientists as leading up to Hubble's discovery, with Lemaître's contribution being regarded historically as the most significant. It was soon recognized that a range of relativistic solutions were available to describe the evolution of an expanding universe. It is customary to describe these solutions in terms of what is known as an R-t diagram, giving the scale function $R(t)$ graphed as a function of t. Three important models during this period were advanced by Lemaître, by Eddington, and by Einstein and de Sitter.

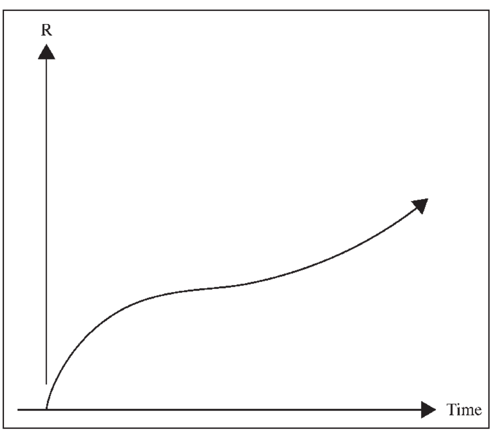

Figure 8.2: Lemaître's universe (1932).

The Lemaître universe was proposed by the Belgian physicist in 1932 and differed somewhat from the solution given in his 1927 paper. One supposes that the world begins from a single point and expands outward, with the rate of expansion decreasing with time. At a certain point the rate of expansion begins to increase with time and continues in this way forever (figure 8.2). An essential feature of this universe is the presence of a positive cosmological constant in the gravitational field equations. Lemaître's model was not a very popular one and was hardly considered at all in the literature of the second half of the twentieth century. However, since 1998 and the discovery of universal acceleration, there has been a marked revival of interest in it.

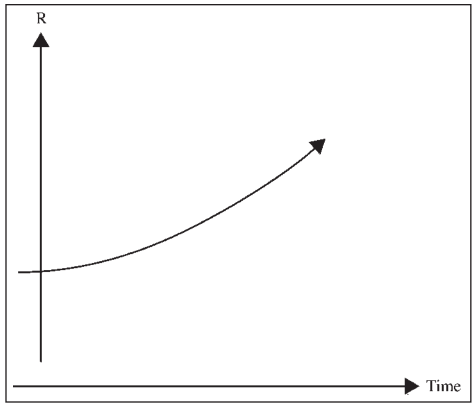

Figure 8.3: Eddington's universe (1933).

Eddington's universe was very similar to the one presented in Lemaître's 1927 paper and was described in his popular 1933 book *The Expanding Universe.* Eddington made essential use of the cosmological constant, a fact that was in keeping with the curious emphasis he placed on the numerical properties of the constants of nature. In this conception the universe begins in a static Einstein state. At a certain instant, there is instability or a disturbance, and the universe begins and continues henceforth to expand outward (figure 8.3). What exactly triggers the expansion is not identified, but the initial static state may be regarded as a time when the stars and galaxies were formed. The cosmic repulsion corresponding to the cosmological constant is responsible for the continued expansion of the universe. Despite the enthusiasm of its founder, the Eddington universe has never enjoyed popularity among cosmologists.

Einstein and de Sitter devised a solution which did away with the cosmological constant. As Einstein saw it, it was no longer necessary to include the constant in the field equations because the universe is no longer assumed to be static. In the Einstein–de Sitter universe the world expands from a point, and the rate of expansion continually decreases with time. This decrease is caused by the braking force of gravity (figure 8.4). In the second half of the twentieth century this solution, or something like it, was the one preferred by cosmologists, although since 1998, its validity has been called into question. From its initial formulation, an objection to the Einstein–de Sitter world concerned the very low age for the universe implied by it. If we assume that the expansion of the universe is slowing down, then an upper bound on the age of the universe is given by $1/H$. This quantity is obtained by running the expansion at its current rate backward and calculating the time from the present to the birth of the universe at a point. Using Hubble's value for H of 550, this procedure leads to an age of less than two billion years, a value that was very difficult to square with the time scales required in geology, much less those required for the universe as a whole.

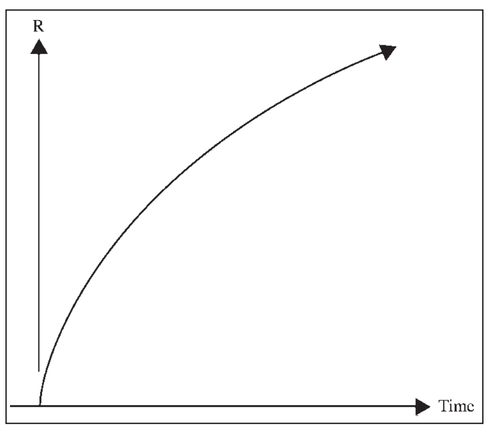

Figure 8.4: Einstein–de Sitter universe (1932).

The so-called age problem would be a recurrent theme in modern cosmology up to the very present.

HUBBLE AND THE RELATIVISTS

In his 1929 paper Hubble had suggested that the red shifts observed in spiral nebulae might be evidence of a de Sitter effect. At the time that he wrote de Sitter's solution was no longer considered very plausible by relativistic cosmologists, but Hubble seems still to have placed some credence in the Dutch astronomer's theoretical work. One thing that stands out in Hubble's subsequent writing is his refusal to accept the interpretation of the red shifts as resulting from actual velocities of recession. He believed that the they might be due to some unknown effect, possibly resulting from energy dissipation as light travels through space. His colleague Fritz Zwicky (1898–1974) at Caltech explored such a "tired light" explanation of the red shift. It seemed natural if a process of this sort existed that the effect would be proportional to the distance traveled by light through space.

Hubble presented a detailed account of his observational work in his 1936 book *Realm of the Nebulae*. He made it very clear that the concept of an expanding universe was something that belonged to the general theory of relativity. In fact, he referred to the "expanding universe of general relativity" (1936, 198) and implied that the acceptance of the red shifts as actual velocity shifts was dependent on acceptance of this theory. Any other explanation of the red shifts would require some new principle of physics, but Hubble felt that this might very well be necessary. His doubts about expansion intensified in the years that followed. Given the very high number assigned to Hubble's constant, it seemed that the universe would have to be very dense, small, and young, much more so than was indicated by general observation. In a 1942 article in *Science* he concluded that "the empirical evidence now available does not favor the interpretation of redshifts as velocity shifts" (214).

Hubble's reservations about expansion were consistent with his distrust of theory and were supported by the difficulties evident in the very young age of the universe implied by the then accepted rate of expansion. This problem was resolved, at least temporarily, by the recognition in the 1950s that the distance scales being used in nebular astronomy were drastically underestimating distances to galaxies. Walter Baade (1893–1960), an astronomer at the Mount Palomar Observatory in California, realized that Cepheid variables actually fall into different classes, with different average intrinsic luminosities. (Baade's research is discussed in chapter 9.) The variables being used to measure the distance to galaxies were in fact of considerably larger intrinsic brightness than was initially assumed to be the case. This in turn implied that they must be farther away, and the revised distance measurements implied that the universe was larger than it was previously believed to be. The time it takes the universe to expand to its current state will depend on how large it is, and the new enlarged distance scales implied that the age of the universe must be much greater than the value of under two billion years derived from

Hubble's original estimate of the rate of expansion. The value of Hubble's constant has continued to be revised as distance estimates have been refined. By the 1980s the accepted value of *H* was around 50 km s^{-1} Mpc^{-1}, implying an age for the universe of 18 billion years, a length of time which seemed to be sufficient to account for the evolution of globular clusters, believed to be the oldest objects in the universe. It would certainly be undesirable to have a universe that is younger than the oldest objects in it. In the 1990s, researchers associated with the Carnegie-Mellon Foundation began a concerted program that is still underway today to evaluate the Hubble constant and remove some of the uncertainty concerning its value. The value obtained by this group is somewhat higher than 50, implying a younger universe and giving rise once again to the age problem. The currently accepted value of the constant (2006) is around 70, with an estimated age of the universe of about 13 to 14 billion years. Astrophysicists working on stellar evolution have struggled to revise their theories in order to accommodate their estimates of stellar ages to this lower value for the age of the universe.

UNIVERSAL EXPANSION AND RELATIVISTIC COSMOLOGY

An important historical question related to Hubble's momentous 1929 discovery concerns its relationship to contemporary work in the general theory of relativity. That the invention of cosmological solutions based on general relativity occurred at precisely the same time that Slipher and Humason were beginning to detect large systematic nebular red shifts was simply a coincidence. The two developments were largely independent. The advances in telescopic instrumentation that made the nebular research possible followed from improvements in technology and the increased financial support for astronomy in America from government and philanthropic foundations. General relativity, by contrast, developed within a central European scientific culture, with a strong emphasis on advanced mathematics and pure theory. In retrospect, it seems that Hubble's relation would have been detected inevitably with improvements in the size, quality, and location of observing facilities; it could well have been discovered earlier or later. It is nonetheless a fact that throughout the decade leading up to the 1929 breakthrough, speculation about the red shifts was often tied in with theorizing in relativistic cosmology. Hubble was aware of de Sitter's writings and explicitly cited the de Sitter effect in the 1929 paper. It was also the case that general relativists such as Eddington were among the first to explore the implications of Hubble's discovery in terms of dynamical world solutions.

Although the observational discoveries of the period were independent of theoretical work in relativistic cosmology, the converse cannot be said to be true. At the time he wrote his 1917 paper de Sitter was aware of Slipher's findings through a report on them published by Eddington in the *Monthly Notices of the Royal Astronomical Society*. A more detailed description of these findings was presented by Eddington in his 1920 book *Space, Time and Gravitation,* where he wrote, "The motions in the line-of-sight of a number of nebulae have been determined, chiefly by Professor Slipher. The data are not so ample as we should like; but there is no doubt that large receding motions greatly preponderate" (161) It is significant that Friedmann, in his 1922 paper, cited both de Sitter's paper and the French translation of Eddington's book. The very high red shifts reported in these sources certainly would have raised doubts about Einstein's assumption of a static universe and suggested the possibility of dynamical cosmological solutions of the field equations. It is also known that Slipher's findings were reported in 1923 in a widely read Russian scientific magazine published in Petrograd, Friedmann's home city. The case of Friedmann is interesting because he, more than Lemaître, is often seen as someone who was uninfluenced by observation and whose geometric solutions represented a prescient achievement of pure theory. It should be noted that one of the key assumptions of his relativistic solution, the dependence of the scale function only on the time, was later found to hold for the universe as a whole. The relativists were not working in complete isolation from observational work, although it is nonetheless the case that the emergence of dynamic theoretical solutions at precisely this time was a highly unusual event of which there are few parallels in the history of science.

In the work during the 1920s on relativistic cosmology, no one, with the possible exception of Carl Wirtz and Howard Robertson, had predicted a linear red shift–distance relation or made an attempt to configure the spectroscopic data to what was then known about distances to nebulae. The very status of the nebulae, much less their distances, was only being clarified during this period. To understand why an expansionist interpretation of the universe was not generally considered before 1930, it is also important to understand the intellectual atmosphere of the 1920s. What most struck scientists of the period about the spectroscopic data was the fact that it might well consist of a verification of Einstein's radical new theory of gravity. It was this theory and its revolutionary implications that excited scientists. The nebular spectral shifts seemed to offer clear and unequivocal evidence for general relativity, much clearer than the fine discriminations involved in interpreting eclipse observations. The focus of scientific attention was on the meaning of the observational data for general relativity and not on the possible fact of universal expansion.

After 1929, when expansion seemed to be the most probable interpretation of Hubble's law, the general relativists were able to turn to the until then neglected dynamical solutions of Friedmann and Lemaître. It is worth noting that Hubble regarded the concept of an expanding universe as a notion rooted in the general theory of relativity. In retrospect, it seems clear that if one accepts the red shifts as due to real velocities—and this is the most obvious explanation—then it follows that the universe is expanding, a conclusion that requires for its warrant no particular theory of gravity, much less the formidable machinery of general relativity. In 1933 Eddington wrote that the theorists had for the past 15 years been expecting something "sensational" along the lines of Hubble's discovery (and there could be no finding more sensational than Hubble's) and seemed almost to be taking some credit on behalf of the theorists for the discoveries coming from the great American observatories.

There seems little doubt that Hubble was concerned with emphasizing the purely phenomenological character of his result: its independence from contemporary theorizing in mathematical cosmology. To concede that the red shifts were recessional velocities was in Hubble's view to accept an underlying theoretical approach to cosmology and possibly to suppose that an achievement of skilled observation owed something to the "invented universe" of the theorist. As Hubble (1936, vii–viii) emphasized, "the conquest of the Realm of the Nebulae is an achievement of great telescopes."

CONCLUSION

The discovery of the red shift law and its confirmation as an expression of universal expansion was the most important event in the history of astronomy since Copernicus and one of the most important events in the whole history of civilization. This discovery has few parallels in the history of science: the revelation of a profound truth about the physical world on the largest scale with absolutely no anticipations in earlier work on the subject. Prior to Slipher's nebular spectroscopic investigations, there was not the slightest hint in the speculative or serious astronomical literature of the idea of an expanding universe. The events in observational cosmology of the 1920s and 1930s were so momentous, unexpected, and singular that the scientists of the period were hardly able to grasp the intellectual and historical magnitude of what had happened.

9

FROM UNIVERSAL EXPANSION TO THE BIG BANG

INTRODUCTION

In the first decade following Hubble's discovery of the red shift law, there was some disagreement concerning its meaning and cosmological significance. There were researchers such as Fritz Zwicky and Hubble himself who believed that the nebular spectral shifts occurred as the result of some physical process acting on light as it travels the long distances in space from source to observer. It was natural to assume that the effect of this process would be proportional to the distance traveled by the light. According to this hypothesis, the universe is static or at least is not subject to any large-scale systematic motions. Zwicky and other proponents of the "tired light" hypothesis were unable to work out a theory to explain the physical process that would lead to the observed pattern of red shifts. Although there have up to the present been occasional proponents of a static interpretation of Hubble's law, the majority of scientists have concluded that the universe is expanding, and this conviction has become the fundamental tenet of modern cosmology.

FORMULATION OF THE BIG BANG THEORY

Both Friedmann and Lemaître had devised cosmological models on the basis of general relativity that hypothesized that the universe had expanded outward from some initial time when the scale factor R was zero. In such a universe the density of matter and energy will increase as one goes backward in time, reaching infinite values at time zero. The general relativistic models were geometric structures satisfying the gravitational field equations, and the analysis of the physical conditions that must hold in them was not a matter of close concern. Friedmann died in 1925, only dimly award of the new developments in nebular astronomy. Lemaître, by contrast, was present in January 1925 at the meeting of the American Astronomical Society, where Hubble's demonstration of the extragalactic character of the spiral nebulae was announced. With the discovery of the red shift law four years later the relativistic models became something more than mathematical curiosities; the very real possibility existed that they described the physical universe in which one lived.

Lemaître believed that if the expansion of the universe was extrapolated backward in time, one was led, other things being equal, to an initial moment of creation involving conditions of extremely high density. He rejected the relativistic model proposed by Eddington, in which the universe began in a static state and then, at some moment, began to expand outward. He hypothesized that the universe began in the radioactive disintegration of a "primeval atom," a fantastic explosion that propelled the subsequent expansion of the universe. Like many cosmological theorists in the 1930s, Lemaître also believed that the expansion might be propelled by a kind of repulsive cosmic force corresponding to the cosmological constant in the field equations of general relativity. With the discovery of universal expansion Einstein himself rejected the cosmological constant, stating that its earlier introduction by him—done to preserve a static cosmos—was a mistake.

Lemaître is regarded as the father of modern physical cosmology. His idea that the universe began with an explosive event in conditions of high density attracted the attention of many theorists in the two decades following 1931. This idea formed the basis for what became known as the big bang theory of the universe. The name itself was coined by British scientist Fred Hoyle (1915–2001) in 1949 in a BBC radio lecture. Ironically, Hoyle was a proponent of an alternative cosmology (the steady state theory discussed below) and used the phrase big bang in a rather disparaging way to criticize his scientific opponents.

A key idea of the big bang theory is that the universe is evolutionary. It originated at a finite time in the past—believed by current estimates to be around 12 to 14 billion years ago—and has undergone a steady expansion and decrease in density since then. As one looks out in space, one looks back in time; it follows, according to the big bang theory, that the universe should look younger and therefore less evolved the farther one looks out. The theory is a historical one since its account of the large-scale structure of the universe is also an account of the temporal origins of the universe. In this respect, the big bang theory stands in striking contrast to both ancient Greek cosmology and to Copernican heliocentric cosmology, both of which involved no assumptions about the origins of the planetary system.

Lemaître's notion of a disintegrating primeval atom was an interesting idea, but it proved difficult to develop into a consistent quantitative model describing conditions in the very early universe. The modern hot big bang theory had its origins in the writings, during the 1940s and early 1950s, of three American specialists in nuclear physics, George Gamow (1904–1968), Ralph Alpher (1921–), and Robert Herman (1914–1997). Of the three, Gamow was most vigorous in promoting big bang cosmology, which he did in research papers as well as in popular writings aimed at a broad scientific audience.

Gamow was initially concerned with the problem of stellar nucleosynthesis, that is, with how heavier elements are synthesized from lighter elements in the interiors of stars. This problem was closely connected to the question of how stars evolve. By the 1930s it was recognized that a star's source of energy involved thermonuclear fusion in its hot and dense core. A major breakthrough occurred in 1938, when Hans Bethe (1906–2005) in the United States explicitly identified the chain of reactions by which hydrogen is converted to helium, the so-called carbon-nitrogen cycle. Carl von Weizsäcker (1912–) in Germany obtained a similar result at roughly the same time. In the proposed sequence of nuclear reactions, hydrogen is converted to helium in the cores of stars, carbon playing the role of a catalyst in the reactions. The carbon-nitrogen cycle is the main source of stellar energy. Serious problems arose when physicists tried to derive corresponding reaction cycles for the heavier elements. Gamow and others were attracted to cosmology and the big bang idea because it allowed in principle for the possibility of prestellar synthesis of the heavier elements.

The essential idea as it developed in the work of Gamow, Alpher, and Herman in the late 1940s was that the very early universe was dominated by radiation, matter being present at this time in the form of a soup consisting of protons, neutrons, and electrons. As the universe expanded, thermonuclear processes produced helium nuclei from the protons and neutrons. Further element formation followed, although the precise mechanisms for this were not spelled out. At a certain time the universe had expanded and cooled to such a degree that the matter density exceeded the radiation density; at this moment, later referred to as the decoupling time, the universe as we know it was born. In a paper in 1948 Alpher and Herman carried out some computations and concluded that "the temperature in the universe at the present time is found to be about 5° K" (Kragh 1996, 119). No one at the time viewed this as a serious empirical prediction subject to test, and the work of Gamow, Alpher, and Herman failed to attract much interest.

BRITISH COSMOLOGY AND THE STEADY STATE THEORY

The two leading figures in British cosmology during the 1930s were Arthur Eddington and the Oxford astrophysicist Arthur Milne (1896–1950). As we saw in the preceding chapter, Eddington embraced the application of the general theory of relativity to analyze nebular motions and the expansion of the universe. By contrast, while Milne took an interest in the exciting discoveries in nebular astronomy, he objected to the use of general relativity as the exclusive theoretical framework to understand Hubble's law. Milne endorsed the concept of an expanding universe but sought a simpler way of understanding this conception. In relativistic accounts the increasing separation between any two bodies in an expanding universe is understood to result from the expansion of space: the two bodies move apart because the space between them is expanding. Milne preferred instead to take space as Euclidean and invariant and to understand the separation of the two bodies as resulting from the motion of the bodies themselves in space. Consider a gas in which each particle of the gas represents a galaxy. The gas begins from a singularity. There is initially a distribution of velocities for the gas particles from zero up to the speed of light. The gas expands outward. After any given time, as viewed from any particle, the velocities of the other particles are distributed in a simple way, in which velocity is proportional to distance: particles with larger initial velocities have traveled proportionally larger distances in the time elapsed since the initial singularity. Hubble's law is thus the simple consequence of the segregation of velocities that occurs as the swarm of particles expands.

In order to develop this conception, Milne made use of the special theory of relativity and something he called the cosmological principle, a conception he attributed to Einstein and that he seemed to regard as a general statement of the relativity idea. As Milne formulated it, this principle asserted that the events and the laws of nature were the same for all observers given that their reference frames were similarly situated with respect to the phenomena. Insofar as universal expansion was concerned, the principle was supported by observation. The magnitude of a nebula's red shift was independent of its direction in the sky, and so the recessional velocity in Hubble's law was proportional only to the distance. Milne seemed to have viewed the cosmological principle more generally as a logical postulate that characterized one's experience of physical reality at the most basic level.

By the very nature of Milne's model the universe expands outward. By contrast, models based on general relativity describe either an expanding or contracting universe, and there is no theoretical reason why the actual universe that we observe happens to be expanding. Milne considered this fact a point in favor of his theory. His work was embedded in a more general approach to the theory of space and time that he called kinematic relativity. This subject took Euclidean geometry as basic and rejected the fundamental idea of the general theory of relativity, according to which the action of gravity is given in terms of the geometrical structure of space and time. Milne rejected such notions as the expansion of space and the curvature of space, arguing that these concepts lack empirical referents and therefore have no scientific meaning. His most important contribution, carried out in collaboration with William McCrea, involved the use of classical Newtonian gravitational theory to determine how the scale factor R changes with time. The two men succeeded in obtaining the Friedmann-Lemaître equation, a differential equation previously derived from general relativity, that gives R as a function of the time t.

Milne's willingness to challenge orthodoxy and to construct cosmologies based on simple rational considerations influenced the British approach to cosmology in the 1930s and 1940s. At the end of the war, many scientists who had engaged in war-related research turned to other scientific projects. Three such figures were the young Cambridge physicists Herman Bondi, Thomas Gold (1920–2004), and Fred Hoyle. Bondi and Gold were Austrian refugees from Hitler's Europe, while Hoyle had studied stellar physics at Cambridge as a scholarship student. Bondi and Gold followed Milne in working outside the confines of general relativity, while Hoyle attempted to situate the theory within a general relativistic framework.

Hoyle recalled that the idea for the steady state theory originated one night in late 1946 or early 1947 after he, Gold, and Bondi had watched a film at a Cambridge cinema. In the film, *Dead of Night*, an architect awakens from a dream one morning to a ringing telephone. A client has called, asking him to go to a country manor to assess a project. Upon his arrival at the manor in his motor car he is overwhelmed by a sense of déjà vu as he joins a group of manor guests. In the scenes that follow the guests recount stories of the bizarre, the psychologically strange, and the paranormal. In a culminating scene, events spiral wildly out of control at the manor, and the man suddenly awakens to discover that it has all been a dream. He picks up the phone, whose ringing had awoken him, only half remembering the details of the dream. The call contains a request to assess a job at a country manor. The final scene of the film shows the architect driving up to the manor, which appears strangely familiar to him.

Following the film, Gold was struck by the question, what if the universe is like that? What if recurrence and a kind of dynamic equilibrium are characteristic of the workings of the cosmos at the largest level? In analyzing the phenomenon of universal expansion quantified by Hubble's law, Bondi and Gold adopted what they called the perfect cosmological principle. This substantially modified Milne's principle by asserting that the universe, on a large scale, looks the same at all points in time as well as in space. The adoption of such a principle amounted to a rejection of the assumption made by both Lemaître and Milne that the universe as we move backward in time is denser, and as we move forward in time will become less dense. Bondi and Gold cited philosophical reasons in support of the perfect cosmological principle, but the primary evidence for it in the 1940s came from empirical considerations. The distance scales that were employed by astronomers at this time implied very high values for Hubble's constant, as high as 500–600 km s^{-1} Mpc^{-1}. According to big bang evolutionary models, this in turn seemed to imply that the age of the universe was fairly small, certainly no more than one or two billion years. It was unclear how the evolution of the stars and the development of the solar system and the Earth itself could have occurred within such a narrow time frame. The so-called "age paradox" would be partially resolved in the 1950s by the introduction of revised distance scales introduced by Walter Baade of the Mount Palomar Observatory (see below). However, the paradox would arise again in cosmology and has proved to be a recurring difficulty for big bang models of the universe.

Bondi, Gold, and Hoyle proposed that the world is in a steady state. As the galaxies recede outward from each other, matter in the form of hydrogen atoms is created spontaneously at a very low rate in the resulting void. Out of this matter, new stars and galaxies form, and so the large-scale density of the universe remains constant in time. The spontaneous creation of matter generates pressure, which propels the expansion of the universe; the galaxies are thrust outward at an ever-increasing rate. Hoyle, who was a gifted popular expositor of astronomy and the steady state theory, explained the mechanism in the following way (1990, 222–223): "Each new object makes room for itself among the previously existing units, forcing the previously existing units to move apart from each other, and so providing a physical *raison d'être* for the expansion of the universe.... Think of the creation as being driven by ascertainable physical processes and of the inexorable introduction of new units of creation as forcing the others apart, much as the introduction of new guests into a cocktail party forces earlier guests to move outwards from the initial gathering point, although as always in cosmology this concept has to be formulated without reference to any particular spatial centre."

It should be noted that the idea of a steady state universe involving the spontaneous creation of matter was not in itself new and had been advanced during the 1920s and 1930s by the American mathematical astronomer William MacMillan (1871–1948) and the German chemist Walther Nernst (1864–1941). These scientists questioned the running down of the universe and the inevitable increase in disorder predicted by the laws of thermodynamics. They proposed that radiation was converted to matter in the ether of space and that this matter accreted as dust to stars and replenished their source of energy. Radiation from space had been first detected by high-altitude balloon experiments in 1912. Macmillan believed that cosmic rays, as this radiation came to be called, emanated from the creation of matter in empty space. MacMillan's universe was static and unchanging at the largest scales. As MacMillan put it, "the universe does not change always in any one direction.... It is like the surface of the ocean, never twice alike and yet always the same" (Kragh 1996, 143). Although the speculative conceptions of MacMillan and Nernst failed to win much favor with scientists, their idea of a steady state universe was important in the general background to the development of the Gold-Bondi-Hoyle theory. By incorporating the crucial observational datum of universal expansion into the theory, the latter became a plausible logical alternative to the big bang hypothesis widely accepted in cosmology.

The Gold-Bondi steady state model of the universe implies a definite relation derived from Hubble's law, giving the distance of a galaxy as a function of time. At any given time the velocities of galaxies will, according to this law, be distributed in a linearly increasing way according to their distances from the observer. Since the universe is in steady state, this same distribution of velocities with distance will also hold at any future time. Consider now a galaxy at time t that is at a distance d with velocity of recession v. At a later time t' it will have moved out to a greater distance d' and must partake of the velocity that a galaxy at that distance possesses in a steady state universe. Hence in moving from d to d' the recessional velocity increases by a factor d'/d. Mathematically, this means that the scale factor is increasing exponentially with time or, equivalently, that the distance of the galaxy increases exponentially with time.

The steady state model describes a highly accelerated expanding universe, one in which the galaxies are whooshing away from us into the far reaches of space. Such a universe is very different from the sort of universe posited in any of the Milne-style or big bang models.

In its detailed form, there were two versions of the steady state theory, one advanced by Bondi and Gold and the other by Hoyle. The former stressed the philosophical basis of the theory and its independence from general relativity; the latter tried to develop the theory using relativity in a way that was consistent with physical cosmology. Hoyle proved to be the most persistent and enduring defender of the steady state world picture. He cited the problem of galaxy formation in the universe. As one looks out in very distant space, galaxies are sighted; according to the big bang theory, they must have been around quite early in the universe. It is not at all clear how compact, gravitationally bound objects such as galaxies could have formed out of diffuse matter in conditions of very high energy, so close in time to the putative initial explosion that created the world. The problem of galaxy formation is today a very thorny one for the big bang theory. In a steady state model, by contrast, matter is formed in the void opened up between the separating galaxies; in these rather placid conditions the formation of galaxies would seem to be a fairly natural event.

Adherents of the big bang idea, unable to account for the production of heavy elements in the interior of stars, had supposed that these elements were synthesized in the conditions of extreme temperature and density in the early universe. The steady state theorists questioned whether it was even permissible to assume that the laws of physics remained valid under such highly singular conditions. Since such a process of element formation was not available to them, they were motivated to investigate more seriously the basic problem of stellar nucleogenesis. Hoyle, in particular, obtained very important results in this direction. During the 1950s Hoyle, William Fowler (1911–1995), Margaret Burbidge (1919–), and Geoffrey Burbidge (1925–) successfully developed a theory to explain the synthesis of elements in stars and supernovae. Similar results were obtained independently at this time by Alastair Cameron (1925–2005), a physicist at a nuclear facility in Canada. According to the resulting theory of stellar evolution, which is now widely accepted, supernovae scatter the heavier elements throughout space, and it is from this debris and existing interstellar matter that a later generation of stars is born. It is believed that the Sun and its planetary retina were born of such a process.

BAADE AND STELLAR POPULATIONS

Walter Baade was an astronomer at the Bergedorf Observatory near Hamburg, where he worked on the photographic study of globular clusters and spiral nebulae. In 1931 he joined the staff at the Mount Wilson Observatory. During the 1930s he carried out work on photographic photometry and published some important results on supernovae. With the outbreak of World

War II Baade was registered as an enemy alien and prohibited from any military duty or research. As other astronomers, including Hubble, entered war service, Baade gained unprecedented access to the Hooker reflector at Mount Wilson. The brownouts of Los Angeles and the San Gabriel Valley resulting from fear of a Japanese attack meant that the night skies were very dark, producing very good observing conditions. Between 1942 and 1944 Baade's research and observations led him to identify two stellar populations. Elliptical galaxies, globular clusters, and the centers of spirals contain population II stars, which possess large velocities and whose brightest members are moderately luminous yellow giants. The Sun, its neighbors, and, more generally, the stars in the spiral arms of galaxies are "ordinary" population I stars. Baade used the Hertzsprung-Russell diagram to analyze the two populations. The main sequence of the diagram and the red giants consist of population I stars, while the population II stars tend to congregate near the lower end of the main sequence. It was subsequently established that population I stars are younger and richer in heavier elements. Baade's identification of stellar populations was a fundamental contribution to the study of galactic structure, kinematics, and stellar evolution.

After the war Baade emerged as the leading researcher at Mount Wilson and one of the most prominent astronomers in the world. The work for which he is best known concerned a fundamental revision of the Cepheid-variable distance scale. The variables in the Small Magellanic Cloud studied by Leavitt were classical Cepheids. These were regarded as similar to a class of variables known as short-term Cepheids, of which the stars RR Lyrae and W Virginis were representative members. Because the short-term Cepheids were often found in star clusters, they were also known as cluster variables. It was assumed that the same period-luminosity relation held for both classical and short-term Cepheids. To calibrate the distance scale, it was necessary to have an independent determination of the distance to at least some of these variables. This could be done for the short-term RR Lyrae variables because several of these were close enough to exhibit measurable proper motions. In the method of statistical parallax the average value of the proper motions of a group of related stars is used to estimate their distance. Knowing the distance to some of these variables, the period-luminosity relation could then be applied to estimate the distances to more far-flung variables, including ones in the Magellanic Clouds and nearby galaxies.

Although Hubble had studied classical Cepheid variables in the Andromeda nebula, the somewhat dimmer, short-term variables were too faint to be visible. Following the establishment of the great 200-inch (450 centimeter) reflector at Mount Palomar in 1948, Baade continued his program of observations with this instrument. The larger Palomar telescope allowed for the detection of fainter stars, and it was determined that short-term Cepheids should be seen in M 31. Careful study at Palomar of photographic plates by Baade failed to turn up any such variables. This fact implied that the Andromeda nebula must be considerably farther than was assumed and also indicated that something was amiss in the use of Cepheid-type variables to estimate distances. Baade concluded that the classical Cepheids and the cluster or RR Lyrae variables obeyed different period-luminosity relations and that the classical Cepheids were considerably more luminous than had previously been thought. In coming to this conclusion, he also made use of his population concept, showing that the classical Cepheids were population I stars, while the RR Lyrae variables belonged to population II. In arriving at this conclusion, he was assisted by Henrietta Swope (1902–1980) and by the research results of his doctoral student Alan Sandage (1926–).

As a result of the revision of the distance scale, Baade announced in 1952 at the General Assembly of the International Astronomical Union in Rome that the distances to the galaxies should all be doubled. This had the positive effect of increasing the size of the spirals relative to the galaxy, the latter still seeming to be puzzlingly large in comparison to its neighbors. (It is noteworthy that Shapley, still a proponent of some form of the big-galaxy model, had persistently opposed the idea that there were different kinds of Cepheids with different period-luminosity relations.) In an expanding universe that was assumed to begin from a singularity the enlarged distance scale also meant an increase in the age of the universe, thus providing some relief from the age problem.

ADVENT OF RADIO ASTRONOMY

A very significant development in twentieth-century astronomy was the invention, beginning in the 1930s, of radio telescopes, which permitted the detection of low-frequency radiation from celestial sources. Radio astronomy did not originate as a concerted program by astronomers but rather emerged by chance in the course of attempts by electrical engineers to identify sources of noise in radio communication. Karl Jansky (1905–1950) was an engineer at Bell Telephone Laboratories in the 1930s, working at a facility in New Jersey on the problem of interference in trans-Atlantic telephone communication. Using a rotating radio receiver, he detected in 1932 “a very steady hiss type static the origin of which is not yet known” (Sullivan 1984, 12). which he was able to show was astronomical in nature and emanated from the band of the Milky Way. He published his results in a journal for radio engineers, although his findings were also reported in popular astronomical periodicals of the day.

During the years 1932–1937 Jansky worked alone on the problem of “star static.” If modern science has any heroes, Jansky’s efforts at this time cast him among them. His radio astronomical research was sometimes acknowledged by professional astronomers but failed to excite serious interest in the research community. After 1937 he returned to work on problems of terrestrial noise in radio communication. He suffered from a debilitating kidney ailment that led, in 1950, to his death.

Jansky’s pioneering astronomical efforts were continued by Grote Reber (1910–1999), another radio engineer, who, in the 1940s, used a backyard paraboloidal dish in a suburb of Chicago to create the first map of celestial radio sources. It would later be established that the radiation detected by Jansky and Greber emanated from the Milky Way galaxy and was the result of a blending together of a large number of sources into what is known as synchroton radiation. The latter is emitted by particles moving in very strong magnetic fields and is typically associated with the remnants of supernovae.

The event that led to radio astronomy on a large, organized scale was the intensive development of radio and electronic technology during World War II. After the war, many of the scientists who had been involved in military projects retooled their radar equipment and receivers and began to carry out research in radio astronomy. Pioneers were Stanley Hey (1909–2000) and his colleagues in Britain's Army Operational Research Group, Bernard Lovell (1913–) at Jodrell Bank, Martin Ryle (1918–1984) and Graham Smith (1923–) at Cambridge's Cavendish Laboratories, John Bolton (1922–1993), Gordon Stanley (1921–2001), and Bernard Mills (1920–) in Sydney, Australia, and Harold Ewen (1922–) at Harvard. Several areas of investigation emerged. One involved the analysis of solar radiation and the investigation of radio waves emitted by the solar corona. Another was initiated by work in 1944 of the Dutch theorist Hendrick van de Hulst (1918–2000), who predicted that neutral hydrogen atoms in space should emit radiation at the 21-centimeter wavelength. In 1951 this radiation was detected by Ewen and the Harvard researchers. Jan Oort in Holland established a program of research in the 1950s that was successful in using the 21-centimeter band to map out the arms of the Milky Way galaxy.

A third area of research focused on objects with very small angular diameters that were strong emitters of radio waves. The first of these powerful discrete sources was identified by Hey and his collaborators in 1946 in the constellation Cygnus and designated as Cygnus A. Another such source, Cassiopeia A, was discovered in 1948 by Ryle and Smith. In the early 1950s Baade and Rudolph Minkowski (1895–1976), working with the Palomar telescope, established that Cassiopeia A was a galactic nebula (an object within the Milky Way galaxy) with unusual filamentary structure. It would later be determined that it was the remnant of a supernova some 11,000 light-years from the Sun. Cygnus A was found to be a 17th-magnitude galaxy with a substantial red shift, indicating that it was a very distant and very energetic source of radio waves. It was the first of the "radio galaxies" to be discovered. The collaboration between optical and radio astronomers would prove to be very fruitful—among the immense number of nebular objects, the radio data enabled the observer to identify particular ones for detailed optical investigation.

As the resolution of radio receivers improved, astronomers began to detect many more very localized or discrete sources of emission. A project to compile a systematic catalog of discrete radio sources was established in the 1950s at Cambridge University under the direction of Ryle. From 1950 to 1955 the Cambridge group carried out several detailed surveys. Many radio sources are known by their designation in the Cambridge catalogs; for example, 3C 273 is the 273rd object in the third Cambridge survey. The Cambridge group also pioneered methods of interferometry, in which the same source is observed by two radio telescopes. The two signals are relayed to a receiver, and the interference between the two enables one to determine the position of the source with an accuracy that is proportional to the separation between the telescopes.

The examination of the optical counterparts of discrete radio sources had revealed that many of them were distant galaxies. Ryle came to believe that the majority of these sources were extragalactic. He became interested in the cosmological implications of radio astronomy and carried out counts of discrete radio sources with distance. In 1955 he announced that his results indicated a statistically anomalous increase of faint sources with distance and therefore with earlier time, a crucial piece of evidence against the steady state theory, which required uniformity in both space and time. Ryle's claims were controversial and were criticized both by Australian researchers in radio astronomy and by the founders of the steady state theory itself. Nevertheless, as Ryle himself observed, his research seemed to show that it was possible, in principle, to distinguish empirically among the competing predictions of the different world models, an exciting fact in itself. Ryle's contributions to science were recognized in 1974, when he and Anthony (1924–) became the first astronomers to receive the Nobel Prize.

With further advances in interferometry the resolution of radio receivers improved. By the early 1960s, fairly accurate coordinates for a large number of discrete sources were available. Examined in the great California reflectors, some of these objects appeared to be star-like, with extremely unusual spectra. Two examples were the sources 3C 48 and 3C 273. Jesse Greenstein (1909–2002) and Maarten Schmidt (1929–) were astronomers at Caltech involved in the analysis of their spectra. In 1963 Schmidt realized that the unusual character of 3C 273's spectrum was a result of the fact that its hydrogen emission lines were shifted by an extremely large amount to the red; the red shift was so large that the spectrum had appeared unrecognizable. The huge red shift implied that it must be extremely distant in space, a very compact and incredibly powerful source of energy. A similar conclusion followed for 3C 48. These objects became known as quasars, short for quasi-stellar radio sources. The name proved to be somewhat misleading since it was soon found that many of the star-like sources with large red shifts were radio-silent. Nevertheless, the name stuck, and quasar astronomy developed into an important field of research.

The discovery of quasars seemed to provide evidence for the big bang theory since it apparently showed that the more distant universe was different from the nearer universe, as one would expect in an evolving cosmology. Astronomers hypothesized that quasars were the active centers of galaxies, possibly associated with the collision of two galaxies. Because the earlier universe was denser and more crowded, such collisions would have been more frequent. These speculations did not impress opponents of the big bang theory, who reasoned that as one looked out into the distant universe, it was natural to encounter diversity; the identification of unusual objects was to be expected. Supporters of the steady state theory also questioned the "cosmological" interpretation of the quasar red shifts as arising from the expansion of the universe described by Hubble's law. They suggested that they may result instead from objects thrust out with great velocity from relatively nearby galactic cores. Although the discovery of quasars was an important event, the debate in cosmology continued, and no consensus was forthcoming.

The event that clinched victory for the big bang theory in the minds of most astronomers was the detection in 1965 of the microwave background radiation. Like many of the major discoveries of twentieth-century astronomy, this event occurred more or less by accident in the course of a project devoted to another purpose. Arno Penzias (1933–) and Robert Wilson (1936–) were working in the early 1960s at Bell Laboratories in Holmden, New Jersey, on the problem of satellite communication. Penzias had a doctorate in physics from Columbia, and Wilson had a doctorate in astronomy from Caltech. They were granted permission by Bell to devote some of their time to astronomical research. They worked with a horn-shaped receiver that had been made surplus following Bell's termination of its involvement with the Echo-satellite communications project (see figure 9.1). They set about preparing the instrument for a project to study sources of microwave emission in the Milky Way galaxy. The intensity of radiation picked up by a radio receiver at a given wavelength is typically measured in terms of the temperature of a blackbody that emits the radiation at this wavelength. A blackbody is an idealized body that absorbs all radiation that falls on it. The radiation emitted by a blackbody depends only on its temperature and

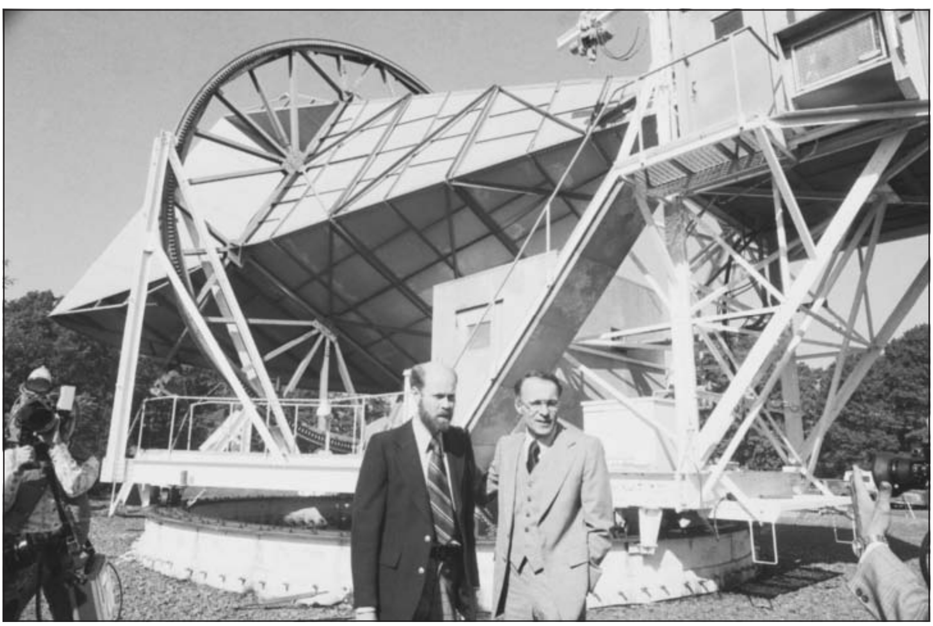

Figure 9.1: Wilson and and the Holmden microwave receiver. Bettmann/ CORBIS.

is characterized by a graph relating wavelength and intensity. Penzias and Wilson were interested in radiation of very low intensity. They fitted the horn receiver with a liquid helium load that could be used as a comparison to accurately measure a low-noise signal coming into the horn.

Penzias and Wilson possessed an instrument of unprecedented sensitivity, capable of making accurate measurements of very weak radiation. They picked up a steady three-degree-Kelvin noise in the microwave band that seemed to emanate from all parts of the sky. In order to make refined observations of galactic sources, it was first necessary to identify the source of this radiation. Despite repeated attempts over a one-year period, they were unable to trace it to any of the likely sources—nearby New York City, contamination on the surface of the receiver, or even radiation from within the galaxy. At the same time they were working on this problem a group of astronomers at Princeton University, under the direction of Robert Dicke (1916–1997), was investigating models of the early universe. A former student of Dicke's, James Peebles (1935–), had discussed the "cosmic electromagnetic radiation" associated with the early universe in a paper delivered at The Johns Hopkins University early in 1965. In effect, Dicke and Peebles were duplicating the research of Gamow, Herman, and Alpher from over 15 years earlier, which had been largely forgotten. Through their contacts in the astronomical community Penzias and Wilson became aware of the work of Dicke, and a meeting was arranged between the Bell scientists and the Princeton group. Dicke realized that the three-degree excess noise in the Holmden horn receiver was consistent with the radiation that would have been emitted following the big bang. The radiation appeared to be a fossil relic left over from the initial cataclysm that created the world.

The discovery of the microwave background radiation turned out to be a turning point in the history of cosmology, comparable to Hubble's 1929 discovery of the red shift relation. It provided concrete physical evidence for the big bang theory. There was no immediate explanation for its existence in steady state or other alternative cosmologies, and the majority of the scientific community was won over to the big bang idea. Following 1965, cosmology began to be taken much more seriously, both scientifically and institutionally. High-energy physicists became interested in the subject, and graduate courses in it became a regular part of astronomy programs in universities. Financial support for research in extragalactic radio and optical astronomy increased. In 1978 Penzias and Wilson were awarded the Nobel Prize in physics for their discovery.

The standard cosmological model accepted today by the astronomical community is the big bang theory. Further evidence for this theory emerged in the 1960s and 1970s from estimates of the frequency of helium in the universe. Surveys of near and distant regions indicate a constant ratio of helium to hydrogen: for every 10 atoms of hydrogen, there is 1 atom of helium. It is believed that this amount of helium could not have been produced in stars and that some of it must have resulted from fusion processes in the primordial conditions of high temperature and density following the initial bang. The frequency of isotopes of such light elements as hydrogen and lithium also appears to point to a prestellar origin in the big bang.

CONCLUSION

Rational cosmology up to the seventeenth century presented a coherent picture of the workings of the planetary system without any concern for the origins and development in time of this system. For Aristotle the heavens were unchanging and incorruptible, possessing no origin and experiencing no evolution. Change of place was the only alternation that could be ascribed to the planets. In the Middle Ages this tenet of Aristotelian philosophy led to some conflict with Christian theology since Christians believed that the world had been created by God and that only God was eternal and changeless. Although theologically unacceptable, the Aristotelian conception largely prevailed in scientific astronomy and was a prominent feature of the Copernican-Newtonian world picture.

In the eighteenth century the nebular hypothesis was advanced by Kant and Laplace to explain the genesis of the solar system. Ideas of origin and evolution became things of increasing scientific concern and achieved a dominant place in such sciences as biology and geology. However, as the study of the universe as a whole shifted its focus from the solar system to the stars and nebulae, these ideas receded into the background; there simply was not enough information about the large-scale nature of the universe, and speculations about its origin were hazy at best. All of this changed with the discovery of universal expansion and the microwave background radiation. The traditionally distinct subjects of cosmology and cosmogony, of the nature and structure of the universe on the one hand and the origin and development of the universe on the other, were shown to be coextensive. For the first time in human history it became possible to move beyond the psychological formulations of religious doctrine and make meaningful statements about the creation and evolution of the whole universe.

10

THE BIG BANG UNIVERSE: FROM 1965 TO THE TWENTY-FIRST CENTURY

TECHNOLOGICAL ADVANCES

Richard Hirsch (1983, 9) begins his history of X-ray astronomy with the statement, "X-ray astronomy is a gift of technology." It would not be an exaggeration to broaden this statement to include all of modern cosmology. The rapid and exciting development of this subject in the past century has been the direct result of advances in technology and engineering. In the past 40 years, sophisticated Earth-based and satellite instrumentation as well as computer simulation and analysis of data have led to unprecedented opportunities for both galactic and extragalactic research.

In the 1970s, charge-coupled devices replaced photographic plates in telescopes, resulting in a great increase in the sensitivity of imaging. A light-sensitive chip stores incoming light from a source as an electrical charge. The chip consists of an array of elements or pixels, on which the image is recorded and relayed to a computer screen. CCD technology is the basis of today's commercial digital cameras and was pioneered in astronomical telescopes.

CCDs are used in all types of telescope today, from the amateur's backyard instrument to the professional observatory. Adaptive optics, a system for canceling the disturbing effect of the atmosphere within a telescope, is a more specialized technology that has been developed for professional-level telescopes and has dramatically enhanced the resolution of images. These advances have been introduced into a new generation of gigantic telescopes situated high on mountaintops in Arizona, Hawaii, Chile, and elsewhere. Notable facilities are the Keck Observatory and Japanese Subaru Telescope in Hawaii, the European Very Large Telescope in Chile, and the Gemini Observatory, which operates telescopes in both Hawaii and Chile. Special telescopes have also been developed to carry out surveys. The Sloan Digital Sky Survey is being conducted under the auspices of an international consortium of astronomers. It employs a specially built 2.5-meter reflector in New Mexico and is producing a map of all celestial sources in a field covering one quarter of the celestial sphere.

Headquartered in New Mexico, the Very Long Baseline Array was built at a cost of $85 million and became operational in 1993. It consists of a system of 10 25-meter radio telescopes located across the continental United States, Hawaii, and the Virgin Islands. The VLBA has enabled researchers to construct radio maps of unprecedented resolution, up to 1000th of a second of arc. It has assisted in the study of very distant quasars and galaxies, yielding images of these objects at very low frequencies and enabling astronomers to determine their motions and masses.

Instruments attached to high-altitude balloons, airplanes, rockets, and orbiting satellites have enabled detailed observations in the infrared, ultraviolet, and X-ray bands not possible from the Earth's surface and have also yielded enhanced resolution of images in the optical range. The following summary of some of these projects is necessarily selective and focuses on developments (most of them ongoing) that are of particular interest for cosmology.

The Hubble Space Telescope, launched by NASA in 1991, is the most expensive scientific instrument in the history of astronomy (figure 10.1). Equipped with a 2.4-meter Casssegrain reflector, the telescope transmits signals to satellites, which then send them on to the Space Telescope Science Institute in Baltimore. Following some repair and corrections to the primary instrument carried out by shuttle missions, the Hubble Telescope has realized its potential, relaying to the Earth a succession of stunning images of the cosmos. From the viewpoint of cosmology the Hubble Telescope has provided data that have allowed a more accurate determination of the value of Hubble's constant and

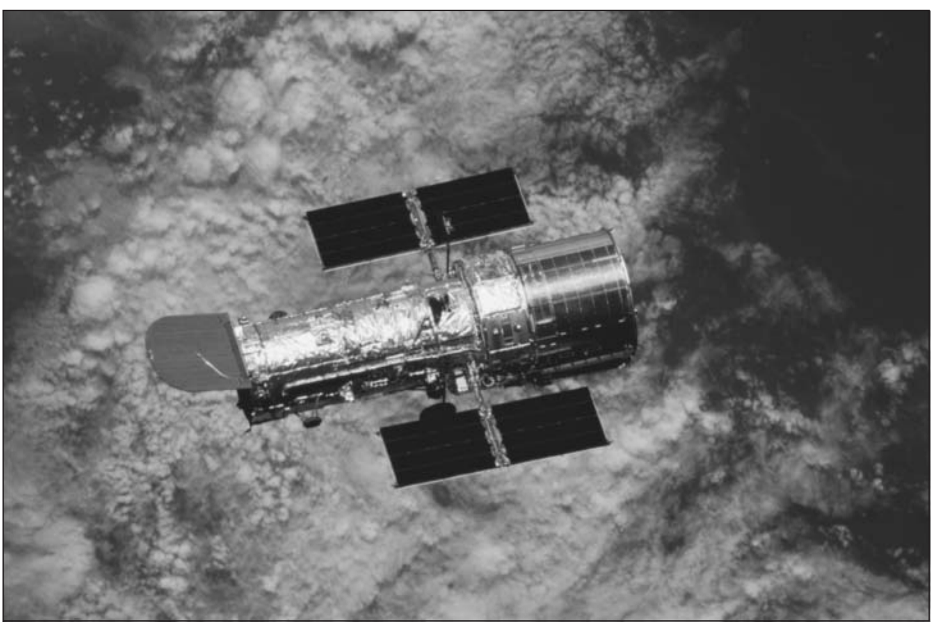

Figure 10.1: The Hubble Space Telescope. NASA.

was also used in the historic research of the late 1990s leading to the discovery of universal acceleration (discussed below). The Hubble has completed two deep-field pictures of small star fields in the northern and southern celestial spheres, obtained through exposures taken over 10 days. The deep fields have provided a glimpse into the formation of galaxies in the early history of the universe and provided observational material for further investigation by advanced Earth-based telescopes and X-ray satellite observatories. Another deep-field study has been conducted by the Suburu Telescope, which, in 2004, identified the farthest cluster of galaxies yet known.

Other major orbiting observatories put into orbit by NASA include the Compton Gamma Ray Observatory, launched in 1991, the Chandra X-Ray Observatory, launched in 1999, and the Space Infrared Telescope facility, launched in 2003. Each of these satellites is returning important data for cosmological research.

In addition to the great orbiting observatories, there have been several special-purpose satellites, whose primary missions involved cosmological research or have had important implications for questions of interest to cosmology. Launched into Earth orbit in 1989, the Cosmic Background Explorer (COBE) was the product of a two-decade effort to study the microwave background radiation. Since the discovery of this radiation in 1965, there was a concerted attempt to detect small variations—what are called anisotropies—in the intensity of the radiation in different parts of the sky. A completely uniform field of radiation would raise questions about how discrete structures could have ever formed in the universe and would raise fundamental problems for the big bang theory. In the 1970s, astronomers had detected a substantial variation, or dipole, in the radiation resulting from the motion of the solar system relative to the rest of the universe. The COBE mission grew out of these efforts and was directed from the Goddard Space Flight Center in Maryland. It involved the collective efforts of engineers, physicists, astronomers, and computer analysts. COBE measured the strength of the microwave background radiation at different frequencies and also attempted to detect anisotropies using its differential microwave radiometer. The satellite returned data for four years. A major result of COBE was to verify the blackbody nature of the microwave background radiation and thereby to provide evidence for its probable origin in the big bang. On the basis of two years of observations and extensive computer analysis of the data, researchers also concluded that there were definite, if extremely small, variations in the intensity of the radiation. Commenting on the significance of this last finding, the lead astronomer, George Smoot (1945–), said, "If you're religious, it's like seeing God" (1993, 289). Despite their somewhat tentative character, the COBE results became a major science news story, attracting intense public interest and being widely viewed as dramatic confirmation of the big bang theory.

A defect of COBE was the very crude resolution of its map of the background radiation. Its work has been continued by high-altitude balloon experiments in Antarctica, the largest of which is BOOMERANG, or Balloon Observations of Millimetric Extragalactic Radiation and Geophysics. Microwave dish detectors have also been installed in Antarctica to study the background radiation. The instruments used in these experiment operate best in the very cold conditions around the south pole. Satellite research has also continued with the launch in 2001 of the Wilkinson Microwave Anisotropy Probe, or WMAP. All of these instruments have confirmed the existence of temperature differences found by COBE and helped to produce a detailed map of the background radiation. Their findings have been interpreted by theorists in terms of cosmological models involving such concepts as inflation, the decoupling of matter from radiation, and the formation of discrete structures. The background radiation is believed to have originated in primordial plasma of the early universe during an initial period of thermalization. The work of the microwave detectors has been complemented by the Sloan Digital Survey. In principle, one should be able to compare the map of the microwave background with a chart of galaxy distribution and gain valuable information on how the universe developed.

The COBE and WMAP projects are representative of major research endeavors in modern cosmology. They are technologically advanced and very expensive, science on a large scale, involving many specialists from different fields—engineering, astronomy, and data analysis. A different sort of scientific venture with many of these characteristics was the European astrometry project High Precision Parallax Collecting Satellite, known as Hipparcos. (The name is similar to Hipparchus, the ancient astronomer who compiled a star catalog around 150 B.C.) This satellite operated from 1989 to 1993 and carried out measurements of unprecedented accuracy of the position and color of over 100,000 stars. All distance methods are piggybacked on one another—statistical parallaxes rely on base lines obtained from trigonometric parallaxes, Cepheid-variable studies are calibrated using parallax methods, while estimates involving red shifts and supernovae work from Cepheid determinations. Estimates of distances to the farthest objects in the universe depend ultimately on trigonometric parallaxes of stars in the Sun's neighborhood, and it was these that Hipparcos measured so accurately. The satellite obtained distances out to 200–300 light-years, three times as far as traditional astrometry had reached, and substantially increased the accuracy of existing stellar parallax measurements. The recalibrated Cepheid distance scale implied by Hipparcos involved an increase in this distance scale of about 10 percent, an important finding for cosmology and one that is currently the subject of further investigation.

The dominance of technology and large collaborative projects has been a characteristic of cosmological research since the beginning of the modern revolution in the early twentieth century. For every research venture mentioned above, there are plans for larger and more ambitious undertakings in the next few decades. The twenty-first century will be a period of investigation on a multitude of fronts, with a diversity of instruments and methods of unprecedented technological and engineering sophistication. If history is any guide, we should expect unforeseen findings and new theoretical developments to modify our picture of the universe in fundamental ways.

DISTRIBUTION OF GALAXIES

In the 1970s and 1980s, scientists at the Harvard-Smithsonian Center for Astrophysics measured the distances to over 30,000 galaxies in selected sectors of the sky. Distances were calculated from red shift data using Hubble's relation. The plan was to construct a three-dimensional map of the universe out to several hundred light-years. Leadership in this venture was provided first by Marc Davis (1947–) and later by John Huchra (1948–), Margaret Geller (1947–), and Valerie de Lapparent. The CFA surveys came as a major surprise. Instead of being distributed more or less uniformly in space, galaxies lie along long sheets and walls that surround large voids. The universe possesses a soap-bubble structure characterized by considerable local unevenness in the distribution of galaxies. Beginning with the uniform conditions indicated by the cosmic background radiation, the universe has evolved into a rather lumpy place—the task of explaining this fact in terms of relativistic models of galaxy formation in an expanding universe has not proved an altogether easy one.

During the 1980s a group of seven astronomers embarked on an international collaborative project to investigate a particular class of galaxies, the elliptical galaxies. The latter are ones that lack spiral arms, are free of gas, and are predominantly oval or round in shape. The astronomers, led by Sandra Faber (1944–) of the Lick Observatory and Trevor Lynden-Bell (1935–) of Cambridge University, were interested in the problem of galactic evolution and concentrated on the ellipticals because they are, as a class, quite uniform in their properties. To map the galaxies, they measured their red shifts and used Hubble's relation to estimate their distances. In the course of their investigation they derived an indicator that appeared to correlate very well with absolute luminosity. The indicator was the velocity dispersion of the galaxy, a quantity that measures the spread of the velocities of stars in the galaxy. Its correlation with luminosity allowed a measure of distance that could be used independently of the Hubble relation. Much to their surprise, the astronomers discovered that the galaxies they were studying possessed large and systematic "peculiar motions," velocities independent of universal expansion arising from the gravitational attraction of neighboring galaxies and matter.

The Milky Way galaxy is part of a larger collection of galaxies known as the Local Group. The Local Group in turn belongs to a system of clusters known as the Local Supercluster. The data of Faber, Lynden-Bell, and their associates indicated that the Local Group and the Supercluster as well as several other clusters are streaming toward a more distant concentration of mass. Dubbed the "Great Attractor" in 1986 by Alan Dressler (1948–), one of the investigators in the project, this mass consists of a large swell in the density of matter arising from a concentration of galaxies and "dark matter" in the direction of the southern constellations Hydra and Centaurus. The existence of the Great Attractor provided further evidence of local inhomogeneity in the distribution of matter in the universe.

GRAVITATIONAL LENS AND GRAVITATIONAL WAVES

Gravitational Lens

Both Newtonian and Einsteinian theories of gravity predict that light will be bent in the presence of strong gravitational fields. Einstein's original prediction of such an effect and the value he obtained for its magnitude were confirmed in the 1960s by experiments in which radar signals were bounced off the planets. When a signal is reflected off Venus or Mercury while this planet is located on the other side of the Sun along the line of sight from the Earth to the Sun, the trajectory of the radar signal will be deflected by the Sun's gravitational field. This deflection is indicated experimentally by a slight change in the time it takes for the radar beam to reach the planet and return to the radio antenna on Earth. The results obtained have served to verify the general theory of relativity and enabled researchers to refine and calibrate its predictions.

As early as the eighteenth century, scientists hypothesized that the light from a distant star may be perturbed by the gravitational action of a closer star lying along the line of sight from the Earth to the distant star. With the advent of the general theory of relativity, there was renewed interest in the possible existence of such a phenomenon. In 1919 the British physicist Oliver Lodge (1851–1940) introduced the term "gravitational lens" to denote the closer star that acts gravitationally on the light from the more distant star. In 1924 Einstein calculated the magnitude of such an effect and concluded that it was too small to be observable. In the 1930s Fritz Zwicky proposed that gravitational lensing would be more likely to be observed in the case of extragalactic nebulae, what are today called galaxies. Because galaxies are very distant, there is a greater probability that an intervening massive body will lie somewhere along the long line of sight from us to any given galaxy. Furthermore, the large mass of an intervening galaxy or cluster of galaxies makes it a likely candidate to act as a lens.

In the 1960s, gravitational lens again became a subject of active theoretical interest, and several papers were published analyzing the optical properties of lens systems and producing calculations to measure their effects. Despite this interest, no systematic observational program emerged to detect such phenomena. In the 1970s a group of researchers led by Dennis Walsh (1933-2005) was attempting to correlate radio sources with objects observed through large optical telescopes. Walsh was working at the Jodrell Bank radio telescope in Britain and prepared a catalog of radio sources. One of these was 0957 + 561, so designated because it was located in Ursa Major at right ascension 9 hours and 57 minutes and declination 56 degrees, one minute. A preliminary survey placed this source very close to a blue double-stellar object of the 17th magnitude; its blueness made it a likely candidate to be a quasar. In 1979 Walsh and his collaborators, Robert Carswell and Raymond Weyman (1934–), examined the object with the two large reflecting telescopes at Kitt Peak Observatory in Arizona. To their surprise, they found that it was a double quasar with identical red shifts and similar emission and absorption spectra. They concluded that the two quasars were in fact one, its light being gravitationally lensed and split by an intermediate object, later to be identified as an elliptical galaxy (possibly part of a cluster of galaxies) located very close in position to one of the quasars. The light from the quasar passes close to the intermediate galaxy, which lies on the line of sight from the Earth to the quasar. The galaxy acts as a gravitational lens, producing the two images of the quasar. The quasar possesses a red shift $z = 1.3$, indicating that it is at a distance of about nine billion light-years.

It is worth noting that neither Walsh nor his collaborators had any prior involvement with the subject of lensing and that the 1979 discovery was essentially a serendipitous event. The realization that 0957 + 561 is a gravitational lensing system created a sensation and led to efforts to discover further such systems. Gravitational lensing became a very active topic in extragalactic astrophysics, a subject of investigation by the most advanced instruments of deep-space astronomy. By 2005, more than one 100 gravitational lensed quasars and galaxies had been detected. Pictures of lensed quasars and galaxies are among the most popular photographs released by the education office of the Hubble Space Telescope. A gravitational lens can act on the image of a distant object in various ways. The image may be distorted, bent into a curved form, or made to consist of multiple component images. Figure 10.2 depicts a distant quasar in the constellation Pegasus that has been lensed into four images by a galaxy that is relatively close to us. In recognition of Einstein's contributions to the theory of gravity, it is called "Einstein's Cross."

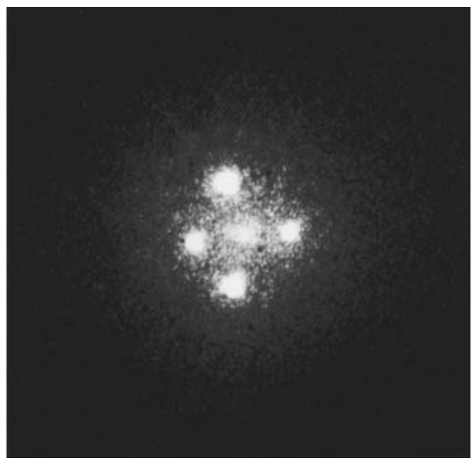

Figure 10.2: Einstein's Cross, a gravitationally lensed quasar. NASA.

It is realized today that gravitational lens systems are not restricted to objects outside the galaxy, and there has been much interest in so-called minilensing events involving stars within the Milky Way system or the nearby Magellanic Clouds. However, from the viewpoint of cosmology it is distant extragalactic lenses that are of the most interest, and they have become an important tool of investigation in modern cosmology. If a lensed quasar undergoes a change in brightness, this change will be relayed to observers at slightly different times in the two images, reflecting the slightly different distances the light has to follow along the two optical trajectories. This difference and an analysis of the geometry of the lensing system allows one to determine the distance to the quasar more precisely and to obtain a more accurate value for Hubble's constant. Gravitational lenses also magnify the image of an object, opening up to investigation distant quasars and galaxies that would otherwise be unobservable. Indeed, the most distant objects in the universe, small galaxies at a distance of 13 billion light-years, have been sighted as part of lens systems, which magnify their brightness by as much as a factor of 100.

GRAVITATIONAL WAVES

When a gravitational source undergoes a change of motion, this change will result in a disturbance in its gravitational field that is propagated through space. The effect will normally be a very small one, not detectible by even the most sensitive instruments. One of the discoveries of modern astrophysics has been the occurrence of extreme events: a supernova explosion, the rapid implosion of matter into a black hole, binary stars spiraling into each other with huge rotational velocities, quasars involving the collision of galactic nuclei, and so on. It is thought that such events should be powerful enough to emit sizeable gravitational waves, and there is currently a concerted effort to detect them using specially built instruments.

Gravitational waves are analyzed today from the viewpoint of the general theory of relativity, the dominant theory of gravity employed in astrophysics. During the 1960s an American physicist named Joseph Weber (1919–2000) constructed instruments to detect gravitational waves. Despite determined efforts over many years, he was unable to convince the scientific community that he had found anything. The first actual empirical evidence for the existence of gravitational waves emerged by accident within astrophysics in the course of an investigation of a new type of body called a pulsar. In 1967 S. Jocelyn Bell (1943–), a graduate student working under Anthony Hewish (1924–) at Cambridge University, detected a celestial source emitting a rapid series of radio pulses at extremely regular intervals. Other such pulsating sources were soon found. Although pulsars were initially seen as an enigma, Thomas Gold arrived in 1968 at the explanation that is now generally accepted. Massive stars are believed to end their lives as very compact and dense objects, so dense that the protons and electrons are fused together as neutrons. Neutron stars possess extremely powerful magnetic fields, which result in the emission of radio waves from the ends of a magnetic axis through the star that is inclined to the star's axis of rotation. The star rotates very rapidly, and as it does so, the beam of radiation periodically crosses the observer's line of sight, resulting in the detection of a regular sequence of radio pulses.

During the 1970s University of Massachusetts astronomer Joseph Taylor (1941–) and his graduate assistant Russell Hulse (1950–) embarked on a search for pulsars at the Arecibo radio astronomical facility in Puerto Rico. The search involved the use of a minicomputer that was programmed to pick out pulsar signals as the large dish at Arecibo scanned the sky. In 1975 Hulse found a source, designated PSR 1913+16, which proved to be a pulsar paired with a companion star at a distance of 14,000 light-years. The pulsar and its companion revolve about each other with a velocity of over 300 kilometers per second. From observation over a period of several years it was found that the binary system was losing rotational energy and that the two stars were spiraling toward each other. A calculation by Taylor based on the general theory of relativity revealed that the energy lost was exactly equal to the energy that would arise from the dissipation of gravitational waves from the system. PSR 1913+16 was the first example ever in which there was real evidence of gravitational waves, and the work of Taylor and Hulse was heralded as a major breakthrough. In 1993 the two researchers were awarded the Nobel Prize for Physics for their research. In his acceptance speech Hulse called attention to the serendipity of their discovery: it occurred not as a result of a search for gravitational waves but as a result of a program within radio astronomy to identify pulsars.

Following the discovery of the Hulse-Taylor pulsar, there were renewed efforts to design and build instruments that could detect gravitational waves directly. The largest facility built for this purpose was LIGO run from Caltech and MIT. LIGO, which stands for Laser Interferometer Gravitational-Wave Observatory, came on-line in 2003 when it joined several other large American and international observatories already in operation. The direct discovery of gravitational waves, should it occur, would be a milestone in the history of astrophysics and theories of gravity and would have fundamental implications for cosmology. The general theory of relativity would be verified and subject to analysis by the procedures and methods of the traditional astronomical observatory. Unlike electromagnetic radiation, gravitational waves are not impeded by intermediate dust and stars that may lie along the line of sight from the observer to the emitting object. The reception of these waves would provide an unprecedented view of distant objects and could yield fundamental information about the gravitational interaction of the early universe.

BLACK HOLES

The term "black hole" was introduced by American theorist John Archibald Wheeler (1911–) in 1968 and refers to an object that is so dense and massive that no electromagnetic radiation can escape from its gravitational field. Although the idea of a black hole goes back to the eighteenth century, the quantitative aspects of such objects were only analyzed in detail using the general theory of relativity. Black holes may be rotating or nonrotating, charged or uncharged, and may even interact thermodynamically with their environment, losing energy. The key variable associated with a black hole is its Schwarzschild radius, named after the general relativist Karl Schwarzschild (1873–1916), who theorized about gravitationally compact objects in 1916. Assume all the mass of an object is contained within a certain radius. If this radius is less than or equal to the Schwarzschild radius then no radiation can escape from the object to the outside. The surface defined by the limiting radius is called the event horizon. All events within the event horizon remain confined to the black hole, locked inside and forever hidden from the outside universe. The Schwarzschild radius is a function of the mass of the black hole;

for an object the mass of the Sun it is about three kilometers; for an object three million times the mass of the Sun it is around 300 astronomical units. A star that collapses and becomes a black hole will interact gravitationally with the rest of the universe but will interact in no other way. Cosmologist Joseph Silk (1980, 264) has likened this situation to the grin of the Cheshire cat: only the gravitational field is left behind.

Theories of stellar evolution have determined that the final stage in the life of a star is dependent on its mass. A star like the Sun will end its life as a white dwarf, while a star much more massive than the sun will end its life as a supernova, and the central part of the supernova will collapse to form a neutron star. If the core is massive enough (more than about two times the mass of the Sun), it will implode and become a black hole. The most studied candidate for a stellar-sized black hole is the X-ray source Cygnus X-1, first detected by satellite in 1965 and investigated extensively in the 1970s and 1980s. Cygnus X-1 was found to be a binary system consisting of a red giant and an invisible companion. From a study of the small perturbations in the position of the visible star it was inferred that the mass of the companion was over five times the mass of the Sun, implying that it must be a black hole. The X rays are emitted, as matter from the visible star is drawn with ever-increasing velocity into the black hole companion, releasing large amounts of energy as it approaches the event horizon. Several other binary star systems similar to Cygnus X-1 have been found. While not proven with certainty, it is believed that each involves a black hole companion star.

The highly speculative discussion around black holes at times has made them seem almost like things from science fiction. What has emerged as observationally sound and theoretically useful is the concept of a very massive, or supermassive, black hole. The current model of quasars and certain types of galaxies conceives of them as a massive central black hole, around which an accretion disk has formed, consisting of matter falling into the central hole and emitting tremendous amounts of energy in the process. Although active galactic nuclei are the most dramatic evidence of the existence of supermassive black holes, it is now believed that such objects also lie at the centers of most galaxies. Indeed, detailed observation of the Milky Way galaxy has pointed to the presence of a very massive black hole at its center. This center is located within the band of the Milky Way at a point known as Sagittarius A, a fairly weak radio source in the constellation of Sagittarius. The center possesses a mass of about 2.7 million solar masses and radiates X rays, the signature characteristic of a black hole. The study of X-ray emissions was greatly aided by the launch of the Chandra orbiting X-ray telescope, and observations with this instrument have contributed greatly to imaging of the galactic nucleus. Infrared observations of stars very close to the galactic center have enabled astronomers to calculate their orbits and periods, and these values have verified the value for the mass of the black hole at this center. The Schwarzschild radius of this hole is about 300 astronomical units.

The concept of a black hole provides an outstanding example of how a notion rooted in theory can come to play an important role in a science largely governed by observation. Black holes as theoretical entities will be an essential part of any detailed explanation of the universe as a whole, both in terms of its current structure and in terms of what we can conclude about its origin and evolution in time. The finding that galactic nuclei consist of black holes has contributed to a theory of the formation and evolution of galaxies. During the early history of a galaxy, very massive stars tend to congregate in its center, leading to the formation of a supermassive black hole and resulting in an object we observe as a quasar, or a Seyfert galaxy. As the galaxy evolves, the accretion disk dissipates, and the energy emitted by the nucleus decreases, resulting eventually in the relatively quiet black hole nucleus of the sort found in the Milky Way system. Black holes have arisen in an entirely different context in theorizing about the first moments of the universe. It has been shown by Stephen Hawking (1942–) that miniature black holes may have developed in the early universe, only to dissipate during the ensuing expansion.

DARK MATTER

A major ongoing challenge to astronomy that has emerged in the last 50 years is the problem of dark matter. As early as the 1970s, the Mount Wilson astronomer Fritz Zwicky had observed that the rotational characteristics of galaxies in the Coma cluster implied that the galaxies must be embedded in a larger quantity of mass. Although this matter was not observable by optical means, its existence could be inferred by the type of rotational motion exhibited by the cluster. In the 1970s Vera Rubin (1928–), William Ford (1931–), and associates at the Carnegie Institution in Washington, D. C., carried out detailed spectroscopic studies of individual galaxies. They measured the angular rate of rotation as a function of distance from the center of the galaxy. They found that beyond a certain distance from the center the curve became flat, indicating that the visible parts of the galaxy were rotating as if they were embedded in a larger quantity of matter. The visible galaxy was apparently contained within a larger halo of dark matter. The discovery of the Great Attractor in the 1980s revealed a region of generalized high-mass density in intergalactic space detectable only by its gravitational attraction. Evidence on many different fronts has accumulated indicating that a very considerable percentage of the universe is present in a "dark" form, emitting no electromagnetic radiation but interacting gravitationally with visible matter.

In considering what dark matter is made of it is customary to distinguish between baryonic and nonbaryonic matter. Baryonic matter consists of ordinary protons, neutrons, and electrons and would be present in dark form in brown dwarfs, black holes, and other objects that are known to exist but emit little or no radiation. (A brown dwarf is a very small star, less than about eight percent the mass of the Sun. Such stars are too small to generate energy through nuclear reaction in their cores but emit some energy in the infrared band as the result of gravitational contraction.) Unfortunately, it is believed that such sources could provide only a small fraction of the dark matter in the universe. Candidates for nonbaryonic dark matter include something known as "weakly interacting massive particles," or WIMPS, none of which have been detected so far. Another more popular candidate is the neutrino, a particle whose existence was predicted in 1931 but was only first detected in 1956. Neutrinos are small, chargeless particles that travel at velocities close to the speed of light and are produced in nuclear reactions. When a neutron decays into a proton and electron, it emits an antineutrino. It is not known conclusively if any of the different types of neutrinos possess mass, although experiments in the late 1990s indicated that some types may have a slight mass. Because there are so many neutrinos in the universe, even a very small nonzero mass would result in a significant contribution to the total bulk of the cosmos.

ACCELERATION

For close to 70 years following the discovery of Hubble's law, astronomers made assumptions about universal expansion that seemed reasonable and conformed to how the universe could be expected to behave. The expansion of the universe that began in the big bang would decrease in strength as the gravitational attraction of matter worked over time to slow expansion. Hubble's constant would be found to be decreasing over time. This fact was built into the standard relativistic models; the R-t graphs giving the scale factor R as a function of t were all concave downward, indicating the steady slowing of the universal expansion. (The Eddington and Lemaître models from the early 1930s were never seriously pursued.) A basic constant of big bang cosmological theories was the deceleration parameter, giving the rate at which the expansion of the universe is slowing down.

In the 1990s, separate teams of astronomers at Harvard University and the University of California at Berkeley embarked on an investigation of how Hubble's red shift–distance law holds up for more distant galaxies. In order to do this, it was necessary to obtain accurate data for the distance to these galaxies that was independent of the values given by the law itself. Since the red shifts were given directly by spectroscopic measurement, the recessional velocities of the galaxies were known. It was then simply a matter of matching the red shifts to the independently determined distances in order to see how well Hubble's law fit the data. The data for this research was provided by observations of distant galaxies made with the Hubble Space Telescope.

To estimate the distance to a galaxy, it is necessary to find a method for determining the intrinsic brightness of some of its stars. For galaxies fairly close to us it has been possible to use the Cepheid-variable method, a method greatly extended by observations with the Hubble Telescope and the new generation of powerful, Earth-based telescopes. To find the distances to more distant galaxies, astronomers have used an object known as a type 1A supernova. Such a supernova consists of a binary-star system, in which one member is a white dwarf and the other is a star excreting matter to the dwarf. The mass of the receiving star increases until it reaches a certain value, at which point the star explodes as a supernova. A key characteristic of this class of supernovae is that each possesses roughly the same intrinsic magnitude and exhibits a similar light curve of brightness variation during its short career as an exploding star. Such differences in intrinsic brightness that do exist are related to the star's period of variation and the shape of its light curve—the longer the period, the brighter the supernova.

Early in 1998 the astronomers involved in the supernova research made a major announcement. Suppose one begins with the Hubble constant as determined for galaxies in the neighborhood of the Milky Way galaxy. If one then takes a more distant galaxy and puts the value of its distance as given by the type 1A supernova yardstick method into the Hubble equation, one obtains a red shift that is larger than the one that is actually observed. The galaxy is receding more slowly than it should according to the red shift law. This implies that the expansion of the universe is accelerating as one moves forward in time and closer in space to the galaxy. It would be an understatement to say that this result came as a surprise.

Astronomers devoted much effort to verifying this result and to checking whether some other effect could account for it. If in fact there were obscuring dust or matter present in space, the supernovae would be closer than they appear to be, and the supernova yardstick method would be overestimating their distances. The reduced distances would then fit properly into the Hubble law, and the conventional picture would be confirmed. However, the optical effects of obscuring matter, such as a reddening of the light over and above the red shift of expansion, have not been observed.

The discovery of acceleration ranks with the Hubble relation and the cosmic background radiation as one of the fundamental findings of modern cosmology. With good reason, *Science* magazine called the discovery the scientific story of the year. One response to the supernovae studies has been to renew the study of relativistic world models, in which the cosmological constant λ is positive. Cosmologists have speculated that space is filled with some kind of "dark energy" measured by λ that propels the expansion of the universe. This idea makes sense if one supposes that expansion is subject to the retarding force of gravity and a propelling force corresponding to the dark energy. When the universe was younger and more dense, the force of gravity was strong relative to the dark energy force; as it has expanded and become less dense, the latter force has gained strength relative to gravity, and the result is the accelerated expansion observed by astronomers. This picture is supported by the most detailed data currently available on distant supernovae, which seem to indicate that the universe was decelerating until about six billion years ago, at which time, expansion entered its current phase of acceleration.

INFLATION

Theoretical cosmology has been strongly influenced by a new conception of the very early universe advanced by MIT physicist Alan Guth (1947–) in 1982. A remarkable characteristic of the cosmic background radiation is its uniformity—points on the sky 180 degrees apart possess the same temperature to an accuracy of 1 part in 100,000. These temperatures correspond to parts of the universe that can have had no contact since the initial bang that created the world. Using general relativity, thermodynamics, and particle physics, Guth devised a theory known as inflation to explain this fact. According to this conception, at the very beginning of its history the universe underwent a phase transition resulting in a period of exponential expansion lasting only a tiny fraction of a second—in an instant the universe inflated, creating the homogeneity and isotropy we observe today in the cosmic background radiation. The rapid expansion also resulted in the disappearance of objects known as a magnetic monopoles, hypothetical particles not encountered in nature but predicted to exist in abundance by conventional big bang models.

Inflation requires the universe to be much more massive than it apparently is. A large percentage of the universe must consist of some form of nonbaryonic mass or energy. An inflationary universe is also a flat universe, one which is perfectly Euclidean and which will expand forever. The exponential expansion in the first instants of the big bang resulted in this flat geometry, in the same way that a balloon inflated to a very large size produces a surface that locally is very close to being flat.

THE NEW SYNTHESIS

In recent years, scientists have been working on a synthesis combining inflationary theory, dark matter, the dark energy supposed to exist from the observed accelerated expansion of the universe, and the pattern of anisotropies revealed in the early universe by balloon detectors and by WMAP. There is a feeling among some theorists that a coherent picture of the early and evolving universe is finally coming together.

Before 1998 it was customary to distinguish three cosmological models for an expanding universe, all based on the general theory of relativity.

1. A closed universe, in which the average density of matter is large enough so that gravity is eventually able to overcome expansion. The expansion slows down, eventually stops, and at some time, contraction begins. Relativistic theory implies that the geometry of this universe is Riemannian curved space. The universe is finite.
2. An open universe, in which the density of matter is small enough so that expansion continues forever. The geometry of this universe is hyperbolic, analogous to a saddle-shaped surface. The universe is infinite.
3. A flat universe, in which the average density of matter is the critical density. If the density were smaller than this value, the universe would be closed, and if it were greater than this value, the universe would be open. In a flat universe the expansion is steadily slowing down but never stops: it is the slowest rate of expansion with this property. The geometry of a flat universe is Euclidean, like ordinary school or surveyor's geometry. The universe is infinite.

In cosmology the symbol Ω (Omega) is used to denote the ratio of the actual density of mass in the universe to the critical density. The three possibilities outlined above correspond to the cases $\Omega > 1$, $\Omega < 1$, and $\Omega = 1$. In all three of these models it is assumed that the cosmological constant is zero. The universe was believed to be described by one of these three models, the case $\Omega = 1$ being favored by many cosmologists.

With the discovery of acceleration it has been necessary to radically revise the standard scheme. Much interest focuses on models in which the cosmological constant is positive, although other solutions have also been sought. It is now believed that the universe contains ordinary matter, dark matter, and dark energy. The introduction of dark energy implies the existence of a much larger class of cosmological models than the three given by the possibilities $\Omega > 1$, $\Omega < 1$, and $\Omega = 1$. The density Ω will be made up of a component Ω_m due to matter (ordinary and dark) and a component Ω_v due to dark energy (also known as the vacuum energy): $\Omega = \Omega_m + \Omega_v$. The nature of the cosmological model will be determined by the value of Ω as well as the relative contributions of Ω_m and Ω_v to Ω. Current observation indicates that the density Ω of the actual universe is the critical density, $\Omega = 1$. It is believed that within a small margin of error the contributions to the critical density consist of 5 percent ordinary baryonic matter, 25 percent dark matter, and the rest dark energy. Hence $\Omega_m = 0.3$ and $\Omega_v = 0.7$. In our universe it is not the case that the expansion is slowing; to the contrary, it is expanding at an ever-increasing rate.

If the critical density is equal to one, then the geometry of the universe is Euclidean. That this is indeed the case is supported by several pieces of empirical evidence. Calculations of the frequency of deuterium and measures of the masses of galaxies at different times in the history of the universe imply a flat geometry. The map of anisotropies in the cosmic background radiation has revealed that the distance between successive peaks in intensity is one degree, exactly the value theoretical calculations predict if Ω is assumed to be one and the universe is flat.

PHILOSOPHICAL QUESTIONS

Cosmology considers the problem of the universe taken as a whole and ponders questions such as the origin of the world and its evolution in time. The logical and basic nature of cosmology means that philosophical questions arise here more than they tend to do in other branches of physical science. It is common to distinguish in science between fundamental laws, such as the law of gravity or the law of valences, and laws that apply to special systems or objects, such as Kepler's law of planetary orbits or the principle of natural selection in biology. Hubble's law combines characteristics of both types of law. It concerns the particular configuration of gravitating matter as it happens to exist in nature. On the other hand, because it describes the universe as a whole, it is about everything that ever was and ever will be.

The cosmological principle provides an example of how philosophical views have influenced scientific theorizing about the universe. The adoption of the principle by Einstein and Milne was based on what appeared to be a reasonable a priori assumption about the relationship of the observer to the universe at large. It turned out that geometric world-building starting with the principle conformed with the exciting discoveries in extragalactic astronomy. By extending the principle to include time as well as space, the steady state theorists obtained the perfect cosmological principle and cited philosophical reasons in support of it. Philosophical issues have also arisen in attempts to justify the general theory of relativity as the necessary framework to describe the universe.

A different sort of question has been raised by the philosopher Ian Hacking (1989) and concerns the essentially passive character of extragalactic astronomy. Hacking has staked out a position in the philosophy of science according to which a theoretical entity in physics is said to be real if it is possible to manipulate this entity in some way. Although electrons and other subatomic particles are not directly observable in the same sense as are macroscopic entities such as colliding balls, electrical sparks, or glowing gases, they are nonetheless just as real because we are able to interfere with them: to eject them from instruments, to change their path by means of magnetic fields, or to accelerate them in particle accelerators. Not even in the boldest visions of science fiction does there seem any prospect of interacting actively with the objects located in the universe beyond our galaxy. The study of the distant universe is destined to be a passive science, and for this reason—according to Hacking—extragalactic objects are deprived of an essential characteristic of what it means for something to be real to us.

The philosophical doctrine advanced by Hacking is related to the more general point concerning the irrelevance of extragalactic astronomy to our daily lives. Concerns about the value of the new science of astrophysics were already raised in the nineteenth century by the pioneer of stellar spectroscopy, William Huggins: "The new astronomy, unlike the old astronomy to which we are indebted for skill in the navigation of the seas, the calculation of the tides, and the daily regulation of the time, can lay no claim to afford us material help in the routine of daily life" (Meadows 1984, 70). More recently, commentators such as journalist John Horgan (1996) have cast a skeptical eye at discussions of the early universe, suggesting that they are part of what he calls "ironic science," science that is subject to multiple interpretations. Entities postulated in cosmology such as monopoles or dark energy are more tentative and theoretical than researchers would sometimes lead one to believe. This conclusion is related to the contrast between the highly technical and specialized character of professional research in cosmology and the descriptive and unrigorous character of popular expositions of the subject.

There is no doubt that large parts of cosmology are bound to remain theoretical. Speculations about the inflationary character of the early universe or the creation of small black holes after the big bang will never be subject to direct confirmation. The goal of theoretical work can only be to find a picture of the primordial universe that is plausible and consistent with what is observed to have occurred from the moment that matter decoupled from radiation. It remains the case that the emergence and subsequent verification of big bang cosmology constitutes one of the most substantial and unexpected developments of modern science. We can now make scientific statements about the origin and evolution of the whole universe, something that only a century ago would have seemed inconceivable.

A consequence of the revolution of the past century has been that cosmology has become an essentially historical science. In order to understand the universe as we see it today, it is necessary from observation and theory to reconstruct one singular event, the primordial big bang 13 or 14 billion years ago, and to trace the subsequent evolution of the universe. Final questions traditionally understood to lie within the domain of religion have become part of science. It is worth noting that the father of the big bang theory, the Abbé Lemaître, became a scientific advisor to the Vatican in the later part of his career. In a 1951 address Pope Pius XII cited big bang cosmology in support of the Christian conception of a Creator and the beginning of the world in a creation event. Lemaître himself had reservations about mixing science and religion, believing that matters of religious faith depend in the final analysis on considerations whose validity is independent of the results of science. Modern cosmology has become a kind of secular theology, coming as close as rational investigation ever can to uncovering the ultimate mysteries of the universe.

TIMELINE

380 B.C.	Plato's *Timaeus*. Planetary motion occurs in uniform motion in circles.
360	Eudoxus's system of concentric spheres. Geometrical modeling of planetary motions.
350	Seleucid Babylonian astronomical tables.
340	Aristotle's *Physics*. Physical basis for a geocentric universe.
150	Hipparchus's solar theory. Geometrical modeling combined with precise quantitative data.
ca. 5 A.D.	*Zhou bi suan jing* and the Chinese umbrella cosmology.
150	Ptolemy's *Almagest* and *Planetary Hypotheses*. Mathematical system of geocentric astronomy and cosmology.
890	al-Battani's *Kitab al-Zij*. Advanced presentation of Ptolemaic astronomy.
1220	Sacrobosco's *On the Sphere*. Medieval account of Ptolemaic astronomy.
1377	Oresme's commentary on Aristotle's *On the Heavens*. Critical appraisal of geocentric cosmology.
1543	Copernicus's *Revolutions of the Heavenly Spheres.* Heliocentric system.
1596	Kepler's *Cosmographic Mystery*. Cosmology of heliocentric system based on nested Platonic solids and planetary spheres.
1609	Kepler's *New Astronomy*. Elliptical motion of Mars. Line from Sun to Mars sweeps out equal areas in

	equal times. First attempt at a physical cosmology based on the concept of force.
1610	Galileo's *Starry Messenger*. First telescopic study of the heavens produces evidence for the heliocentric system. Multitude of stars found.
1632	Galileo's *Dialogue on Two World Systems*. Account of Ptolemaic and Copernican systems, in which the latter is depicted as superior.
1687	Newton's *Principia Mathematica*. Mathematical dynamics and the law of universal gravitation. Synthesis of Galilean inertial physics and Keplerian heliocentric astronomy.
1718	Halley identifies the proper motion of stars. The "fixed stars" are not fixed.
1755	Kant's *Universal Natural History and Theory of Heavens*. Island-universe theory. Oval-shaped white nebulae are autonomous star systems similar to the Milky Way but much more distant.
1781	Messier's catalog of nebulae.
1782	Herschel's first catalog of double stars. Most double stars are physically connected systems.
1826	Olber publishes the dark-night-sky paradox. The concept of the universe as a whole gives rise to questions about what one should observe.
1840	Lord Rosse detects the spiral character of some of the white nebulae.
1842	Doppler's formula for wavelength shifts in the spectra of moving sources.
1864	Huggins obtains emission spectrum for planetary nebula in Draco. Showed that some nebulae are not resolvable into stars, a fact that was seen as evidence against the island-universe theory.
1885	Nova in Andromeda nebula. If very distant, it would have to be of a brightness considered inconceivable to physicists of the period. Construed as evidence against the island-universe theory. Later identified as a supernova.
1912	Leavitt finds period-brightness relation for Cepheid variables in Smaller Magellanic Cloud. Basis for a method for determining distances to distant objects.
1914	Slipher detects large spectral shifts in white nebulae. These large speeds differentiate such nebulae from objects in the galaxy.

1917	Hooker 100-inch reflecting telescope at Mount Wilson.
1917	Einstein publishes paper on cosmology and general relativity.
1920	Great debate between Shapley and Curtis on the island-universe theory.
1922	Friedmann's dynamical solution of the relativistic cosmological equations.
1925	Extragalactic character of M 31 established by Hubble. Victory of the island-universe theory.
1927	Lemaître and relativistic solution for an expanding universe. Relates geometric models to results in nebular astronomy.
1929	Hubble publishes the red shift–distance law. Further research with Humason confirms law.
1931	Lemaître and the primeval atom. Universe begins in dense and hot state in an explosive event and has expanded ever since.
1932	Jansky detects radio waves from sources in the Milky Way.
1933	Eddington publishes *The Expanding Universe*.
1935	Milne's nonrelativistic theory of the expanding universe. Introduces the cosmological principle.
1944	Badde identifies two populations of stars. Population I stars are found in the arms of spiral nebulae and include the Sun and the stars in its neighborhood. Population II stars are found throughout galaxies but are concentrated in globular clusters, elliptical galaxies, and the centers of galaxies.
1948	Bondi and Gold's steady state theory. Hoyle formulates another version of the theory in terms of the general theory of relativity.
1948	Gamow, Bethe, and Alpher formulate the hot big bang theory.
1952	Baade revises Cepheid-variable distance method. Classical Cepheids are population I, and RR Lyrae are population II. Different period-luminosity relations hold for the two classes. Distances to galaxies increased by a factor of two.
1955	Ryle and the third Cambridge survey of radio sources. Ryle finds increased number of extragalactic radio sources with increased distance. Seen as evidence for the big bang–type theories.
1957	Burbidge, Burbidge, Hoyle, and Fowler's theory of nucleosynthesis in stars.

1963	Schmidt finds first quasar. Large red shift indicates great distance. Quasars are compact and extremely energetic sources of radiation.
1965	Penzias and Wilson detect the microwave background radiation. Dicke explains its cosmic origin. Victory for the big bang theory.
1968	Wheeler introduces the term "black hole."
1970	Rubin and Ford's study of galactic rotation curves indicates existence of dark matter.
1979	Walsh, Carswell, and Weyman discover gravitational lensing.
1982	Guth's theory of inflation for the very early universe.
1990	Hubble Space Telescope launched.
1991	COBE satellite probe confirms blackbody character of the cosmic background radiation. COBE and later instruments confirm existence of anisotropies in the radiation.
1998	Two teams of astronomers discover universal acceleration.
2003	Hubble's constant found to be 72 kilometers per second per megaparsec, with uncertainty of 10 percent.

GLOSSARY

acceleration: In 1998, astronomers found that the red shifts of distant galaxies were smaller than predicted by Hubble's law. This fact implied that the expansion of the universe was slower in the past and that the rate of expansion has increased with time. The discovery of acceleration has caused cosmologists to revise their models and to posit the existence of "dark energy" to account for the acceleration.

adaptive optics: A system used in large telescopes to cancel the disturbing effect of the atmosphere on the image of a star or galaxy. Pioneered in the 1980s, adaptive optics are employed in the giant telescopes located in Chile and Hawaii.

al-Tusi couple: A geometric device used by the mathematician al-Tusi in which straight-line motion is produced by a combination of circular motions. Used in a planetary model to replace the equant. The al-Tusi couple also raised conceptual questions about the Aristotelian opposition of straight-line and circular motion.

annual parallax: Change in angular position of a star as observed from Earth during its annual orbit about the Sun. Also known as trigonometric parallax because the distance of the star can be determined in terms of astronomical units by trigonometry from the angle of parallax.

astronomical unit: The distance from the Earth to the Sun. Distances within the solar system are often measured in terms of astronomical units. Abbreviated to a.u.

big bang theory: The standard cosmological theory accepted by most scientists today. Posits that the universe began in primordial conditions of extremely high density and temperature approximately 13 billion years ago and that the universe has been expanding ever since.

blackbody: An ideal body that emits all the radiation that it absorbs. For a given temperature of the body, there is a characteristic graph giving the intensity of the emitted radiation as a function of the frequency of the radiation. The concept of a blackbody arose in 1900 in Planck's quantum theory. The microwave background radiation displays a characteristic graph for a blackbody at three degrees Kelvin, indicating its cosmic origin in the primordial big bang.

black hole: An object so dense that no radiation can escape its gravitational field. The term was introduced by Wheeler in 1968. It is believed that supermassive black holes lie at the centers of quasars and galaxies.

celestial equator: The circle on the celestial sphere that is 90 degrees from the north celestial pole. The celestial equator is inclined at an angle of about 23 degrees to the ecliptic.

Cepheid variable: A type of variable star in which the period of variation is related to the absolute brightness or luminosity of the star. Named after the star δ Cephei in the northern constellation of Cepheus. By comparing the absolute brightness of a star with its apparent brightness, one can determine its distance. Cepheid variables are an important tool for determining distances to stars and nearby galaxies.

charge-coupled device: An electronic device for recording the image of an object in a telescope. Such devices have replaced photography. Pioneered in astronomy in the 1970s, CCDs are used in today's digital cameras.

cosmological constant: A constant introduced into the gravitational equations to produce a roughly static system in an extended system of masses. The presence of the constant corresponds to a repulsive tendency that acts over large distances. It was introduced by Neumann in 1896 and again by Einstein in 1917. With the discovery in 1998 that expansion is accelerating, cosmological models containing the cosmological constant have been a subject of renewed interest. The cosmological constant is sometimes known as the λ constant because this was the notation used by Einstein.

cosmological principle: The principle that the universe on a large scale looks the same from every point within it. The term was introduced by the British cosmologist Arthur Milne in the 1930s. Modern theories of cosmology of every stripe accept the cosmological principle.

dark matter: Matter that does not manifest itself in the form of electromagnetic radiation. Studies of the rotational motions of galaxies and clusters of galaxies have indicated the existence of large amounts of dark matter in the universe. Inflationary versions of the big bang theory also predict the existence of dark matter.

deferent: In Ptolemy's model for the motion of a planet the deferent is a large circle whose center is the Earth or a point near the Earth. The center of the planet's epicycle lies on the deferent. In the case of Venus and Mercury the epicycle center revolves on the deferent once each year. For Mars, Jupiter, and Saturn the epicycle center revolves on the deferent with a period characteristic of each planet.

diurnal parallax: The change in position of an object as observed from the surface and center of the Earth. The daily or diurnal motion of each of the planetary bodies—its rising and setting—occurs on a circle whose center is the center of the Earth. Because we observe the body from the surface of the Earth, at a substantial distance from the center, it appears to shift in direction during a 24-hour period. Only the Moon exhibits diurnal parallax that is large enough to be observable with the naked eye.

Doppler effect: A shift in the length of waves emitted from a source moving with respect to the observer. If the source is approaching the observer, the wavelength increases, and if it is moving away from the observer, it decreases. Light from a receding source is shifted to the red, and light from an approaching source is shifted to the blue.

eccentric circle: A circle whose center is close to but does not coincide with the Earth. Such a circle is said to be situated eccentrically with respect to the Earth. Used by Hipparchus and Ptolemy to represent the motion of the Sun.

ecliptic: The path traced by the Sun during its annual eastward circuit around the celestial sphere. The Moon and the planets also move eastward on the celestial sphere within a narrow band surrounding the ecliptic.

elliptical geometry: A geometry different from Euclid's, characterized by the property that the angles of a triangle add up to more than 180 degrees. First discussed by Riemann in 1854, such a geometry is also known as Riemannian geometry. The surface of a sphere in which a line is defined as a great circle—the intersection of a plane through the center of the sphere and the surface—is a model for elliptical geometry. The first cosmological solutions of the field equations of general relativity by Einstein and Friedmann assumed that the universe was finite and that the geometry of space was elliptical. In such a world, light from a given source will eventually travel around a great circle in space and return to the source. Cosmologists today believe elliptical geometry is an unlikely choice as the geometry of space.

epicycle: In Ptolemy's model for planetary motion the planet revolves on a small circle called an epicycle whose center revolves on a larger circle known as the deferent.

equant: A device used by Ptolemy to account for small variations in planetary motion. The equant is a point offset from the center of the planetary deferent, the latter being situated eccentrically with respect to the Earth. The motion of the center of the epicycle on the deferent is uniform with respect to the equant. Hence neither the center of uniform angular motion nor the Earth itself is at the center of the deferent.

equivalence principle: The action of a uniform gravitational field on a system of bodies may be regarded as equivalent to the same system in which no force acts but in which the bodies are subjected to a uniform acceleration. Adoption of the equivalence principle was the first step in Einstein's development of the general theory of relativity.

Euclidean geometry: The traditional geometry of space that was described by Euclid of Alexandria around 300 B.C. In Euclidean geometry the angles of a triangle add up to 180 degrees and the relationship between the sides and diagonal of a triangle has the very simple Pythagorean form $a^2 + b^2 = c^2$.

Euclidean geometry is currently favored by cosmologists as the geometry of the universe.

extragalactic: Until well into the twentieth century the word galaxy referred exclusively to the Milky Way galaxy. Hence any object that was outside of the Milky Way was said to be extragalactic. This meaning has persisted even as the term galaxy has been extended.

field equations: Equations that describe the action of gravity in Einstein's general theory of relativity. The field equations are written in terms of tensor notation using methods from a branch of mathematics known as differential geometry.

galaxy: The word comes from the Greek for "milky way" and originally referred to the Milky Way system of stars. With Hubble's discovery in 1925 that spiral nebulae and other white nebulae are star systems similar to the Milky Way and external to it, the word has come to refer to any of these large star systems.

globular cluster: A spherically shaped cluster containing a very large number of stars. The galaxy is framed symmetrically by globular clusters. This fact was used by Shapley in 1920 to infer that the Sun is situated some distance away from the center of the Milky Way galaxy.

gravitational lens: An object whose gravitation bends the light from a more distance source. The image of a distant galaxy or quasar may be distorted, magnified, or multiplied as a result of the action of an intervening galaxy or cluster of galaxies lying along the line of sight to the more distant object. Gravitation lenses enable some of the most distant objects in the universe to be studied more closely than would be possible otherwise. They can also be used to determine the distance to the lensed object and therefore to determine the value of Hubble's constant.

gravitational wave: A rapid and sudden change in motion of a massive object will result in the propagation of gravitational waves, undulations in the gravitational field of the object that are propagated at the speed of light. Such waves are analyzed using the general theory of relativity. They have never been observed directly but have been inferred to exist from the observation of radio waves emitted by binary pulsars. Currently, there are large scientific projects in place to detect gravitational waves directly.

Great Attractor: An increase in density of matter at a distance of 150 million light-years in the directions of the constellations Hydra and Centaurus. In analyzing large numbers of galactic red shifts and distances during the 1980s, astronomers identified a departure from pure Hubble flow, leading them to infer the existence of what they called the Great Attractor. Its existence indicates that there is significant local inhomogeneity in the universe.

Hertzsprung-Russell diagram: A graph in which the brightness of a star is plotted against surface temperature. For most stars, there is a linear relationship between brightness and surface temperature. There is also a class of bright red giants with a lower relative surface temperature. The H-R diagram is used to study the evolution of stars.

Hubble's constant: The constant *H* in Hubble's law, measured in units of kilometers per second per megaparsec. For every increase in distance of one megaparsec the recessional velocity of an object increases by *H* kilometers per second. Hubble's constant is currently believed to have a value of 70 with an uncertainty of ±15 percent.

Hubble's law: The red shift of an object such as a galaxy is linearly proportional to its distance from the Earth. Red shift is conventionally measured as the velocity of recession for the corresponding Doppler shift. The law is $v = Hd$, where v is the nominal recessional velocity of the object, d is its distance, and H is a constant known as Hubble's constant. If the red shift represents an actual recessional velocity—as most astronomers believe—then Hubble's law implies that the universe is expanding. In relativistic cosmology the motion of recession is understood to result from the expansion of space. At a certain distance the recessional velocity dominates any local or peculiar motions, giving rise to what is known as pure Hubble flow.

hyperbolic geometry: A geometry different from Euclid's characterized by the property that the angles of a triangle add up to less than 180 degrees. The functions describing the relationship between sides and angles in this geometry are the hyperbolic functions. Hyperbolic geometry was the first non-Euclidean geometry to be studied, with published accounts by Lobachevsky and Bolyai in the first part of the nineteenth century. In 1924 Friedmann constructed an infinite relativistic world model in which the geometry of space is hyperbolic.

inertia: The tendency of a body in the absence of external forces to continue in uniform straight-line motion. The principle was the basis of the new physics of the seventeenth century that replaced traditional Aristotelian physics.

inflation: In the big bang theory a very short-lived event involving exponential expansion of the whole universe in the first instant of the big bang. It is believed that inflation produced the homogeneity seen today in the microwave background radiation. It also explains the absence in the universe of a particle known as a magnetic monopole.

interferometer: A device that enables one to locate the position of a source by analyzing the interference patterns generated by a signal arriving at two different receivers. Very long base lines between receivers have been used in radio astronomy to provided unprecedented levels of resolution.

island-universe theory: The white and oval-shaped nebulae such as M 31 and M 51 are autonomous star systems, or "island universes," similar to the Milky Way galaxy and external to it.

light-year: The distance light travels in one year. The nearest star is just over four light-years away. The galaxy is 100,000 light-years in diameter, and the Andromeda galaxy is 2.5 million light-years distant.

Mach's principle: The inertial properties of matter are determined by the distribution of matter throughout the universe. It influenced Einstein in adopting a form of what later became known as the cosmological principle as a basis for his cosmological solutions of the field equations.

Messier catalog: A catalog of 103 nebulae published by the French astronomer Charles Messier in 1781. The most prominent nebulae in the sky are identified by their Messier number. For example, M 31 is the Andromeda galaxy, and M 13 is the globular cluster in Hercules.

microwave background radiation: Radiation coming from all parts of the sky, possessing a temperature corresponding to a blackbody at three degrees Kelvin. The discovery of the microwave background radiation in 1965 was the event that confirmed (for most scientists) the big bang theory of the universe. The radiation is believed to have been emitted in the primordial bang that created the universe. The radiation is also known as the cosmic background radiation.

nebula: A fuzzy or milky object, from the Latin for "cloud." There are several different types of nebula, based on their appearance under telescopic examination: planetary nebulae, white nebulae (spiral and elliptical), reflection nebulae, globular clusters, and open clusters. The class of white nebulae consists of galaxies external to the Milky Way.

nesting principle: Adopted by Ptolemy in developing his planetary cosmology, the principle asserts that there are no empty spaces between the successive spherical shells within which the planets move. The principle enabled Ptolemy to determine the dimensions of the planetary system.

nucleosynthesis: Process in which the nuclei of elements fuse together and form heavier nuclei, releasing energy as they do so. The energy emitted by a star comes from thermonuclear fusion at the star's center. For stars on the main sequence of the H-R diagram, protons fuse to form helium nuclei. This process is known as the carbon-nitrogen cycle because carbon and nitrogen are formed temporarily at one step in the sequence of reactions.

Olber's paradox: If the sky is evenly populated by Sun-like stars and the universe is very large or infinite, then the total radiation reaching the Earth should be very large. The night sky should be bright, but paradoxically is not. Modern solutions of Olber's paradox use the fact that in the big bang model the amount of radiation reaching the Earth is constrained by the limited number and age of radiant bodies that have formed since the creation of the universe.

opposition: If the Sun, Earth, and the planet lie in a straight line, then the planet is said to be in opposition. At opposition the planet reaches its highest point in the sky at midnight.

parsec: The distance of an object exhibiting an annual parallax of one second of arc. A parsec is approximately 3.26 light-years. The distances to galaxies are typically given in megaparsecs, or units of distance equaling one million parsecs.

perfect cosmological principle: The universe on a large scale is the same at all points in space *and* in time. Foundation of the steady state theory of the universe, developed by Bondi, Gold, and Hoyle in the 1940s and 1950s.

photometry: The measurement of the brightness of stars and galaxies. Brightness is measured on the logarithmic magnitude scale, where each increase in magnitude corresponds to a 2.5-fold increase in brightness. Photometry has moved from optical methods, to photography, to highly sensitive electronic devices.

Platonic axiom: The motion of all celestial bodies is circular and uniform (constant angular motion). The basic axiom of Greek mathematical astronomy, it dominated the study of planetary motions up to the time of Kepler.

Platonic solid: A polyhedron that is convex (no indentations) and in which each face is a congruent regular polygon. Euclid showed around 300 B.C. that there are only five such solids, the tetrahedron, cube, octahedron, dodecahedron, and icosahedron. Kepler constructed a heliocentric cosmology by fitting the six planetary orbits within a nested sequence of the Platonic solids.

precession: The slow movement in a westward direction of the two points of intersection of the celestial equator and the ecliptic.

proper motion: Real as opposed to apparent motion of a star. In proper motions the stars are actually moving in space. First detected by Halley in 1718. Proper motions were one of the first subjects studied in stellar astronomy, indicative of the beginning of an interest in the universe beyond the solar system.

pulsar: A rapidly rotating neutron star that emits radio waves along the ends of an axis inclined to its axis of rotation. The radio waves are received on the Earth as a sequence of pulses. Study of a pulsar-star binary system in the late 1970s led to the detection of what were inferred to be gravitational waves.

quasar: A quasi-stellar radio source, an extremely luminous and very distant object. The first quasars to be discovered were energetic emitters of radio waves and possessed extremely large red shifts. Radio-silent quasar-like objects were subsequently found. The existence of quasars has been used as evidence that the universe is evolutionary.

red shift: A shift to lower frequency in the spectrum of a star or galaxy. For optical spectra a shift to the red. A red shift may result from the action of gravitation on the emitted light. Most often, it is caused by the motion of the light source away from the Earth. The recessional velocity of an object may be peculiar (arising from its motion through space relative to the Earth) or cosmological (arising from the expansion of space according to the general theory of relativity). Cosmological red shifts are described by Hubble's law and are very large, increasing linearly with distance.

reflector: A telescope in which the light from a source is reflected and focused by a primary mirror. The focused image is examined by an eyepiece, spectroscope, or photometric receiver. Reflecting telescopes today are the largest optical instruments in astronomy and are the main instruments used to study faint and distant objects. Hubble's law was discovered through observations of nebulae made with the Hooker 100-inch (250 centimeter) reflector at the Mount Wilson Observatory in southern California.

refractor: A telescope in which the light from a source passes through a main objective lens and is focused and examined by an eyepiece, spectroscope, or photometric receiver. Because the objective lens can only be supported around its circumference, the size of refractors is limited.

retrograde motion: The planets move eastward along the ecliptic, except for certain periods, when they move backward for a while before resuming their forward motion. The backward motion is called retrograde motion.

spectroscopy: The study of spectra. Spectroscopy gives information about the chemical constitution, temperature, and motion of a star or nebula.

spiral nebula: A nebula possessing a spiral structure. First identified by Lord Rosse in 1840, spiral nebulae were found to be external galaxies similar to the Milky Way galaxy.

statistical parallax: Other things being equal, the size of proper motion in a star is inversely proportional to its distance. By measuring the average proper motion of a group of stars, one obtains an estimate of the distance to this group. If the stars are relatively close together and their number is large, the estimate will be very accurate. Distances measured in this way are said to be obtained by the method of statistical parallax.

steady state theory: A theory of an expanding universe that supposes that the average density of matter remains constant in time. As the universe expands, new atoms are created in space, compensating for the decrease in density that would otherwise occur. The universe is in a steady state, unchanging on a large scale in both space and time.

stellar aberration: The apparent direction of starlight received by an observer on Earth is affected by the motion of the Earth as it moves about the Sun. First discovered in the 1720s, aberration confirmed that light propagates with finite velocity.

supernova: A star that suddenly appears and shines for a short period of time with a brightness many millions of times the brightness of a normal star. Such an event results from a sudden and massive explosion in the star and has causes related to changes in energy processes within the star.

zone of avoidance: A region centered around the equator of the Milky Way characterized by the absence of spiral and other white nebulae. Many nineteenth-century astronomers believed that the existence of the zone indicated that the nebulae were systemically connected to the Milky Way galaxy and so were not external island universes. Curtis showed in the first part of the twentieth century that many galaxies contain bands of obscuring dust and gas in their equatorial regions. The presence of such matter in the Milky Way explains the zone of avoidance.

BIBLIOGRAPHY

Aaboe, Asger. 2001. *Episodes in the Early History of Astronomy.* New York: Springer. Good expository account of the technical side of Babylonian astronomy and Greek astronomy and cosmology.

Arnaldes, R., and A. Z. Iskandar. 1975. "Ibn Rushd." In *Dictionary of Scientific Biography,* ed. C. C. Gillispie. Vol. 12, 1–9. Leading Islamic commentator on Aristotle, was a major influence in the Latin West, where he was called Averroës. Opposed Ptolemy's astronomical system, upholding instead the Eudoxan-Aristotelian conception of homocentric spheres.

Bartusiak, Marcia. 2000. *Einstein's Unfinished Symphony: Listening to the Sound of Space-Time.* Washington, D.C.: Joseph Henry Press. A popular history of Einstein's general theory of relativity, with emphasis on attempts to confirm the theory's predictions about gravity.

Bernstein, Jeremy. 1973. *Einstein.* Fontana Collins. A nontechnical account of Einstein's life and science.

Collins, Harry, and Trevor Pinch. 1993. *The Golem: What Everyone Should Know about Science.* New York: Cambridge University Press. Contains essays on Eddington's 1919 eclipse expedition, Weber's attempt to find gravitational waves, and the problem of missing solar neutrinos.

Crowe, M. J. 1994. *Modern Theories of the Universe: From Herschel to Hubble.* New York: Dover Publications. With excerpts from original sources and exercises. Very good on Herschel and on the "great debate" between Shapley and Curtis.

———. 2001. *Theories of the World from Antiquity to the Copernican Revolution.* New York: Dover Publications. Contains excerpts from original sources and useful classroom exercises.

Cullen, Christopher. 1996. *Astronomy and Mathematics in Ancient China: The Zhou bi suan jing.* Cambridge: Cambridge University Press. A readable account of an important work as well as a good introduction to ancient Chinese astronomy.

Dijksterhuis, E. J. 1961. *The Mechanization of the World Picture*. Oxford: Oxford University Press. Most detailed single-volume history of physical science from the Greeks to Newton.

Dressler, Alan. 1994. *Voyage to the Great Attractor: Exploring Intergalactic Space*. New York: Vintage Books. A first-person account of research in extragalactic astronomy.

Dreyer, J.L.E. 1953. *A History of Astronomy from Thales to Kepler*. New York: Dover Publications. Reprint of 1905 edition; revised with a forward by W. H. Stahl. Classic still useful today. Reflects the instrumentalist view of Greek astronomy current among historians around 1900.

Eddington, Arthur. 1920. *Space Time and Gravitation: An Outline of the General Theory of Relativity*. Cambridge University Press. The first detailed English-language exposition of Einstein's theory.

Evans, James. 1998. *The History and Practice of Ancient Astronomy*. New York: Oxford University Press. Expository account including technical details of all aspects of ancient astronomy.

Galileo Galilei. 1957. *Discoveries and Opinions of Galileo*. Translated and with introduction and notes by Stillman Drake. New York: Doubleday Anchor Books. Contains a selection of Galileo's writings, including *Starry Messenger* (1610) and *Letter to the Grand Duchess Christina* (1615). In the latter, Galileo argues for the compatibility of biblical teaching and the new astronomy.

Gillispie, C. C. 1970. *Dictionary of Scientific Biography*. New York: Scribner. Compendium of biographical information on scientists of the past. An essential reference work.

Gingerich, Owen, ed. 1984. *Astrophysics and Twentieth-Century Astronomy to 1950: Part A*. Vol. 4. *The General History of Astronomy*. Cambridge: Cambridge University Press. Collection of essays covering many aspects of astronomy in the twentieth century.

Goldstein, Bernard. 1972. "Theory and Observation in Medieval Astronomy." *Isis* 63: 39–47. Contends that there was some willingness on the part of medieval commentators to revise Ptolemaic astronomical models in light of observation.

———. 2002. "Copernicus and the Origin of His Heliocentric System." *Journal for the History of Astronomy* 23: 219–35. Maintains that Copernicus rejected the Ptolemaic system because the angular velocities of the Sun and the inferior planets do not decrease with distance from the center. Suggests that Copernicus's belief in material spheres led him to adopt a heliocentric theory, and that otherwise he would have developed a Tychonic system.

Hacking, Ian. 1989. "Extragalactic Reality: The Case of Gravitational Lensing." *Philosophy of Science* 56: 555–81. Provocative essay discussing the status of galaxies and quasars as real objects.

Harrison, E. R. 1981. *Cosmology: The Science of the Universe.* Cambridge: Cambridge University Press. Textbook on cosmology that makes good use of historical material.

Hetherington, Norriss, ed. 1993. *Encyclopedia of Cosmology: Historical, Philosophical, and Scientific Foundations of Modern Cosmology*. New York: Garland Publishing. Articles on a range of topics in cosmology, from antiquity to the twentieth century.

Hirsch, R. F. 1983. *Glimpsing an Invisible Universe: The Emergence of X-Ray Astronomy.* Cambridge: Cambridge University Press. Well-documented history of a new field of astronomical investigation.

Horgan, John. 1996. *The End of Science: Facing the Limits of Knowledge in the Twilight of the Scientific Age.* Reading, Mass.: Addison-Wesley. Provocative critique of science by a *Scientific American* journalist. Suggests that modern cosmology often lacks empirical meaning and is more tentative than practitioners would lead one to believe.

Hoyle, Fred. 1952. *The Nature of the Universe.* Oxford: Blackwell. Contained the first popular exposition of the steady state theory.

———. 1990. "An Assessment of the Evidence against the Steady-State Theory." In *Modern Cosmology in Retrospect*, ed. B. Bertotti et al., 221–31. Cambridge: Cambridge University Press. The author suggests that it is possible to salvage something from the ashes of the steady-state theory. Some interesting historical remarks on the theory.

Hubble, Edwin. 1936. *The Realm of the Nebulae.* New Haven, CT: Yale University Press. An account of nebular astronomy addressed to a wide audience. Written from the viewpoint of an observationalist, it stands in contrast to contemporary popular books on astronomy and cosmology by such theoreticians as Eddington and Jeans.

———. 1942. "The Problem of the Expanding Universe." *Science* 95: 212–15. Discussion of difficulties Hubble discerned in the theory of the expanding universe.

Huggins, William. 1909. *The Scientific Papers of William Huggins,* ed. William and Lady Huggins. London: William Wesley and Son. Original papers of Huggins on spectroscopy. Includes editorial remarks by the author as he looks back on his work.

Jammer, Max. 1969. *Concepts of Space: The History of Theories of Space in Physics.* Cambridge, Mass.: Harvard University Press. Interesting as a survey of a single concept in the history of physics. Author wrote similar histories for the concepts of mass and force.

Koestler, Arthur. 1963. *The Sleepwalkers: A History of Man's Changing Vision of the Universe.* New York: Universal Library. First published 1959. Contains biographies of Copernicus, Tycho, Kepler, and Galileo. Critical of Copernicus and laudatory of Kepler.

Koyré, Alexandre. 1957. *From the Closed World to the Infinite Universe.* New York: Harper & Row. Classic that has influenced modern interpretations of the Scientific Revolution.

Kragh, Helge. 1996. *Cosmology and Controversy: The Historical Development of Two Theories of the Universe.* Princeton, N.J.: Princeton University Press. Detailed history of twentieth-century cosmology in both its theoretical and observational aspects.

Kuhn, T. S. 1957. *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought.* Cambridge, Mass.: Harvard University Press. This study of the Copernican revolution prepared the way for Kuhn's integration of social and conceptual aspects of the history of science in his famous 1962 *Structure of Scientific Revolutions.*

Meadows, A. J. 1984. "The New Astronomy." In *Astrophysics and Twentieth-Century Astronomy to 1950: Part A.* Vol. 4. *The General History of Astronomy*, ed. O. Gingerich, 59–72. Cambridge: Cambridge University Press. Account of the emergence of astrophysics in the nineteenth century.

Murschel, Andrea. 1995. "The Structure and Function of Ptolemy's Physical Hypotheses of Planetary Motion." *Journal for the History of Astronomy* 26: 33–61. In-depth study of Ptolemy's *Planetary Hypotheses.*

Neugebauer, Otto. 1968. "On the Planetary Theory of Copernicus." *Vistas in Astronomy* 10: 89–103. A critique of Copernicus from a modern mathematical viewpoint.

———. 1969. *The Exact Sciences in Antiquity.* New York: Dover Publications. Originally published in 1957. A masterful introduction to ancient Babylonian and Greek astronomy, with emphasis on its mathematical aspects.

North, J. D. 1965. *The Measure of the Universe: The History of Modern Cosmology.* Oxford: Clarendon Press. A scholarly history of twentieth-century cosmology up to 1965. Excellent on mathematical and theoretical aspects of cosmology, the book covers parts of the history that are relatively little known today.

———. 1994. *The Norton History of Astronomy and Cosmology.* New York: W. W. Norton. An informative one-volume survey.

Olcott, W. T. 1954. *Olcott's Field Book of the Skies,* rev. R. N. Mayall and M. W. Mayall. 4th ed. New York: G. P. Putnam's Sons. A classic American guide to the planets and constellations.

Pedersen, Olaf. 1993. *Early Physics and Astronomy: A Historical Introduction.* Cambridge: Cambridge University Press. Revised edition; first published 1974. Covers a range of topics in the history of the exact sciences.

Price, Derek. 1959. "Contra-Copernicus: A Critical Re-estimation of the Mathematical Planetary Theory of Ptolemy, Copernicus and Kepler." In *Critical Problems in the History of Science,* ed. Marshall Clagett, 197–211. Madison: University of Wisconsin Press. Thought-provoking attempt at a reassessment of Ptolemy and Copernicus.

Schofield, C. J. 1981. *Tychonic and Semi-Tychonic World Systems.* New York: Arno Press. An account of a neglected part of the history of cosmology.

Silk, Joseph. 1980. *The Big Bang: The Creation and Evolution of the Universe.* San Francisco, Calif.: W. H. Freeman and Company. Widely used textbook on cosmology.

Smith, R. W. 1982. *The Expanding Universe: Astronomy's 'Great Debate' 1900–1931.* Cambridge: Cambridge University Press. Readable and well-documented history of the famous developments of the 1920s.

Smoot, George, and Keay Davidson. 1993. *Wrinkles in Time.* New York: William Morrow and Company. An account of the COBE satellite venture by the project leader.

Sullivan, W. T. 1984. *The Early Years of Radio Astronomy.* Cambridge: Cambridge University Press. Very good on Jansky's early work. Includes essays by many of the leading figures in the development of radio astronomy.

Taub, L. C. 1993. *Ptolemy's Universe: The Natural Philosophical and Ethical Foundations of Ptolemy's Astronomy.* Chicago, Ill.: Open Court. A good discussion of the nontechnical aspects of Ptolemy's astronomy.

Thorndike, Lynn. 1949. *The Sphere of Sacrobosco and Its Commentators.* Chicago, Ill.: University of Chicago Press. Contains the Latin text and English translation of Sacrobosco's book.

van Helden, Albert. 1985. *Measuring the Universe: Cosmic Dimensions from Aristarchus to Halley.* Chicago, Ill.: University of Chicago Press. An informative survey of the subject.

Weinberg, Steven. 1977. *The First Three Minutes: A Modern View of the Origin of the Universe.* New York: Basic Books. Publication of this book indicated a growing interest among particle theorists in cosmology during the 1970s.

Westman, R. S. 1975. "Three Responses to the Copernican Theory: Johannes Praetorius, Tycho Brahe, and Michael Maestlin." In *The Copernican Achievement,* ed. R. S. Westman, 285–345. Berkeley: University of California Press. Argues that an instrumentalist construal of Copernicus's achievement played a major role in its reception.

INDEX

About the Author

CRAIG G. FRASER is a professor in the Institute for History and Philosophy of Science and Technology at the University of Toronto. He is the editor-in-chief of the journal *Historia Mathematica*, and has written numerous articles on the history of modern cosmology.